NB Uttale: A Norwegian Pronunciation Lexicon with Dialect Variation

Marie Iversdatter Røsok, Ingerid Løyning Dale
The Language Bank at the National Library of Norway,
Oslo, Norway
{marie.rosok, ingerid.dale}@nb.no

Abstract

We present a Norwegian pronunciation lexicon with Bokmål orthographic word forms and up to eight alternate phonological transcriptions per word form. The lexicon covers dialectal variations for five geographical areas, as well as pronunciation variations for spontaneous and manuscript-read speech. It is based on the NST Bokmål lexicon for East Norwegian, whose original phonological transcriptions have been corrected, before they were converted with dialect specific regular expression rules. To evaluate the quality and consistency of the new, rule-generated transcriptions, we trained grapheme-to-phoneme (G2P) models and report our results with word- (WER) and phoneme-error-rate (PER) metrics. We found that the G2P models trained on lexica for Southwest and West Norwegian *close-to-written* transcriptions have the lowest WER scores, and that all error-corrected, *close-to-written* lexica yield better WER scores than the original NST lexicon. The lexicon is available under an open license, and can be used for various language technology applications and in linguistic research.

Keywords: Lexicon, Lexical Database, Phonetic Databases, Phonology, Speech Resource/Database, Validation of LRs, Corpus

1. Introduction

Pronunciation lexica have long been an essential linguistic resource in speech recognition and speech synthesis systems, and serve as training data for grapheme-to-phoneme (G2P) models as well as forced alignment models, commonly used to generate segmented transcriptions for further analysis in phonetic and phonological studies. A lot of effort has been put into improving the design (Lamel and Adda, 1996), linguistic quality, and coverage (Schlippe et al., 2014) of pronunciation lexica. G2P models trained on pronunciation lexica were found to improve system outputs by suggesting candidate transcriptions for out-of-vocabulary (OOV) words (Jouvet et al., 2012). Many speech synthesis systems still rely on linguistic expertise in the form of a pronunciation lexicon and a G2P model, despite recent years' stark improvements of end-to-end systems (Radford et al., 2023; Wang et al., 2017). Regardless of system architecture, a recurring issue for speech technology systems is the lack of enough representative data, especially for low resource languages. Thus, speech technology systems and tools reaching state-of-the-art levels of performance tend to cater to English, and most systems in other languages are developed only for a mainstream standard, or the most commonly spoken dialect (Rehm and Way, 2023). Not only does this reinforce the dominance of the most common language variant (spoken or written) in technological advancements, but also diminishes the tools' usability for research on language variations and low(er) resource languages.

We present NB Uttale[1], a pronunciation lexicon for Norwegian with up to eight distinct dialectal transcription variants per word entry. It covers five dialect areas, each with a *close-to-spoken* and a *close-to-written* style variant. The pronunciation variants in NB Uttale were created by applying transformation rules to the East Norwegian transcriptions in NST Pronunciation Lexicon for Norwegian Bokmål (Nordisk Språkteknologi, 2003). The updates include both error-fix rules that correct incorrect transcriptions in the NST lexicon, and dialect rules that change the transcriptions into other dialectal variants. NB Uttale provides a machine readable overview of dialectal variation in Norwegian phonology as of 2022, and the resource can be of use in other, dialect specific tools or in research on spoken Norwegian. We present the lexicon and the development of the transformation rules in section 2, and describe the evaluation criteria and method we used to assess the lexicon quality in section 3. We present and discuss our findings in section 4, and section 5 concludes the paper.

2. Pronunciation Lexicon for Dialectal Variation in Norwegian

Norwegian is a relatively small language, spoken in Norway with a population of 5.5 million (Statis-

[1] https://www.nb.no/sprakbanken/en/resource-catalogue/oai-nb-no-sbr-79/

tics Norway (SSB), 2023). The written language has two separate grammatical standards, Nynorsk (13%) and Bokmål (86.5%)[2]. The spoken language has no official standard, and dialectal variation can be observed in syntax, vocabulary, phonemic patterns, tonemes, as well as prosody.

NB Uttale is a pronunciation lexicon for Norwegian Bokmål with phonemic dialectal transcriptions for five dialect areas. The lexicon builds on the NST pronunciation lexicon for Norwegian Bokmål, which was created in 2003 and consists of approximately 785 000 word entries with transcriptions in Standard East Norwegian. The NB Uttale lexicon adds transcriptions for four more dialect areas, as well as 25 000 new words. The five broad areas are East (e), Central (t), North (n), Southwest (sw) and West Norwegian (w). The dialects are mutually intelligible, but they have characteristic differences in inflectional paradigms and phonological processes on the phonemic level that can change the meaning of a word between dialects.

2.1. Lexicon Updates

The dialectal lexica were created by applying string transformation rules to the original NST lexicon, written by trained linguists at the Language Bank at the National Library of Norway. Regular expressions were formulated to match and substitute phoneme sequences in a group of words that share a phonemic pattern. Most of the transformation rules changed the transcriptions into dialectal variants, but some rules were dedicated to correcting transcription errors in the NST lexicon. The rules were implemented in a cumulative manner and many transcriptions were matched by more than one rule to arrive at the final dialectal form. The transformation rules consisted of search-and-replace expressions for the phonemic transcriptions, plus optional constraints on grammatical tags and grapheme sequences that needed to be present in the orthographic word for the rule to apply. Wordforms that should not be matched by a rule were added to a list of exemptions. The final lists of rules, exemptions, new words, as well as the python tool Lexupdater that was used to apply the rules and generate the new transcriptions are all openly available on Github[3]. All rules and the resulting new transcriptions were reviewed before they were added to the lexicon to ensure that they were in line with the respective dialects.

Each dialect has both a *close-to-spoken* and a *close-to-written* transcription variant. The *close-to-spoken* variants represent the true dialectal pronunciations in unplanned speech, while the *close-to-written* variants represent the pronunciation of manuscript-read Bokmål. The resulting pronunciation lexicon consists of ten dialectal lexica. However, two pairs of lexica are identical (*close-to-written* sw and w, *close-to-written* n and t).

2.2. Error Correction

The lexicon updates for NB Uttale included corrections of wrong transcriptions in the NST lexicon. The error correction rules modified both wrong transcriptions that misrepresented actual pronunciation, as well as inconsistencies in the transcriptions. An example of the first is the wrong use of the syllabic retroflex consonant /ɳ̩/ in some single definite noun suffixes such as in *bakeren* ("the baker"), where the correct phoneme is the non-syllabic /ɳ/. The incorrect nasal resulted in an extra syllable for an entire category of nouns. An example of the second is the mixed use of the retroflex fricative /ʂ/ and the postalveolar fricative /ʃ/. Both are present in the lexicon and represent the same sound, but have clear and separate uses.

Postalveolar /ʃ/ is part of the core phonological inventory of Norwegian (Kristoffersen, 2000), while /ʂ/ is the result of assimilation between an adjacent /r/ and /s/ sound (see section 2.3.) and is closer to the phonetic realisation in Eastern Norwegian. While the two variants can be considered allophones of the same coronal fricative phoneme, and therefore mutually exclusive in a single phonemic system, the choice to keep both was a practical and operational one. Firstly, the retroflex phoneme retains information about the assimilation process and was necessary for the dialectal updates. Secondly, the two sounds correspond to different grapheme sequences, which is relevant for grapheme-to-phoneme alignment. /ʂ/ corresponds to *rs*, while /ʃ/ typically corresponds to *sj* and *skj*. The two phonemes were sometimes used interchangeably, which caused an inconsistent mapping between graphemes and phonemes. This was corrected in the updated lexica.

2.3. Dialectal Transcription Variants

Before adding the dialectal transcriptions, the linguists consulted available literature on Norwegian dialects to identify the dialectal phenomena that should be covered, as well as dialectal speech corpora for examples of pronunciation, particularly Nordic Dialect Corpus (Johannessen et al., 2009). The dialect rules changed the East Norwegian transcriptions into dialectal pronunciation variants.

Many of the dialect rules dealt with retroflex phonemes in Southwest and West Norwegian. East, North and Central Norwegian have retroflexion. This is to a large extent a predictable phonological process where /r/ (or a retroflex consonant) in

[2] Usage among pupils in primary and lower secondary school (Statistics Norway (SSB), 2002-2022)

[3] https://github.com/Sprakbanken/nb_uttale/tree/v1

front of /s, l, n, t, d/ assimilate into the retroflex version of the latter consonant, i.e. /ʂ, ɭ, ɳ, ʈ, ɖ/. This phonological process is not present in the western dialects, and so retroflex phonemes in the NST lexicon were broken up into /r/ and the following consonant. These dialect rules were applied to the western *close-to-spoken* and *close-to-written* lexica, alike.

Other dialect rules target grammatical suffixes, like noun, verb and adjective suffixes. The morphological variation in these grammatical groups are important markers that differentiate Norwegian dialects (Skjekkeland, 1997). For instance, North and Central Norwegian are marked by apocope, where the final syllable of many nouns, verbs and adjectives is deleted. The plural definite noun suffix *-ene*, for example, is pronounced /ənə/ in East Norwegian, but /ɑn/ in Central Norwegian, where the first vowel is changed and the second deleted. The NST lexicon is annotated with grammatical information which enable the rules to target specific groups of words, e.g. only plural definite nouns.

Some rules specify the words or grapheme sequences the rule should apply for, particularly in cases where the phonological variation is unpredictable and lexically conditioned. An example is vowel changes in North and Central Norwegian. These dialects see lowering of root vowels in many words, like /ɪ/ to /ɛ/ such as in *fisk* ("fish"), /ɛ/ to /æ/ such as in *veldig* ("very") and more. These changes are not predictable for a certain phonemic environment, nor tied to a grammatical category, and are therefore restrained to an explicit list of words.

2.4. Close-to-spoken and Close-to-written Lexica

The bulk of the dialect rules apply to the *close-to-spoken* transcription variants, which consequently differ more across dialects. The *close-to-written* transcriptions, on the other hand, represent manuscript-read pronunciations. Since the transcriptions are phonemic, not allophonic, the differences between the dialect areas for manuscript-read speech are minimal, and the *close-to-written* lexica are more similar to East Norwegian, and to each other. The North and Central *close-to-written* lexica are affected by error correction rules and the same three dialect rules, and are identical. The Southwest and West *close-to-written* lexica are affected by error correction rules and dialect rules that break up retroflexes, and are also identical. Hence there are eight distinct dialectal lexica, rather than ten.

3. Evaluation

We wanted to investigate the effect the error correction and dialect rules have had on the transcriptions in the pronunciation lexica. We considered correctness and consistency as the main criteria for evaluation of the quality of our lexica, and have trained and evaluated G2P models as a quantitative proxy measure of these features. We report word-error-rates (WER) and phoneme-error-rates (PER) for the models, one model per lexicon variant.

Consistency relates to both phonemic consistency and grapheme-phoneme transparency. The same phoneme should be transcribed for the same speech sound across the lexicon, meaning that different phonemes should not be used interchangeably to represent the same sound. Also, grapheme sequences should have the same phonemic realisation for similar phonological contexts. This leads to a more transparent relationship between graphemes and phonemes in the lexicon. The pronunciation lexicon will never have complete consistency, because the the same letter can have different pronunciations and this variation is not always predictable from the orthographic or phonemic environment. Still, the more consistent the phonemic transcriptions are, the more likely it is that the statistical model will generalise well and reproduce the expected phoneme sequences for a given grapheme sequence.

Correctness refers to how the correct phoneme should be transcribed for a given speech sound. It is difficult to quantitatively evaluate the transcriptions' correctness against real, spoken, dialectal realisations of each word, without a full dataset of recorded and transcribed speech for all the words and all the dialects. The PER metric serves as a measure of how well a statistical G2P model trained on the lexicon aligned and generalised the patterns for grapheme-phoneme sequence mapping from the lexicon entries. Hence, transcription errors in the lexicon would result in higher PER numbers, and the model might reproduce these same errors. The error correction rules were developed to improve both correctness and consistency of the transcriptions, and we assume that these rules improve the G2P models' performance. The dialect rules ensure correct phonemic realisations of dialectal pronunciation variants, but because dialectal variation can be complex and unpredictable, they sometimes compromise on consistency. Here, we refer back to the example of vowel lowering in section 2.3. In this and other cases, a dialectal update has only been implemented for a specified list of words, not for all matching phonemic contexts in the lexicon. Similarily, dialectal pronunciations that are further removed from the orthographic standard can make the grapheme-phoneme relationship even less transparent. For instance, the letter *a* is not typically realised as the sound /ɪ/. The plural definite neuter noun suffix *-a* is pronounced /ɑ/ in East Norwegian *husa* ("the houses"). In West Norwegian, the final vowel is /ɪ/, an unconventional realisation of the letter *a* in the lexicon.

Yet, other dialectal updates might help grapheme-phoneme transparency. In East Norwegian some final plosives are silent after liquids, e.g. /d/ after /n/ in *kveld* ("night"), pronounced /kʋɛl/. West Norwegian has retained the final plosive /kʋɛld/, and the updated transcription therefore results in a one-to-one mapping between graphemes and phonemes.

In order to check how well the lexicon transcriptions could be generalised, and report our results quantitatively, we trained weighted finite state transducer (WFST) G2P models with Phonetisaurus (NOVAK et al., 2016). We chose the Phonetisaurus framework [4] for training G2P models due to ease of access and short training time, even on a CPU. The transcriptions and orthographic sequences were first aligned, allowing for many-to-many symbol mappings, e.g. *ng* to /ŋ/, or *x* to /ks/. The aligned dictionaries were used to train 8-gram models with the mitlm toolkit[5]. We have used the same configuration settings for all the models we trained, to ensure that we compare differences between the input data, and not the model architectures.

We split the lexicon in a train (80%) and test set (20%) based on word form IDs and ensured that there were no overlapping word forms in the two partitions. Since we were most interested in evaluating the effect of error correction and dialect conversion, we kept new word entries out of the partitions. Each G2P model we report on has been trained on a partition of the lexicon with transcriptions pertaining to a single dialect area (e.g. *e*) and a single pronunciation variant (e.g. *written*). Our baseline model is trained on the original transcriptions in the NST lexicon, which are most similar to the North and Central Norwegian *close-to-written* transcriptions, measured in how many transformation rules were applied.

4. Results

We report WER and PER for the trained G2P models for all the updated dialectal lexica. The results in Table 1 show that the models derived from the *close-to-written* lexica perform better in terms of WER and PER than the model derived from the original NST lexicon. The East, North and Central *close-to-written* lexica were mostly affected by error corrections and got only three (North, Central) or four (East) dialect updates. These changes to the lexica improved G2P model performance.

Lexicon	WER	PER
NST original	14.60	2.82
e_written	14.05	2.70
w_written, sw_written	**13.54**	**2.54**
t_written, n_written	14.06	2.72
e_spoken	14.05	2.72
w_spoken	17.19	3.41
sw_spoken	16.73	3.34
t_spoken	16.80	3.47
n_spoken	17.54	3.75

Table 1: The G2P models trained on Southwest and West Norwegian *close-to-written* transcriptions have the lowest error rates. These two lexica are identical, and so are the North and Central Norwegian *close-to-written* lexica. The North Norwegian *close-to-spoken* model has the highest error rates.

Interestingly, the sw_written and w_written models perform the best of all. The Southwest and West *close-to-written* lexica were subject to quite a few dialect rules in addition to error corrections, namely the rules that break up retroflex phonemes. This is the main difference between the Southwest and West *close-to-written* lexica and the East, North and Central *close-to-written* lexica. The added performance improvement can be attributed to the lack of retroflex phonemes in the first pair of lexica. Among the wrongly predicted transcriptions for sw_written/w_written, we found no occurrences of incorrect predictions of *rs, rl, rn, rd* and *rt* sequences. This stands in contrast to the e_written and n_written/t_written model predictions where incorrect realisations of the same sequences are common. The improvement is explained by differences in alignment. Transcriptions with retroflex phonemes require a many-to-one alignment between grapheme and phoneme sequences (*rn* to /ɳ/). Since the Southwest and West transcriptions do not have retroflex phonemes, they retain a one-to-one mapping between graphemes and phonemes (*rn* to /rn/).

The e_spoken model performs on par with the *close-to-written* models. This is expected because the e_spoken lexicon has very few dialectal transformation rules and is very similar to its *close-to-written* counterpart. Only one dialect rule differentiate these two lexica.

The remaining four *close-to-spoken* models show the highest error rates, and they are also derived from the lexica that undergo the most substantial dialectal transformations. What we read from this is that the dialect updates (beyond the ones that break up retroflex phonemes) introduce some complexities or imperfections to the lexica that negatively

[4] https://github.com/AdolfVonKleist/Phonetisaurus

[5] https://github.com/mitlm/mitlm

impact performance of the derived G2P models. This is to be expected because dialectal variation is complex.

The dialect rules, more often than not, filter words by grammatical features and change the phonemic realisation for only a specific category of words (e.g. present tense verbs, singular definite nouns). These changes are structured when considered in light of their grammatical categories, however, the G2P model does not have access to these grammatical annotations. It only examines the relationship between the orthographic words and phonemic transcriptions, and this relationship is made more complex by the dialect rules, which leads to higher error rates.

Here, we note that the WER and PER scores of all the models are in the same vicinity. None of the models show a dramatic increase or decrease in performance, and the discrepancies between the worst and best performing models are no more than 4 and 1.21 for WER and PER, respectively. The results show that the models are able to generalise over the transcriptions and produce correct dialectal transcriptions for over 80% of the words in the test set. We take from this that the dialectal updates have had the intended effect on the lexica and have not inadvertently created major problems with the correctness and consistency of the transcriptions.

5. Conclusion

In this paper we have evaluated the ten (eight distinct) dialectal lexica for Norwegian Bokmål in NB Uttale to investigate the effects of the lexicon updates on the transcriptions. We trained WFST G2P models on each lexicon using the Phonetisaurus framework and looked at the WER and PER scores as a measure of the consistency and correctness of the transcriptions. The results show that the error corrections make transcriptions more correct and consistent, and improve G2P model performance. Despite added complexity in some of the dialect updates, all models still produce correct transcriptions for more than 80% of the words. Importantly, transcriptions without retroflex phonemes yield even greater benefit to performance because of the one-to-one mapping between graphemes and phonemes.

Our investigations show that there are no unintended consequences that severely reduce the quality of the transcriptions. We conclude, therefore, that the lexica overall maintain correct and consistent transcriptions at the same time as they ensure dialectal pronunciation variants for Norwegian dialects. This makes NB Uttale a valuable resource with information about features of Norwegian dialects in a machine-readable format, which can contribute to for instance dialect specific speech synthesis, or more accurate forced alignment for dialectal speech in Norwegian.

6. Acknowledgements

The work on the pronunciation lexicon NB Uttale has in part been supported by the IKTPLUSS grant for the SCRIBE project financed by the Norwegian Research Council (KSP2IPD).

Thank you also to Per Erik Solberg for his support and feedback during the work on this paper.

7. Bibliographical References

Maximilian Bisani and Hermann Ney. 2008. Joint-sequence models for grapheme-to-phoneme conversion. *Speech Communication*, 50(5):434–451.

Awni Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, and Andrew Y. Ng. 2014. Deep speech: Scaling up end-to-end speech recognition.

Janne Bondi Johannessen, Joel Priestley, Kristin Hagen, Tor Anders Åfarli, and Øystein Alexander Vangsnes. 2009. The nordic dialect corpus - an advanced research tool. In *Proceedings of the 17th Nordic Conference of Computational Linguistics NODALIDA 2009*, NEALT Proceedings series volume 4. `https://tekstlab.uio.no/nota/scandiasyn/`.

Denis Jouvet, Dominique Fohr, and Irina Illina. 2012. Evaluating grapheme-to-phoneme converters in automatic speech recognition context. In *ICASSP - 2012 - IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4821 – 4824, Kyoto, Japan.

Zhenzhen Kou, Daisy Stanton, Fuchun Peng, Françoise Beaufays, and Trevor Strohman. 2015. Fix it where it fails: Pronunciation learning by mining error corrections from speech logs. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4619–4623.

Gjert Kristoffersen. 2000. The phonology of norwegian.

L. Lamel and G. Adda. 1996. On designing pronunciation lexicons for large vocabulary continuous speech recognition. In *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP '96*, volume 1, pages 6–9 vol.1.

JOSEF ROBERT NOVAK, NOBUAKI MINEMATSU, and KEIKICHI HIROSE. 2016. Phonetisaurus: Exploring grapheme-to-phoneme conversion with joint n-gram models in the wfst framework. *Natural Language Engineering*, 22(6):907–938.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518. PMLR.

Georg Rehm and Andy Way, editors. 2023. *European Language Equality: A Strategic Agenda for Digital Language Equality*. Cognitive Technologies. Springer International Publishing, Cham.

Tim Schlippe, Wolf Quaschningk, and Tanja Schultz. 2014. Combining grapheme-to-phoneme converter outputs for enhanced pronunciation generation in low-resource scenarios. In *Proc. 4th Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU 2014)*, pages 139–145.

Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. 2019. wav2vec: Unsupervised pre-training for speech recognition.

Martin Skjekkeland. 1997. Dei norske dialektane : tradisjonelle særdrag i jamføring med skriftmåla.

Statistics Norway (SSB). 2002-2022. Pupils in primary and lower secondary school. `https://www.ssb.no/en/statbank/table/03743/`.

Statistics Norway (SSB). 2023. Population in norway. Updated 23 August, 2023. `https://www.ssb.no/en/befolkning/folketall/statistikk/befolkning`.

Herman J. M. Steeneken and David A. van Leeuwen. 1995. Multi-lingual assessment of speaker independent large vocabulary speech-recognition systems: THE SQALE-PROJECT. In *Proc. 4th European Conference on Speech Communication and Technology (Eurospeech 1995)*, pages 1271–1274.

Xu Tan, Jiawei Chen, Haohe Liu, Jian Cong, Chen Zhang, Yanqing Liu, Xi Wang, Yichong Leng, Yuanhao Yi, Lei He, Frank Soong, Tao Qin, Sheng Zhao, and Tie-Yan Liu. 2022. Naturalspeech: End-to-end text to speech synthesis with human-level quality.

Yuxuan Wang, R.J. Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J. Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, Quoc Le, Yannis Agiomyrgiannakis, Rob Clark, and Rif A. Saurous. 2017. Tacotron: Towards End-to-End Speech Synthesis. In *Proc. Interspeech 2017*, pages 4006–4010.

8. Language Resource References

Janne Bondi Johannessen and Joel Priestley and Kristin Hagen and Tor Anders Åfarli and Øystein Alexander Vangsnes. 2009. *The Nordic Dialect Corpus - an Advanced Research Tool*. Proceedings of the 17th Nordic Conference of Computational Linguistics NODALIDA 2009, NEALT Proceedings Series Volume 4. `https://tekstlab.uio.no/nota/scandiasyn/`.

Nordisk Språkteknologi. 2003. *NST Pronunciation Lexicon for Norwegian Bokmål*. Språkbanken, National Library of Norway. PID hdl:21.11146/23. `https://www.nb.no/sprakbanken/ressurskatalog/oai-nb-no-sbr-23/`.