From Sounds to Symbols: A Survey of Phonetically and Phonologically Informed Language Generation

Tyler Loakman[1*, 2], **Joseph James**[1], **Chenghua Lin**[2]
[1]Department of Computer Science, The University of Sheffield, UK
[2]Department of Computer Science, The University of Manchester, UK
chenghua.lin@manchester.ac.uk

Abstract

Text, although written using symbols and characters, is not void of connection to the predominantly spoken form of language. Whilst syntax and semantics are the cornerstones of meaningful text, a substantial portion of human communication relies heavily on the specific sounds and sound patterns observed in language. In this survey paper, we present an outline of the areas of computational linguistics, specifically natural language generation (NLG), that are either dependent upon, or benefit from, an additional understanding of phonetics and phonology. We cover a range of subdisciplines, from practical areas such as and grapheme-to-phoneme conversion (G2P) to creative endeavours such as poetry and humour generation. We aim for this overview to situate the role of phonetics and phonology in NLG, and invite the community at-large to consider how the study of such features may be able to positively impact different domains through fusing characteristics of both speech and text.

1 Introduction

To successfully generate coherent language, computational models must understand several key areas of linguistics. Fundamentally, the area of syntax dictates the ordering of words and the grammatical relationships between constituents (Moravcsik, 2006; Van Valin, 2001), whilst semantics addresses the meaning of language and the message that is to be conveyed (Kroeger, 2018; Hofmann, 1993). The former is essential to not generate random assortments of words, turning "cat the sat on mat his" into "his cat sat on the mat", whilst the latter is what explains the nonsensical nature of the phrase "They quickly ran *upwards down* the stairs". Both of these elements are essential to the majority of written language. However, as computational approaches such as large language models (LLMs) become adept at handling these elements of language, the challenge presented by other areas of linguistics becomes much clearer. Of these, the related areas of phonetics and phonology concern the production, perception, and distribution of speech sounds across human languages (Rogers, 2013; De Lacy, 2007; Wells, 1982b). Such elements are essential to many aspects of human communication, even in its written form, complementing syntax and semantics and being the fundamental elements that separate prose from poetry and other creative uses of language.

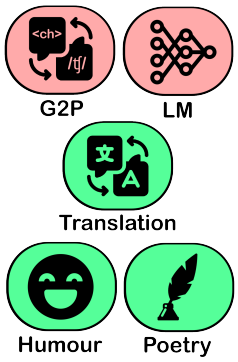


Figure 1: Illustration of the areas surveyed in this work. G2P refers to grapheme-to-phoneme conversion, whilst LM refers to the general domain of language modelling. We distinguish between upstream foundational tasks (red) and downstream applications (green).

1.1 Existing Surveys

To the best of our knowledge, this work presents the first survey paper dedicated to the topic of phonetically and phonologically motivated developments in NLG. However, there exists a range of survey papers that cover individual areas in depth, particularly those where conventional approaches do not rely explicitly on such characteristics. Cheng et al. (2024) provides an overview of G2P in multilingual settings, covering methods, datasets and evaluation metrics, Gonçalo Oliveira (2017) cover the area of intelligent poetry generation, Watanabe and Goto (2020) address the automatic generation of lyrics, and Loakman et al. (2025d) and Amin and Burghardt

[*]Work performed while at the University of Sheffield.

Proceedings of the 19th International Natural Language Generation Conference: Long Papers, pages 623–638
October 17-21, 2026 ©2026 Association for Computational Linguistics

(2020) survey approaches to humour generation. We focus on novel approaches to incorporating phonetic and phonological knowledge into these tasks, rather than overall paradigms and state-of-the-art approaches, where the novelty is not linguistically motivated. The primary aim of this survey is one of breadth concerning the inclusion of phonetic and phonological knowledge in natural language generation, demonstrating the richness of research in this subdiscipline for an domain-adjacent readership, rather than per-topic depth for area specialists.

1.2 Survey Outline

In this survey, we focus primarily on works where phonetic and phonological knowledge has been incorporated into the process of natural language *generation*. We do, however, take a broad view of generation to also include elements such as the natural language explanation of phonetically and phonologically motivated text. Regarding paper inclusion, we utilised a systematic backward snowball approach beginning with articles found on the ACL Anthology and Google Scholar using a list of domain-relevant keywords (see Appendix A for the full list of terms). From there, cited works were obtained and considered for inclusion. In §2.1 we outline grapheme-to-phoneme conversion, a fundamental step for most other phonetically and phonologically informed works. In §2.2 we cover the broad topic of language modelling and approaches to incorporating relevant knowledge without a specific downstream task (including prompting methods, knowledge probing, and general evaluation). Following this, we address the large scale generative tasks of poetry generation in §3.1 and humour generation in §3.2, followed by translation in §3.3.

2 Upstream Foundational Tasks

2.1 Grapheme-to-Phoneme Conversion

Grapheme-to-phoneme conversion (G2P) is a common prerequisite for the processing of text using phonetic and phonological knowledge for downstream tasks, particularly for languages with opaque deep orthography. G2P centres on the process of converting *graphemes*, the character-level writing units of a language, to *phonemes*, the smallest meaningful units of sound. As such, it allows the investigation of the sounds that would be produced if written text was to be read aloud, assuming a specific accent. For instance, G2P may convert the word *bath* into /bɑːθ/ for a British speaker of Southern Standard British English (SSBE, Wells, 1982b), /baθ/ for a speaker of General Northern English (Wells, 1982b), or /bæθ/ for a speaker of General American (Wells, 1982a).[1] It is therefore paramount to the task of G2P that phonetic and phonological knowledge is available in the form of transcribed training data such as pronunciation dictionaries or aligned speech and text (Ribeiro et al., 2023; Route et al., 2019; Aquino et al., 2019).

G2P is commonplace in tasks such as automatic speech recognition (ASR) and speech synthesis (i.e., text-to-speech), particularly in multilingual and low-resource settings (Cheng et al., 2024). The inclusion of phonetic transcriptions provides the spoken form of words, allowing the transfer of knowledge across languages with different orthographic conventions but shared phoneme inventories. This has motivated approaches that exploit phonological similarity and cross-lingual transfer to extend models to under-resourced languages (Matogawa et al., 2024; Hammond, 2023; Li et al., 2022; Deri and Knight, 2016). In situations where textual data is scarce, the incorporation of multimodal speech data to learn pronunciations has also been well researched (Ribeiro et al., 2023; Route et al., 2019; Aquino et al., 2019), facilitating cross-modal transfer to alleviate resource constraints.

Owing to the ever-growing scale of model requirements, including quantities of training data, a growing focus in G2P is computational efficiency. Recent works such as Kim et al. (2022) present methods to make G2P more appropriate for complex, real-time applications. They propose a non-autoregressive transformer for bilingual G2P, achieving accuracy comparable to autoregressive baselines whilst offering substantially faster inference, making them suitable for low-latency applications. Similarly, LiteG2P (Wang et al., 2023) presents a highly parameter-efficient and parallelised hybrid connectionist temporal classification (CTC) model that incorporates G2P mapping dictionaries alongside maximum mapping length restrictions and masking rules to suppress invalid phoneme predictions for each grapheme.

A persistent challenge that faces G2P, and has measurable downstream impacts (particularly regarding puns and rhyme) is that of homographs. Homographs concern words that have identical orthography yet differ in their pronunciation based on context (e.g., "take a bow" versus "a bow and arrow"). G2P systems capable of homophone disambiguation are becoming increasingly common, such as SoundChoice (Ploujnikov and Ravanelli, 2022), which leverages BERT-

[1] These constitute three distinct collections of accents, each with variation in practice.

based semantic embeddings to incorporate contextual information, allowing for sentence-level disambiguation of homographs. Similarly, Han et al. (2024) present an in-context knowledge retrieval approach to G2P, where an LLM is prompted with the sentence and its phonetic transcription, using semantic context to find the correct pronunciation of homographs.

Similar to the linguistic ambiguity posed by homographs, the prevalence of code-mixing in many language communities presents an increased need for multilingual G2P models capable of handling switches between languages and writing systems. To this end, Cho et al. (2020) present a multi-stage code-mixed G2P model for Korean which applies transliteration at the sentence level to handle agglutinative morphology. They decompose the task into stages, which includes detecting and separating non-Korean segments in *eojeols*[2] and transliterating them into Korean pronunciations to improve the handling of code-mixed orthographies. Additionally, Zhu et al. (2022) demonstrated that byte-level models (e.g., ByT5) outperformed standard token-level models (e.g., mT5) trained on a large scale multilingual G2P dataset. Rather than relying on predefined subword tokenisation, ByT5 works directly on UTF-8 bytes which avoids segmentation errors and vocabulary mismatches, thereby reducing the phoneme error rate, especially in low-resource languages.

A further challenge to G2P is presented with underspecified orthographies, seen in languages such as Hebrew, Arabic, and Persian, where the given writing systems do not directly reflect all phonological contrasts present in a language's spoken form. In particular, Persian (i.e., Farsi) has become a popular focus for G2P research due to its underspecified orthography, in which short vowels (e.g., ezāfe particles) are not typically represented in writing. Rahmati and Sameti (2024) present *GE2PE*, a Persian G2P model to handle unwritten short vowels, ezāfe detection, and homograph disambiguation. GE2PE handles inputs at the sentence level using ByT5 to incorporate morphological, syntactic, and semantic contexts which would otherwise not be possible at the word-level. Similarly, Qharabagh et al. (2025) show *Finglish* prompting (i.e., romanised Farsi) to improve G2P accuracy without additional training. They further introduce a sentence-level benchmark for polyphone disambiguation and ezāfe prediction, also releasing Kaamel-Dict, an open-licensed Persian G2P dictionary. Further research has also explored alternative disambiguation strategies for handling Persian homographs using intermediaries such as *Finglish* (**?**Bertina et al., 2025).

[2]An *eojeol* is a fundamental linguistic unit of Korean, consisting of a sequence of characters separated from the rest of a text with whitespace.

2.2 Language Modelling

Language modelling concerns the development of computational models of linguistic structure and meaning, most essentially via being able to predict the most probable next word given a previous sequence (Chang and Bergen, 2024). We first discuss more general literature that addresses the inclusion of phonetic and phonological knowledge in the language modelling process, abstracted away from any particular downstream application, covering aspects such as prompting approaches, phonetic/phonological knowledge probing, and general phonetics/phonology benchmarks.

Rather than relying entirely on retroactive knowledge injection, phonetic and phonological knowledge can be incorporated into language models from the offset via being included in the pretraining stage. To this end, Goriely et al. (2024) present a pretraining paradigm where LLMs are trained on continuous streams of phonemes by removing word boundaries and applying character-level tokenisation. This approach allows model to learn enhanced phoneme-based knowledge at the cost of a slight decrease in performance on traditional language understanding tasks. Similarly, Bunzeck et al. (2024) compare grapheme- and phoneme-based Llama models and observe that grapheme models perform better overall on grammatical tasks, whereas phoneme models trade some performance for improved understanding of phonological tasks. Along the same lines, Gale et al. (2023) present BORT, a BART-based model pre-trained on orthographic and phonemic data through a self-supervised phoneme-to-grapheme (P2G) translation task that aims to restore orthography. Experimenting with pronunciation, spelling, and noise transforms, BORT is fine-tuned on clinically relevant tasks in speech language pathology and shows substantial gains over baselines. Performing pretraining on phonemic representations has also been explored for ASR (Masumura et al., 2020). Tsvetkov et al. (2016) introduce *polyglot* language models, which train a single recurrent neural network (RNN) over phonetic symbol sequences from multiple languages using shared representations. They condition the models on typological features of the target language to guide cross-lingual transfer. Their results show that polyglot models generalise better to held-out phonetic sequences than comparable monolingual models, and

phonetic feature representations learned in multilingual settings are of higher quality than those trained in isolation. An analysis of the embedding space shows that the models capture features such as vowel length, nasality, and consonant-vowel distinctions, with similar sounds clustering across languages.

Abstracted from downstream tasks, the phonetic and phonological understanding of models can also be assessed through observing the characteristics learnt during training, and the development of explicit phonetic word embeddings. For example, Gutscher and Pucher (2024) investigate whether articulatory characteristics, dialectal differences, and speaker variances can be encoded in embeddings from audio segments, using quinphones[3] in the embeddings of Austrian German speakers. They show that characteristic phones are captured effectively in the embedding space, allowing for better classification of dialects. Finally, Malisz et al. (2024) release PRODIS, a large multispeaker Polish speech database with a phoneme-level language model, enabling the study of how prosody and discourse factors shape acoustic patterns. There also exists a range of BERT-based encoding models for speech tasks using phonetic features (Nguyen et al., 2023; Zhang et al., 2022; Sundararaman et al., 2021). McLaughlin et al. (2025) investigate how Llama 3.2 internally represents phonetic information despite not explicitly training on such data. By probing and using causal intervention techniques, they show that embeddings can encode recoverable phonetic features, identifying what they coin a "phoneme mover head" that can promote phonetic similarity in rhythm tasks. They further find that the model's internal vowel space partially aligns with the IPA vowel quadrilateral, indicating that the model can converge on phonetic representations without explicit training. Furthermore, Sharma et al. (2021) present work on learning phonetic word embeddings. Using X-SAMPA mappings for English (Wells, 1995) and mappings proposed by Malviya et al. (2016) for Hindi, they develop a distance metric based on perceptual similarity that takes into account the salience of different sound changes (such as vowel changes being more salient than consonants). Phonetic representations of words within the resulting embedding space are shown to correctly group perceptually similar words.

Work has also been conducted on how to integrate phonetic knowledge into prompting strategies, alleviating the necessity of model fine-tuning. Lei et al. (2025) present a phonetic retrieval-augmented framework for ASR with LLMs, where entities are selected using normalised phonetic distance and provided as context. By including phonetically similar entities in the prompt, the model is able to re-rank and refine candidate ASR outputs, improving the performance of named entity recognition. Nguyen et al. (2025) examine the performance gap between Latin and non-Latin script languages, demonstrating that even state-of-the-art multilingual models underperform on languages that are not written in the Latin script. By proposing a prompting method that includes phonemic transcriptions, it allows for language-independent representations that would be otherwise obscured by graphemic differences. This method, therefore, improved multilingual generation and translation, thereby reducing the performance disparity between languages with different writing systems.

With the increasing use of LLMs as the basis of phonetic and phonological tasks, it becomes crucial for such systems to be evaluated. Suvarna et al. (2024) introduce PhonologyBench, a benchmark that evaluates LLMs across three different tasks, including G2P, syllable counting, and rhyme generation. Overall, even the strongest models underperformed human baselines, with the largest discrepancy observed in syllable counting. This weakness extends even to simpler orthographic counting tasks such as counting the frequency of <r> in the word "strawberry" (Xu and Ma, 2025). Zouhar et al. (2024) present PWESuite, a benchmark for phonetic word embeddings that evaluates how models capture phonological relationships. Two levels of understanding are evaluated: *intrinsic* aspects, which include word retrieval and correlation with sound similarity and *extrinsic* aspects, which include cognate detection and sound analogies.[4] Experiments show that whilst subword-based embeddings encode some phonetic structure, dedicated phonetic embeddings consistently outperform them across all tasks. Finally, Gao et al. (2024) present SLIQ-LL, a benchmark targeting spoken language learning by evaluating LLMs on conceptual and applied phonological questions. In this dichotomy, conceptual questions refer to things such as "*How many possible places of articulation are there?*", whilst applied questions asked for this knowledge to be used in practice, such as asking "*What syllable carries primary stress in [word]?*". Results demonstrated that LLMs can learn conceptual

[3] A type of phone representation that takes into account coarticulation from preceding and proceeding sounds.

[4] For example, a sound analogy may be that /tɪn/ is to /dɪn/, what /sɪn/ is to /zɪn/, as a result of the change in voicing within the first segment.

knowledge reasonably well but fail to apply it directly to tasks involving pronunciation, stress, or intonation.

3 Downstream Application Tasks

3.1 Poetry

In this section, we define poetry as an umbrella term encompassing traditional poetry such as sonnets (Spiller, 1992), haikus (Yasuda, 1973), and limericks (Vaughn, 1904), alongside the generation of song lyrics that rely on linguistic techniques and areas of phonetics and phonology such as rhyme, metre, alliteration, consonance, and assonance. Poetry and lyrics of these kinds exploit patterns in phonetics and phonology for literary effects such as sound symbolism (Loakman et al., 2024; Ramachandran and Hubbard, 2001; Sapir, 1929), and to meet constraints imposed by the given medium. For example, a haiku traditionally consists of three lines, in which the first line has five syllables, the second has seven, and the third again has five, whilst song lyrics are often designed to complement a musical backing track and are therefore length-restricted.

Owing to the integral nature of characteristics such as rhyme to poetry, most approaches focus on training phoneme-aware models for poetry generation. McGregor et al. (2016) present an early attempt at using a phoneme-based model for the task of poetry generation. They modelled the patterns of phoneme co-occurrence in poetry with an information-theoretic metric that uses phoneme transcriptions and syllable information from the CMUDict. Whilst this model was trained to have phonetic and phonological awareness to meet form constraints, the approach was at the expense of not maintaining coherent semantics. Hopkins and Kiela (2017) present a similar approach, also training a phoneme-level model. Using a combination of CMUDict transcriptions and finite state automata based on the G2P rules presented in Black et al. (1998), they consider the output semantics by incorporating a stage to convert the learnt phoneme sequences back into words using an HMM and n-gram language model, overcoming the semantic incoherence that could be caused by incorrect homophone selection during transliteration. They increase the robustness of their approach to unseen poetry forms by decoupling the content and form aspects of poetry, using a cascade of weighted finite state transducers to identify syllable stress patterns, and therefore rhythm. Ghazvininejad et al. (2016) likewise explore meeting rhythm and rhyme constraints alongside topic-conditioned poetry generation. Despite the efficacy of incorporating linguistic knowledge, some approaches have also forgone the use of phonetically transcribed data, opting instead to train on lines of unlabelled orthographic text annotated with control codes (Keskar et al., 2019) associated with the metre and rhyme scheme, leading to moderate implicit learning (Ormazabal et al., 2022). Regarding rap lyric generation, Nikolov et al. (2020) incorporate an explicit rhyme enhancement step into generation, using BERT to mask the final word in a line and then predicting the 200 most likely replacements, selecting the candidate that maximises the sequence of overlapping vowels with the last word of the previous line. Other approaches to song lyric generation have focused on backwards decoding - predicting the rhyming words for each line before decoding the rest of a given line, therefore prioritising the presence of rhyme (Sun et al., 2023; Liu et al., 2022; Xue et al., 2021; Van de Cruys, 2020).

A common limitation of NLP approaches to processing language such as poetry is the focus on broad subword tokenisation strategies that break up phonetically relevant sequences. One approach to remedying this is with character-level tokenisation strategies. For example, Tikhonov and Yamshchikov (2018) trained both orthographic character-based and phoneme-based LSTMs to generate poetry that imitates the styles of specific well-known English- (e.g., Shakespeare and Poe) and Russian-language (e.g., Pushkin and Esenin) poets - confirming that the phonetic characteristics of the output are integral to correctly mimicking poetic voice. Potash et al. (2015) also train LSTMs for rap lyric generation, incorporating rhyme knowledge via adding special tokens to the start and end of a line or verse to aid the model in implicitly learning that final tokens in different lines are likely to rhyme. Similar to Tikhonov and Yamshchikov (2018), they incorporate additional features to aid in mimicking the styles of different artists. Potash et al. (2018) extends this work into a benchmark for rap lyric ghostwriting. Building upon the abilities of character-based models for phonetic tasks, Belouadi and Eger (2023) train an end-to-end token-free poetry generation model, ByGPT5, using rhyme and metre classifiers alongside a novel dataset of English and German poetry quatrains that was annotated with an alliteration metric.

Finally, as the fields of music generation (Pathariya et al., 2024; Zhu et al., 2023) and singing speech synthesis (Bai et al., 2025; Cho et al., 2021) expand, so too has the generation of combined lyric and melody generation. For example, Watanabe et al. (2018) present a melody-conditioned lyrics model that

exploits syllable counts to ensure alignment between melody and lyrics whilst taking into account the discourse-level semantic structure of lyrics. Further work to this end has been performed (Chen and Lerch, 2020; Sheng et al., 2020), including for Portuguese (Gonçalo Oliveira, 2015; Gonçalo Oliveira et al., 2007), Tamil (Ramakrishnan A et al., 2009), and Chinese (Lee et al., 2019). More recently, Tian et al. (2023) presented an unsupervised approach leveraging automatic alignment between syllables and musical melody. They use a plan-to-lyric approach, using GPT-2 to generate full words before retrieving the word's syllable stress from CMUDict, identifying whether or not it meets the required stress-duration alignment constraint to be included in the final lyrics. Additional research has also focused on the reverse task, generating appropriate melodies for a pre-existing set of lyrics (Yu et al., 2025; Duan et al., 2023).

In recent years, there has been increased work on benchmarking the performance of pre-trained language models in poetry generation. Abdibayev et al. (2021) present BPoMP, a poetry-focused variant of Warstadt et al. (2020)'s minimal-pair specific benchmark, BLiMP. BPoMP tests models' abilities to distinguish between correct and corrupted limericks from the OEDILF project,[5] the latter of which are created by shuffling rhyming words, shuffling entire lines, or replacing an element of a rhyming word pair with a non-rhyming substitute. Of the three tasks, the tested models struggled the most at identifying corrupted limericks when a rhyming word was replaced with a non-rhyming substitute, demonstrating that these pre-trained models did not understand the phonetic characteristics of the texts, focusing instead on the semantics, which remained constant in this condition. Additionally, Walsh et al. (2024) present a benchmark to test LLMs at identifying different poetic forms.

3.2 Humour

Whilst there are myriad forms of humour, features of phonetics and phonology are paramount to the forms of heterographic puns and tongue twisters. Heterographic puns comprise wordplay that exploits homophony and near-homophony (i.e., words that sound similar, yet are spelt differently). For instance, "*What do you call a fake noodle? An impasta*" relies on the phonetic similarity between "pasta" [pastə] and *"-poster"* [pɒstə]. To generate such humour, computational models must be able to identify words that share similar phonetic characteristics, despite having differing orthography. On the other hand, tongue twisters are difficult to articulate due to the rapid succession of phonemes that share similar phonological features (Clements and Ridouane, 2011; De Lacy, 2007; Klausenburger, 1970) which cause confusion when being realised as phones - for example, alternating between /s/ and /ʃ/ in "*She sells sea shells by the sea shore*". Incorporating phonetic and phonological knowledge enables models to construct sequences that exploit these characteristics.

Midway between poetry/lyric generation and humour, tongue twister generation aims to create a grammatically valid sequence of words that features high levels of alliteration, generating humour through articulatory difficulty. Keh et al. (2023) present PANCETTA, a series of approaches involving training naive and phoneme-aware models in topic-to-twister and style-transfer settings. They include phoneme-level awareness by pre-training models on phonemic representations of WikiText data. Other approaches have also been taken to observe whether features of tongue twisters can be learnt implicitly from orthography (Loakman et al., 2025b, 2023). Two phoneme-aware evaluation metrics were introduced in Loakman et al. (2025b) based on phoneme edit distance (McCoy and Frank, 2018; Mortensen et al., 2016), which assessed the similarity of output phonemes in phonological feature space.

Elsewhere in computational humour, early work in pun generation used template-based joke schema that facilitated the easy injection of specialised linguistic knowledge, such as phonetic transcriptions (Hong and Ong, 2009; Ritchie et al., 2007; Binsted, 1996; Binsted and Ritchie, 1994). For example, in order to create heterographic puns, Binsted (1996) and Binsted and Ritchie (1994) made use of a pre-existing corpus of English homophones, substituting parts of a word with semantically different yet phonetically similar elements, such as turning "tranquillizer" into "trunkquillizer" within a joke concerning elephants. Furthermore, Stock and Strapparava (2005) present HAHAcronym, a method for producing acronym-based jokes that leverages CMUDict to identify candidate words that can replace the original whilst maintaining phonetic similarity, such as using "intimidation" in place of "investigation" for the Federal Bureau of Investigation (FBI). Rule-based approaches remain in present works, such as Dehouck and Delaborde (2025), who present work on generating *Monsieur et Madame* puns based on French given names. Their approach involves taking a phonetically transcribed name and mapping it to a French term

[5] https://www.oedilf.com/db/Lim.php

from Wiktionary using a prefix tree. Further concessions are made for the discrepancy in written and spoken French, with phonological distinctions being relaxed in order to generate a larger pool of pun word candidates whilst using less accent-specific matching. More modern approaches to pun generation have relied on neural architectures, combining phonetics (e.g., homophone pairs) with humour mechanisms such as incongruity (Chen et al., 2024; Sun et al., 2022; Tian et al., 2022b; He et al., 2019).

Baluja (2025) observe that prompting LLMs with synthetic speech of puns being read aloud alongside the original text results in an improvement in generated pun explanations. This suggests that exploiting the audio of the heterograph's pronunciation, rather than identifying similar-sounding words through orthography alone, is beneficial. In the case of homographs, this is likely the result of the homonym subset, whereby two words are spelt the same, yet both their meaning and pronunciation differ (e.g., "*I like to read*", /ɹiːd/ versus "*Yesterday, I read the book*", /ɹɛd/). Furthermore, recent work has shown that LLMs still have considerable difficulty with phonetically motivated humour, struggling more to both generate (Xu et al., 2024) and explain (Zangari et al., 2025; Loakman et al., 2025c) heterographic puns than homographic puns.

3.3 Translation

Finally, we turn to the topic of translation. In particular, phonetic and phonological knowledge is incorporated into translation systems that are specifically designed to process texts of the formats previously discussed. In such cases, performing translation such that it can be understood by a target audience requires not only the preservation of meaning, but also phonetic and phonological constraints, such as maintaining the same metrical structure or location of rhyming couplets, which may be achieved through techniques such as constrained decoding (Bogoychev and Chen, 2023; Nakazawa et al., 2021; Susanto et al., 2020; Post and Vilar, 2018).

Early work into the translation of loan words found that including both grapheme- and phoneme-based methods helped mitigate issues that each faced when applied separately. Oh et al. (2006) showed that whilst grapheme-based models could capture regularities in spelling, they often failed when pronunciation diverged from what would be expected through orthography. In such cases, phoneme-based models showed better performance in preserving sound information that was not encoded in graphemes. For example, when the English loan word "data" (/deɪtə/ or /dætə/) is transliterated, in Korean it becomes "de-i-teo" and in Japanese it becomes "de-e-ta", reflecting each language's phonology. Tian et al. (2022a) present a transliteration model for English to Arabic proper nouns that utilises a phonemic memory module to guide the mapping between English phonemes and Arabic graphemes during decoding. This results in outputs being biased towards phonetically similar transliterations, making them more accurate in transliterating names without changing the pronunciation.

Liu et al. (2019) integrate textual and phonetic embeddings to improve robustness to homophones, whilst Mhaskar et al. (2024) present a method that aligns phoneme counts between source and target sentences. With the introduction of phoneme-count compliance, Mhaskar et al. (2024) demonstrate that their approach significantly outperforms baselines for English-Hindi translation. In addition, Rosales Núñez et al. (2019) present a phonetic normalisation approach to improve the translation of noisy, user-generated content in French. They utilise character-based G2P, with the output being used to generate phonemic candidates for translation. This allows for non-standard spellings, accent variations, and typographical errors to be corrected.

Humour translation remains an actively investigated area, having dedicated shared tasks such as CLEF's JOKER (Ermakova et al., 2026). Taylor et al. (2025) present work on translating puns from English to French, presenting the highest scoring submission. To achieve this, they use contrastive learning with phonetic-semantic embeddings learnt from the databases Lexique (New et al., 2004) and PanPhon (Mortensen et al., 2016). Chakrabarty et al. (2021) show that a multilingual fine-tuned poetry model significantly outperforms models 30x larger that were not exposed to poetry during training. Whilst the results focus on semantics, the findings also suggest links to phonetics and phonology due to poetry frequently relying on rhyme. Similar links have been identified in the processing of Sanskrit and Persian (Sandhan et al., 2025; Khanmohammadi et al., 2023).

4 Challenges & Future Directions

Linguistic Bias. As a research domain, a pressing concern remains the inadvertent promotion of specific linguistic characteristics over others (Fleisig et al., 2024; Ryan et al., 2024; Joshi et al., 2020). Whilst accent robustness is a common concern in areas such as automatic speech recognition (Yong et al., 2025;

Markl and McNulty, 2022), much research in phonetically and phonologically grounded language generation fails to address these concerns. For instance, most G2P systems are trained to represent high-resource, idealised language varieties (e.g., Southern Standard British English or General American), and gold-standard resources such as the CMU Pronouncing Dictionary remain static. As a result, the understanding of phonetically and phonologically informed language overlooks differences in phoneme inventories and articulatory tendencies of non-standard, non-prestige varieties, resulting in language technologies that do not adequately serve such communities. Outside of accent variation, similar representational limitations are observed across different languages. Whilst English is known for having opaque orthography, the sheer volume of available data (including data relating to phonetic and phonological characteristics) facilitates implicit learning to a much greater extent than would be expected from a language that is both low-resource and orthographically opaque, in contrast to languages with transparent orthography and lower levels of inter-speaker variation (Marjou, 2021; Borleffs et al., 2017). To this end, research must continue in the direction of the wider NLP trend towards customisable, personalised models that work for a language community or individual, rather than against (Guan et al., 2025).

Tokenisation Approaches. Whilst existing research has demonstrated that phonetic and phonological knowledge can be implicitly gained by LLMs with significant scale increases in computational power and data, lower resource application settings require more deft approaches. Of these, the limits of subword tokenisation for many of the tasks presented in this work are evident, with the broadness of tokenisation resulting in a loss of understanding concerning the relationship between orthography and pronunciation. Several works have found success in applying alternative tokenisation approaches, such as character-level tokenisers as used in ByT5 (Loakman et al., 2025b; Rahmati and Sameti, 2024; Zhu et al., 2022).

Knowledge Integration. Current foundation models demonstrate remarkable understanding of a wide range of complex topics. However, despite information on phonetics, phonology, and orthography being widely available for high-resource languages, many of the processing difficulties observed in LLMs indicate an inability to integrate this knowledge in practice (Shojaee et al., 2025; Loakman et al., 2025a; Jiang et al., 2024). Consequently, research toward active reasoning using latent knowledge remains fruitful.

Audio-Enabled Models. Finally, a pivotal future direction concerns LLMs explicitly trained on speech representations, such as audio-enabled LLMs or speech-to-speech understanding models (Cui et al., 2025; Wang et al., 2025; Yang et al., 2024; Shan et al., 2025). Whilst such models remain in their relative infancy, these approaches alleviate the need for labour-intensive handcrafted phonetic and phonological knowledge, such as gold-standard transcriptions, whilst additionally avoiding the lack of representational fidelity that is inherent in learning sound-based characteristics from broad phonemic transcriptions or directly from orthography. However, accent diversity concerns remain with such models, with many of the most popular speech datasets being relatively homogenous in language variety (Markl and McNulty, 2022). Multimodal prompting approaches demonstrate promise in facilitating phonetic and phonological understanding, lending support to future advancements as multimodal fusion improves and the handling of speech by LLMs transitions from speech recognition front-ends for text-based understanding models to speech-to-speech models.

5 Conclusion

In this survey we have outlined a range of key foundational and application-based areas where incorporating knowledge of phonetics and phonology is either integral (e.g., heterographic puns) or of substantial benefit (e.g., spelling error correction). We have focussed primarily on unique use cases of phonetic/phonological knowledge in these domains, aiming to provide the reader and the NLP/CL communities as a whole, a concise and accessible introduction to a subfield of NLG where linguistic knowledge of this type is paramount.

Limitations

We aim for this work to act as an overview of the ways in which the NLG community has incorporated knowledge from the areas of phonetics and phonology into different tasks. As a result, this survey does not aim for high levels of depth in any specific subdiscipline, instead we refer the reader to relevant recent works in the respective areas of interest for deeper architectural insight (especially where these details do not pertain to the incorporation of phonetic/phonological knowledge itself). The purpose of this survey is to act as an overview for works incorporating knowledge from a specific branch of linguistics which is less frequently considered in common generation tasks.

Although we have strived to include a wide range of works in the target areas we explore, it remains possible that we have missed some works in the area and this survey is not to be taken as a definitive list of all of the research in NLG that has incorporated phonetic/phonological knowledge, including within the subdisciplines we cover.

Acknowledgments

Tyler Loakman and Joseph James were supported by the Centre for Doctoral Training in Speech and Language Technologies (SLT) and their Applications funded by UK Research and Innovation [grant number EP/S023062/1].

References

Almas Abdibayev, Allen Riddell, and Daniel Rockmore. 2021. BPoMP: The benchmark of poetic minimal pairs – limericks, rhyme, and narrative coherence. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1–9, Held Online. INCOMA Ltd.

Miriam Amin and Manuel Burghardt. 2020. A survey on approaches to computational humor generation. In *Proceedings of the 4th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 29–41, Online. International Committee on Computational Linguistics.

Angelina Aquino, Joshua Lijandro Tsang, Crisron Rudolf Lucas, and Franz de Leon. 2019. G2p and asr techniques for low-resource phonetic transcription of tagalog, cebuano, and hiligaynon. In *2019 International Symposium on Multimedia and Communication Technology (ISMAC)*, pages 1–5.

Peng Bai, Meizhen Zheng, and Xiaodong Shi. 2025. A survey of singing voice synthesis. In *Proceedings of the 10th Conference on Sound and Music Technology*, pages 19–30, Singapore. Springer Nature Singapore.

Ashwin Baluja. 2025. Text is not all you need: Multimodal prompting helps LLMs understand humor. In *Proceedings of the 1st Workshop on Computational Humor (CHum)*, pages 9–17, Online. Association for Computational Linguistics.

Jonas Belouadi and Steffen Eger. 2023. ByGPT5: End-to-end style-conditioned poetry generation with token-free language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7364–7381, Toronto, Canada. Association for Computational Linguistics.

Abbas Bertina, Shahab Beirami, Hossein Biniazian, Elham Esmaeilnia, Soheil Shahi, and Mahdi Pirnia. 2025. Bridging the gap: An intermediate language for enhanced and cost-effective grapheme-to-phoneme conversion with homographs with multiple pronunciations disambiguation. *arXiv preprint arXiv:2505.06599*.

Kim Binsted. 1996. *Machine humour : an implemented model of puns*. Ph.D. thesis, University of Edinburgh.

Kim Binsted and Graeme Ritchie. 1994. An implemented model of punning riddles. In *Proceedings of the Twelfth AAAI National Conference on Artificial Intelligence*, AAAI'94, page 633–638. AAAI Press.

Alan W Black, Kevin Lenzo, and Vincent Pagel. 1998. Issues in building general letter to sound rules. In *3rd ESCA/COCOSDA Workshop on Speech Synthesis (SSW 3)*, pages 77–80.

Nikolay Bogoychev and Pinzhen Chen. 2023. Terminology-aware translation with constrained decoding and large language model prompting. In *Proceedings of the Eighth Conference on Machine Translation*, pages 890–896, Singapore. Association for Computational Linguistics.

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

Bastian Bunzeck, Daniel Duran, Leonie Schade, and Sina Zarrieß. 2024. Graphemes vs. phonemes: battling it out in character-based language models. In *The 2nd BabyLM Challenge at the 28th Conference on Computational Natural Language Learning*, pages 54–64, Miami, FL, USA. Association for Computational Linguistics.

Tuhin Chakrabarty, Arkadiy Saakyan, and Smaranda Muresan. 2021. Don't go far off: An empirical study on neural poetry translation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7253–7265, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tyler A. Chang and Benjamin K. Bergen. 2024. Language model behavior: A comprehensive survey. *Computational Linguistics*, 50(1):293–350.

Yang Chen, Chong Yang, Tu Hu, Xinhao Chen, Man Lan, Li Cai, Xinlin Zhuang, Xuan Lin, Xin Lu, and Aimin Zhou. 2024. Are U a joke master? pun generation via multi-stage curriculum learning towards a humor LLM. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 878–890, Bangkok, Thailand. Association for Computational Linguistics.

Yihao Chen and Alexander Lerch. 2020. Melody-Conditioned Lyrics Generation with SeqGANs . In *2020 IEEE International Symposium on Multimedia (ISM)*, pages 189–196, Los Alamitos, CA, USA. IEEE Computer Society.

Shiyang Cheng, Pengcheng Zhu, Jueting Liu, and Zehua Wang. 2024. A survey of grapheme-to-phoneme conversion methods. *Applied Sciences*, 14(24):11790.

Won Ik Cho, Seok Min Kim, and Nam Soo Kim. 2020. Towards an efficient code-mixed grapheme-to-phoneme conversion in an agglutinative language: A case study on to-Korean transliteration. In *Proceedings of the 4th Workshop on Computational Approaches to Code Switching*, pages 65–70, Marseille, France. European Language Resources Association.

Yin-Ping Cho, Fu-Rong Yang, Yung-Chuan Chang, Ching-Ting Cheng, Xiao-Han Wang, and Yi-Wen Liu. 2021. A Survey on Recent Deep Learning-driven Singing Voice Synthesis Systems . In *2021 IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR)*, pages 319–323, Los Alamitos, CA, USA. IEEE Computer Society.

G. Nick Clements and Rachid Ridouane. 2011. *Where Do Phonological Features Come From? : Cognitive, Physical and Developmental Bases of Distinctive Speech Categories.*, 1st ed. edition. Language Faculty and Beyond Series. John Benjamins Publishing Company, Amsterdam/Philadelphia.

Wenqian Cui, Dianzhi Yu, Xiaoqi Jiao, Ziqiao Meng, Guangyan Zhang, Qichao Wang, Steven Y. Guo, and Irwin King. 2025. Recent advances in speech language models: A survey. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13943–13970, Vienna, Austria. Association for Computational Linguistics.

Paul V. De Lacy. 2007. *The Cambridge handbook of phonology*. Cambridge University Press, Cambridge.

Mathieu Dehouck and Marine Delaborde. 2025. Rule-based approaches to the automatic generation of puns based on given names in French. In *Proceedings of the 1st Workshop on Computational Humor (CHum)*, pages 18–22, Online. Association for Computational Linguistics.

Aliya Deri and Kevin Knight. 2016. Grapheme-to-phoneme models for (almost) any language. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 399–408, Berlin, Germany. Association for Computational Linguistics.

Wei Duan, Yi Yu, Xulong Zhang, Suhua Tang, Wei Li, and Keizo Oyama. 2023. Melody generation from lyrics with local interpretability. *ACM Trans. Multimedia Comput. Commun. Appl.*, 19(3).

Liana Ermakova, Ricardo Campos, Anne-Gwenn Bosser, and Tristan Miller. 2026. Overview of the clef 2025 joker lab: Humour in machine. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 315–337, Cham. Springer Nature Switzerland.

Eve Fleisig, Genevieve Smith, Madeline Bossi, Ishita Rustagi, Xavier Yin, and Dan Klein. 2024. Linguistic bias in ChatGPT: Language models reinforce dialect discrimination. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13541–13564, Miami, Florida, USA. Association for Computational Linguistics.

Robert C. Gale, Alexandra C. Salem, Gerasimos Fergadiotis, and Steven Bedrick. 2023. Mixed orthographic/phonemic language modeling: Beyond orthographically restricted transformers (BORT). In *Proceedings of the 8th Workshop on Representation Learning for NLP (RepL4NLP 2023)*, pages 212–225, Toronto, Canada. Association for Computational Linguistics.

Yingming Gao, Baorian Nuchged, Ya Li, and Linkai Peng. 2024. An investigation of applying large language models to spoken language learning. *Applied Sciences*, 14(1).

Marjan Ghazvininejad, Xing Shi, Yejin Choi, and Kevin Knight. 2016. Generating topical poetry. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1183–1191, Austin, Texas. Association for Computational Linguistics.

Hugo Gonçalo Oliveira. 2017. A survey on intelligent poetry generation: Languages, features, techniques, reutilisation and evaluation. In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 11–20, Santiago de Compostela, Spain. Association for Computational Linguistics.

Hugo Gonçalo Oliveira. 2015. Tra-la-lyrics 2.0: Automatic generation of song lyrics on a semantic domain. *Journal of Artificial General Intelligence*, 6(1):87–110.

Hugo Gonçalo Oliveira, F Cardoso, Amialcar Cardoso, and Francisco C. Pereira. 2007. Tra-la-lyrics: An approach to generate text based on rhythm. *Proceedings of the 4th International Joint Workshop on Computational Creativity*, pages 47–55.

Zébulon Goriely, Richard Diehl Martinez, Andrew Caines, Paula Buttery, and Lisa Beinborn. 2024. From babble to words: Pre-training language models on continuous streams of phonemes. In *The 2nd BabyLM Challenge at the 28th Conference on Computational Natural Language Learning*, pages 37–53, Miami, FL, USA. Association for Computational Linguistics.

Jian Guan, Junfei Wu, Jia-Nan Li, Chuanqi Cheng, and Wei Wu. 2025. A survey on personalized Alignment—The missing piece for large language models in real-world applications. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 5313–5333, Vienna, Austria. Association for Computational Linguistics.

Lorenz Gutscher and Michael Pucher. 2024. Exploring phonetic features in language embeddings for unseen language varieties of Austrian German. In *Proceedings of the 20th Conference on Natural Language Processing (KONVENS 2024)*, pages 317–325, Vienna, Austria. Association for Computational Linguistics.

Michael Hammond. 2023. Low-resource grapheme-to-phoneme mapping with phonetically-conditioned transfer. In *Proceedings of the 20th SIGMORPHON workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 245–248.

Dongrui Han, Mingyu Cui, Jiawen Kang, Xixin Wu, Xunying Liu, and Helen Meng. 2024. Improving grapheme-to-phoneme conversion through in-context knowledge retrieval with large language models. In *2024 IEEE 14th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, pages 631–635. IEEE.

He He, Nanyun Peng, and Percy Liang. 2019. Pun generation with surprise. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1734–1744, Minneapolis, Minnesota. Association for Computational Linguistics.

Thomas R. Hofmann. 1993. *Realms of meaning : an introduction to semantics*. Learning about language. Longman, London.

Bryan Anthony Hong and Ethel Ong. 2009. Automatically extracting word relationships as templates for pun generation. In *Proceedings of the Workshop on Computational Approaches to Linguistic Creativity*, pages 24–31, Boulder, Colorado. Association for Computational Linguistics.

Jack Hopkins and Douwe Kiela. 2017. Automatically generating rhythmic verse with neural networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 168–178, Vancouver, Canada. Association for Computational Linguistics.

Ruili Jiang, Kehai Chen, Xuefeng Bai, Zhixuan He, Juntao Li, Muyun Yang, Tiejun Zhao, Liqiang Nie, and Min Zhang. 2024. A survey on human preference learning for large language models.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Sedrick Scott Keh, Steven Y. Feng, Varun Gangal, Malihe Alikhani, and Eduard Hovy. 2023. PANCETTA: Phoneme aware neural completion to elicit tongue twisters automatically. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 491–504, Dubrovnik, Croatia. Association for Computational Linguistics.

Nitish Shirish Keskar, Bryan McCann, Lav R. Varshney, Caiming Xiong, and Richard Socher. 2019. Ctrl: A conditional transformer language model for controllable generation.

Reza Khanmohammadi, Mitra Sadat Mirshafiee, Yazdan Rezaee Jouryabi, and Seyed Abolghasem Mirroshandel. 2023. Prose2poem: the blessing of transformers in translating prose to persian poetry. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(6):1–18.

Hwa-Yeon Kim, Jong-Hwan Kim, and Jae-Min Kim. 2022. Fast bilingual grapheme-to-phoneme conversion. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Track*, pages 289–296.

Jürgen Klausenburger. 1970. *French Prosodics and Phonotactics : An Historical Typology.*, 1st ed. edition. Beihefte Zur Zeitschrift Für Romanische Philologie Series. Walter de Gruyter GmbH, Tübingen.

Paul Kroeger. 2018. *Analyzing meaning: An introduction to semantics and pragmatics*. Textbooks in Language Sciences. Language Science Press, Place of publication not identified.

Hsin-Pei Lee, Jhih-Sheng Fang, and Wei-Yun Ma. 2019. iComposer: An automatic songwriting system for Chinese popular music. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 84–88, Minneapolis, Minnesota. Association for Computational Linguistics.

Zhihong Lei, Xingyu Na, Mingbin Xu, Ernest Pusateri, Christophe Van Gysel, Yuanyuan Zhang, Shiyi Han, and Zhen Huang. 2025. Contextualization of asr with llm using phonetic retrieval-based augmentation. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Xinjian Li, Florian Metze, David R Mortensen, Shinji Watanabe, and Alan W Black. 2022. Zero-shot learning for grapheme to phoneme conversion with language ensemble. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2106–2115.

Hairong Liu, Mingbo Ma, Liang Huang, Hao Xiong, and Zhongjun He. 2019. Robust neural machine translation with joint textual and phonetic embedding. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3044–3049, Florence, Italy. Association for Computational Linguistics.

Nayu Liu, Wenjing Han, Guangcan Liu, Da Peng, Ran Zhang, Xiaorui Wang, and Huabin Ruan. 2022. ChipSong: A controllable lyric generation system for Chinese popular song. In *Proceedings of the First Workshop on Intelligent and Interactive Writing Assistants (In2Writing 2022)*, pages 85–95, Dublin, Ireland. Association for Computational Linguistics.

Tyler Loakman, Joseph James, and Chenghua Lin. 2025a. Seeing isn't hearing: Benchmarking vision language models at interpreting spectrograms. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 79–86, Mumbai, India. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.

Tyler Loakman, Yucheng Li, and Chenghua Lin. 2024. With ears to see and eyes to hear: Sound symbolism

experiments with multimodal large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 2849–2867, Miami, Florida, USA. Association for Computational Linguistics.

Tyler Loakman, Chen Tang, and Chenghua Lin. 2023. TwistList: Resources and baselines for tongue twister generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 579–589, Toronto, Canada. Association for Computational Linguistics.

Tyler Loakman, Chen Tang, and Chenghua Lin. 2025b. Train and constrain: Phonologically informed tongue twister generation from topics and paraphrases. *Computational Linguistics*, 51(2):415–466.

Tyler Loakman, William Thorne, and Chenghua Lin. 2025c. Comparing apples to oranges: A dataset & analysis of LLM humour understanding from traditional puns to topical jokes. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 9502–9518, Suzhou, China. Association for Computational Linguistics.

Tyler Loakman, William Thorne, and Chenghua Lin. 2025d. Who's laughing now? an overview of computational humour generation and explanation. In *Proceedings of the 18th International Natural Language Generation Conference*, pages 780–794, Hanoi, Vietnam. Association for Computational Linguistics.

Zofia Malisz, Jan Foremski, and Małgorzata Kul. 2024. PRODIS - a speech database and a phoneme-based language model for the study of predictability effects in Polish. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 13068–13073, Torino, Italia. ELRA and ICCL.

Shrikant Malviya, Rohit Mishra, and Uma Shanker Tiwary. 2016. Structural analysis of hindi phonetics and a method for extraction of phonetically rich sentences from a very large hindi text corpus. In *2016 Conference of The Oriental Chapter of International Committee for Coordination and Standardization of Speech Databases and Assessment Techniques (O-COCOSDA)*, pages 188–193.

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

Nina Markl and Stephen Joseph McNulty. 2022. Language technology practitioners as language managers: arbitrating data bias and predictive bias in ASR. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6328–6339, Marseille, France. European Language Resources Association.

Ryo Masumura, Naoki Makishima, Mana Ihori, Akihiko Takashima, Tomohiro Tanaka, and Shota Orihashi. 2020. Phoneme-to-grapheme conversion based large-scale pre-training for end-to-end automatic speech recognition. In *Interspeech 2020*, pages 2822–2826.

Yuhi Matogawa, Yusuke Sakai, Taro Watanabe, and Chihiro Taguchi. 2024. Japanese rule-based grapheme-to-phoneme conversion system and multilingual named entity dataset with international phonetic alphabet. In *Proceedings of the 21st SIGMORPHON workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 77–86.

Richard T. McCoy and Robert Frank. 2018. Phonologically informed edit distance algorithms for word alignment with low-resource languages. In *Proceedings of the Society for Computation in Linguistics (SCiL) 2018*, pages 102–112.

Stephen McGregor, Matthew Purver, and Geraint Wiggins. 2016. Process based evaluation of computer generated poetry. In *Proceedings of the INLG 2016 Workshop on Computational Creativity in Natural Language Generation*, pages 51–60, Edinburgh, UK. Association for Computational Linguistics.

Oliver McLaughlin, Jack Merullo, and Arjun Khurana. 2025. I have no mouth, and i must rhyme: Uncovering internal phonetic representations in llama 3.2. In *ICML 2025 Workshop on Assessing World Models*.

Shivam Mhaskar, Nirmesh Shah, Mohammadi Zaki, Ashishkumar Gudmalwar, Pankaj Wasnik, and Rajiv Shah. 2024. Isometric neural machine translation using phoneme count ratio reward-based reinforcement learning. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3966–3976, Mexico City, Mexico. Association for Computational Linguistics.

E.A. Moravcsik. 2006. *An Introduction to Syntax: Fundamentals of Syntactic Analysis*. Bloomsbury Academic.

David R. Mortensen, Patrick Littell, Akash Bharadwaj, Kartik Goyal, Chris Dyer, and Lori Levin. 2016. PanPhon: A resource for mapping IPA segments to articulatory feature vectors. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3475–3484, Osaka, Japan. The COLING 2016 Organizing Committee.

Toshiaki Nakazawa, Hideki Nakayama, Chenchen Ding, Raj Dabre, Shohei Higashiyama, Hideya Mino, Isao Goto, Win Pa Pa, Anoop Kunchukuttan, Shantipriya Parida, Ondřej Bojar, Chenhui Chu, Akiko Eriguchi, Kaori Abe, Yusuke Oda, and Sadao Kurohashi. 2021. Overview of the 8th workshop on Asian translation. In *Proceedings of the 8th Workshop on Asian Translation (WAT2021)*, pages 1–45, Online. Association for Computational Linguistics.

Boris New, Christophe Pallier, Marc Brysbaert, and Ludovic Ferrand. 2004. Lexique 2 : A new french lexical database. *Behavior Research Methods, Instruments, & Computers*, 36(3):516–524.

Hoang H Nguyen, Khyati Mahajan, Vikas Yadav, Julian Salazar, Philip S. Yu, Masoud Hashemi, and Rishabh

Maheshwary. 2025. Prompting with phonemes: Enhancing LLMs' multilinguality for non-Latin script languages. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 11975–11994, Albuquerque, New Mexico. Association for Computational Linguistics.

Linh The Nguyen, Thinh Pham, and Dat Quoc Nguyen. 2023. XPhoneBERT: A Pre-trained Multilingual Model for Phoneme Representations for Text-to-Speech. In *Proceedings of the 24th Annual Conference of the International Speech Communication Association (INTERSPEECH)*.

Nikola I. Nikolov, Eric Malmi, Curtis Northcutt, and Loreto Parisi. 2020. Rapformer: Conditional rap lyrics generation with denoising autoencoders. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 360–373, Dublin, Ireland. Association for Computational Linguistics.

J Oh, K Choi, and Hitoshi Isahara. 2006. A comparison of different machine transliteration models. *Journal of Artificial Intelligence Research*, 27:119–151.

Aitor Ormazabal, Mikel Artetxe, Manex Agirrezabal, Aitor Soroa, and Eneko Agirre. 2022. PoeLM: A meter- and rhyme-controllable language model for unsupervised poetry generation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3655–3670, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Mohammed Johar Pathariya, Pratyush Basavraj Jalkote, Aniket Maharudra Patil, Abhishek Ashok Sutar, and Rajashree L. Ghule. 2024. Tunes by technology: A comprehensive survey of music generation models. In *2024 International Conference on Cognitive Robotics and Intelligent Systems (ICC - ROBINS)*, pages 506–512.

Artem Ploujnikov and Mirco Ravanelli. 2022. Soundchoice: Grapheme-to-phoneme models with semantic disambiguation. In *Interspeech 2022*, pages 486–490.

Matt Post and David Vilar. 2018. Fast lexically constrained decoding with dynamic beam allocation for neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1314–1324, New Orleans, Louisiana. Association for Computational Linguistics.

Peter Potash, Alexey Romanov, and Anna Rumshisky. 2015. GhostWriter: Using an LSTM for automatic rap lyric generation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1919–1924, Lisbon, Portugal. Association for Computational Linguistics.

Peter Potash, Alexey Romanov, and Anna Rumshisky. 2018. Evaluating creative language generation: The case of rap lyric ghostwriting. In *Proceedings of the Second Workshop on Stylistic Variation*, pages 29–38, New Orleans. Association for Computational Linguistics.

Mahta Fetrat Qharabagh, Zahra Dehghanian, and Hamid R Rabiee. 2025. Llm-powered grapheme-to-phoneme conversion: Benchmark and case study. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Elnaz Rahmati and Hossein Sameti. 2024. GE2PE: Persian end-to-end grapheme-to-phoneme conversion. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3426–3436, Miami, Florida, USA. Association for Computational Linguistics.

V S Ramachandran and Edward M Hubbard. 2001. Synaesthesia — a window into perception, thought and language. *Journal of Consciousness Studies*, 8(12):3–34.

Ananth Ramakrishnan A, Sankar Kuppan, and Sobha Lalitha Devi. 2009. Automatic generation of Tamil lyrics for melodies. In *Proceedings of the Workshop on Computational Approaches to Linguistic Creativity*, pages 40–46, Boulder, Colorado. Association for Computational Linguistics.

Sam Ribeiro, Giulia Comini, and Jaime Lorenzo-Trueba. 2023. Improving grapheme-to-phoneme conversion by learning pronunciations from speech recordings. In *Interspeech 2023*, pages 999–1003.

Graeme D Ritchie, Ruli Manurung, Helen Pain, Annalu Waller, Rolf Black, and Dave O'Mara. 2007. A practical application of computational humour. In *Proceedings of the Fourth International Joint Conference on Computational Creativity (Goldsmith's, London)*, pages 91–98.

Henry Rogers. 2013. *The sounds of language : an introduction to phonetics*, 1st ed. edition. Learning about Language. Routledge, London ;.

José Carlos Rosales Núñez, Djamé Seddah, and Guillaume Wisniewski. 2019. Phonetic normalization for machine translation of user generated content. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 407–416, Hong Kong, China. Association for Computational Linguistics.

James Route, Steven Hillis, Isak Czeresnia Etinger, Han Zhang, and Alan W Black. 2019. Multimodal, multilingual grapheme-to-phoneme conversion for low-resource languages. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*, pages 192–201, Hong Kong, China. Association for Computational Linguistics.

Michael J Ryan, William Held, and Diyi Yang. 2024. Unintended impacts of LLM alignment on global representation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16121–16140, Bangkok, Thailand. Association for Computational Linguistics.

Jivnesh Sandhan, Amruta Barbadikar, Malay Maity, Pavankumar Satuluri, Tushar Sandhan, Ravi M Gupta, Pawan Goyal, and Laxmidhar Behera. 2025. Aesthetics of Sanskrit poetry from the perspective of computational linguistics: A case study analysis on śikṣāṣṭaka. In

Computational Sanskrit and Digital Humanities - World Sanskrit Conference 2025, pages 15–36, Kathmandu, Nepal. Association for Computational Linguistics.

Edward Sapir. 1929. A study in phonetic symbolism. *Journal of Experimental Psychology*, 12(3):225–239.

Weiqiao Shan, Yuang Li, Yuhao Zhang, Yingfeng Luo, Chen Xu, Xiaofeng Zhao, Long Meng, Yunfei Lu, Min Zhang, Hao Yang, Tong Xiao, and JingBo Zhu. 2025. Enhancing speech large language models with prompt-aware mixture of audio encoders. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 19305–19320, Suzhou, China. Association for Computational Linguistics.

Rahul Sharma, Kunal Dhawan, and Balakrishna Pailla. 2021. Phonetic word embeddings.

Zhonghao Sheng, Kaitao Song, Xu Tan, Yi Ren, Wei Ye, Shikun Zhang, and Tao Qin. 2020. Songmass: Automatic song writing with pre-training and alignment constraint. In *AAAI Conference on Artificial Intelligence*.

Parshin Shojaee, Iman Mirzadeh, Keivan Alizadeh, Maxwell Horton, Samy Bengio, and Mehrdad Farajtabar. 2025. The illusion of thinking: Understanding the strengths and limitations of reasoning models via the lens of problem complexity. In *NeurIPS*.

M. R. G. Spiller. 1992. *The Development of the Sonnet: An Introduction*. Routledge.

Oliviero Stock and Carlo Strapparava. 2005. HAHAcronym: A computational humor system. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 113–116, Ann Arbor, Michigan. Association for Computational Linguistics.

Jiao Sun, Anjali Narayan-Chen, Shereen Oraby, Shuyang Gao, Tagyoung Chung, Jing Huang, Yang Liu, and Nanyun Peng. 2022. Context-situated pun generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4635–4648, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yusen Sun, Liangyou Li, Qun Liu, and Dit-Yan Yeung. 2023. SongRewriter: A Chinese song rewriting system with controllable content and rhyme scheme. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12863–12880, Toronto, Canada. Association for Computational Linguistics.

Mukuntha Narayanan Sundararaman, Ayush Kumar, and Jithendra Vepa. 2021. Phonemebert: Joint language modelling of phoneme sequence and asr transcript. In *Interspeech 2021*, pages 3236–3240.

Raymond Hendy Susanto, Shamil Chollampatt, and Liling Tan. 2020. Lexically constrained neural machine translation with Levenshtein transformer. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3536–3543, Online. Association for Computational Linguistics.

Ashima Suvarna, Harshita Khandelwal, and Nanyun Peng. 2024. PhonologyBench: Evaluating phonological skills of large language models. In *Proceedings of the 1st Workshop on Towards Knowledgeable Language Models (KnowLLM 2024)*, pages 1–14, Bangkok, Thailand. Association for Computational Linguistics.

Russell Taylor, Benjamin Herbert, and Michael Sana. 2025. Pun intended: Multi-agent translation of wordplay with contrastive learning and phonetic-semantic embeddings.

Yuanhe Tian, Renze Lou, Xiangyu Pang, Lianxi Wang, Shengyi Jiang, and Yan Song. 2022a. Improving English-Arabic transliteration with phonemic memories. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3262–3272, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yufei Tian, Anjali Narayan-Chen, Shereen Oraby, Alessandra Cervone, Gunnar Sigurdsson, Chenyang Tao, Wenbo Zhao, Yiwen Chen, Tagyoung Chung, Jing Huang, and Nanyun Peng. 2023. Unsupervised melody-to-lyrics generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9235–9254, Toronto, Canada. Association for Computational Linguistics.

Yufei Tian, Divyanshu Sheth, and Nanyun Peng. 2022b. A unified framework for pun generation with humor principles. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3253–3261, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Aleksey Tikhonov and Ivan P. Yamshchikov. 2018. Sounds wilde. phonetically extended embeddings for author-stylized poetry generation. In *Proceedings of the Fifteenth Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 117–124, Brussels, Belgium. Association for Computational Linguistics.

Yulia Tsvetkov, Sunayana Sitaram, Manaal Faruqui, Guillaume Lample, Patrick Littell, David Mortensen, Alan W Black, Lori Levin, and Chris Dyer. 2016. Polyglot neural language models: A case study in cross-lingual phonetic representation learning. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1357–1366, San Diego, California. Association for Computational Linguistics.

Tim Van de Cruys. 2020. Automatic poetry generation from prosaic text. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2471–2480, Online. Association for Computational Linguistics.

Robert D. Van Valin. 2001. *An introduction to syntax*, first edition. edition. Cambridge University Press, Cambridge.

S. Vaughn. 1904. *Limerick Lyrics: A Collection of Over 700 Choice Versifications, to which is Added a Number*

of Short Verses, from Many Sources. T.J. Carey & Company.

Melanie Walsh, Anna Preus, and Maria Antoniak. 2024. Sonnet or not, bot? poetry evaluation for large models and datasets. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15568–15603, Miami, Florida, USA. Association for Computational Linguistics.

Bin Wang, Xunlong Zou, Geyu Lin, Shuo Sun, Zhuohan Liu, Wenyu Zhang, Zhengyuan Liu, AiTi Aw, and Nancy F. Chen. 2025. AudioBench: A universal benchmark for audio large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4297–4316, Albuquerque, New Mexico. Association for Computational Linguistics.

Chunfeng Wang, Peisong Huang, Yuxiang Zou, Haoyu Zhang, Shichao Liu, Xiang Yin, and Zejun Ma. 2023. Liteg2p: A fast, light and high accuracy model for grapheme-to-phoneme conversion. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. BLiMP: A benchmark of linguistic minimal pairs for English. In *Proceedings of the Society for Computation in Linguistics 2020*, pages 409–410, New York, New York. Association for Computational Linguistics.

Kento Watanabe and Masataka Goto. 2020. Lyrics information processing: Analysis, generation, and applications. In *Proceedings of the 1st Workshop on NLP for Music and Audio (NLP4MusA)*, pages 6–12, Online. Association for Computational Linguistics.

Kento Watanabe, Yuichiroh Matsubayashi, Satoru Fukayama, Masataka Goto, Kentaro Inui, and Tomoyasu Nakano. 2018. A melody-conditioned lyrics language model. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 163–172, New Orleans, Louisiana. Association for Computational Linguistics.

J. C. Wells. 1982a. *Accents of English: Beyond the British Isles*, volume 3 of *Accents of English*. Cambridge University Press, Cambridge.

J. C. Wells. 1982b. *Accents of English: The British Isles*, volume 2 of *Accents of English*. Cambridge University Press, Cambridge.

John C. Wells. 1995. Computer-coding the ipa: a proposed extension of sampa.

Nan Xu and Xuezhe Ma. 2025. LLM the genius paradox: A linguistic and math expert's struggle with simple word-based counting problems. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3344–3370, Albuquerque, New Mexico. Association for Computational Linguistics.

Zhijun Xu, Siyu Yuan, Lingjie Chen, and Deqing Yang. 2024. "a good pun is its own reword": Can large language models understand puns? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11766–11782, Miami, Florida, USA. Association for Computational Linguistics.

Lanqing Xue, Kaitao Song, Duocai Wu, Xu Tan, Nevin L. Zhang, Tao Qin, Wei-Qiang Zhang, and Tie-Yan Liu. 2021. DeepRapper: Neural rap generation with rhyme and rhythm modeling. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 69–81, Online. Association for Computational Linguistics.

Qian Yang, Jin Xu, Wenrui Liu, Yunfei Chu, Ziyue Jiang, Xiaohuan Zhou, Yichong Leng, Yuanjun Lv, Zhou Zhao, Chang Zhou, and Jingren Zhou. 2024. AIR-bench: Benchmarking large audio-language models via generative comprehension. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1979–1998, Bangkok, Thailand. Association for Computational Linguistics.

Kenneth Yasuda. 1973. *The Japanese haiku : its essential nature, history, and possibilities in English, with selected examples*, tut book ed.. edition. Charles E. Tuttle, Rutland, Vermont.

Zheng Xin Yong, Vineel Pratap, Michael Auli, and Jean Maillard. 2025. Effects of Speaker Count, Duration, and Accent Diversity on Zero-Shot Accent Robustness in Low-Resource ASR. In *Interspeech 2025*, pages 1148–1152.

Jiaxing Yu, Xinda Wu, Yunfei Xu, Tieyao Zhang, Songruoyao Wu, Le Ma, and Kejun Zhang. 2025. Songglm: lyric-to-melody generation with 2d alignment encoding and multi-task pre-training. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'25/IAAI'25/EAAI'25. AAAI Press.

Alessandro Zangari, Matteo Marcuzzo, Andrea Albarelli, Mohammad Taher Pilehvar, and Jose Camacho-Collados. 2025. Pun unintended: LLMs and the illusion of humor understanding. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 27936–27971, Suzhou, China. Association for Computational Linguistics.

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. In *Interspeech 2022*, pages 456–460.

Jian Zhu, Cong Zhang, and David Jurgens. 2022. Byt5 model for massively multilingual grapheme-to-phoneme conversion. In *Interspeech 2022*, pages 446–450.

Yueyue Zhu, Jared Baca, Banafsheh Rekabdar, and Reza Rawassizadeh. 2023. A survey of ai music generation tools and models.

Vilém Zouhar, Kalvin Chang, Chenxuan Cui, Nate B. Carlson, Nathaniel Romney Robinson, Mrinmaya Sachan, and David R. Mortensen. 2024. PWESuite: Phonetic word embeddings and tasks they facilitate. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 13344–13355, Torino, Italia. ELRA and ICCL.

A Survey Inclusion Criteria

The following seed terms were used to determine paper inclusion in the followed backward snowball approach: *G2P, grapheme to phoneme conversion, poetry generation, rap generation, lyric generation, melody generation, pun generation, joke generation, tongue twisters, phonetics, phonology, phonological, poetry translation, joke translation, phonetic translation.*