Bornholmsk Natural Language Processing: Resources and Tools

Leon Strømberg Derczynski
ITU Copenhagen
Denmark
ld@itu.dk

Alex Speed Kjeldsen
University of Copenhagen
Denmark
alex@hum.ku.dk

Abstract

This paper introduces language processing resources and tools for Bornholmsk, a language spoken on the island of Bornholm, with roots in Danish and closely related to Scanian. This presents an overview of the language and available data, and the first NLP models for this living, minority Nordic language.

Sammenfattnijng på borrijnholmst: Dæjnna artikkelijn introduserer natursprågsresurser å varktoi for borrijnholmst, ed språg a dær snakkes på ön Borrijnholm me rødder i danst å i nær familia me skånst. Artikkelijn gjer ed åuersyn åuer språged å di datan som fijnnes, å di fosste NLP modællarna for dætta læwenes nordiska minnretålsspråged.

1 Introduction

Bornholmsk is a language spoken on Bornholm, an island in the Baltic Sea, the easternmost land mass of Denmark.[1] Bornholmsk is an endangered language. Inhabitants of Bornholm have been changing to using standard Danish over the past century – a development that has escalated within the last 20 years or so; cf. Larsen (2019). In total the island has around 40.000 residents, though there is notable migration to and from the other Danish islands and the mainland, leading to a Bornholmer diaspora.

Given the endangered status of the language, it is important to capture knowledge about it now. One way of doing this is to create tools for working with the language. In particular, we attempt to build machine translation support for Bornholmsk, to not only assist with understanding the language, but also to enable users of it to stick with Bornholmsk instead of being forced to switch to standard Danish – a factor in language erosion – while helping open access to Bornholmsk to those who use standard Danish. Additionally the development of such tools could give higher linguistic status to Bornholmsk among its potential users.

(1) *Fârijn kjöre te böjn å fikkj âu ejn*
Faren kørte til byen og fik også en
fæzelia nætter kjâul kjefter te 'na
utrolig pæn kjole købt til hende
'The father drove to town and got her a really beautiful dress'

(2) *Horrana hâ løvved ætte dæjn piblijn hela dâjn*
Drengene har løbet efter den pige hele dagen
'The boys have run after that girl the entire day'

Figure 1: Sentences in Bornholmsk

Code switching between Danish and Bornholmsk remains common and has been for some time (Baumann-Larsen, 1973).

Historically Bornholmsk is categorised as East Danish (along with the language spoken in Skåne, Halland and (part of) Blekinge) of which it is the only representative in present-day Denmark. Examples of distinctive linguistic features are: 1) the existence of three grammatical genders (the gender inflection is not limited to the definite article, but is also manifested in adjectives, past participles and possessive pronouns). 2) An enclitic form of the third person personal pronoun, namely masculine *-(i)jn* "him" and feminine *-na* "her". 3) The occurrence of *a* in unstressed syllables along with *e* (as well as *i* and *u* in certain contexts). 4) So-called "double definiteness" like in Norwegian and

[1] Following the most common usage on Bornholm we refer to Bornholmsk as a separate language and not a variant of Danish. Although Bornholmsk is normally described as a Danish dialect (the language code for Bornholmsk under IETF BCP-47 is da-bornholm), this shouldn't pose any problems in the context of this paper.

Swedish. Of other, perhaps less distinctive features, one could mention: 5) A special intonation (neither glottal stop nor pitch-accent is used). 6) Two (long) *a* variants. 7) Palatal variants of *g*, *k*, *l* and *n*. 8) A voiced variant of *s* (*z*). 9) A more archaic verbal inflectional system. 10) Different usage of the reflexive *sig*/*dem* compared to standard Danish. 11) Many lexical differences compared to standard Danish (including very common words). Examples of Bornholmsk are given in Figure 1.

A detailed description of Bornholmsk phonology (*Lautlehre*) and morphology is given by Thomsen and Wimmer in their introduction to Espersen et al. (1908). Shorter, general introductions and descriptions, some of which are of more popular nature, are found in Møller (1918, 25–70), Prince (1924) (many errors and misunderstandings), Rohmann (1928), Koefoed (1944, 1969) Sonne (1957), and Pedersen (2018). An exploration of the syntax of Bornholmsk can be found in Pedersen (2009). See also Pedersen (2013, 31–32) on the *s*-passive in Bornholmsk.

Compared to other Danish dialects Bornholmsk has been utilised much more frequently in writing. The 1920s–1940s is considered the Golden Age for written Bornholmsk, but the tradition dates back to the 19th century, and writings in Bornholmsk have continued to be published until this day, e.g. in local newspapers. In recent years the language has also found its way to social media (generally in a less canonical form). In spite of the lack of normative (spelling) dictionaries and formal training most speakers of Bornholmsk find it reasonably easy to read Bornholmsk. The reason for this is at least fourfold: 1) familiarity/tradition (users have been exposed to the language in its written form in newspapers etc.). 2) there is generally a fairly straightforward mapping between spoken and written Bornholmsk, presumably also to a greater extent than for other Danish dialects.[2] 3) Regional variation is very limited (when excluding the so-called "Rønna-fint"). 4) Until very recently the language has changed quite slowly compared to most other Danish dialects (the sound system is e.g. still more or less identical to the system described in Espersen et al. 1908). For the same reasons most of the orthographic variation found in actual examples of written Bornholmsk is of a kind that can be normalised fairly easily without losing any actual linguistic information.

[2]If other Danish dialects were to be transcribed using somewhat similar principles the result would deviate to a greater or lesser extent from both Bornholmsk and Standard Danish, depending on the dialect in question.

Name	Genre	Tokens
Otto J. Lund:		
"Bråfolk" å Stommene	Fiction	35K
"Lyngblomster"	Fiction (poetry)	5.6K
"Vår Larkan ryggar"	Fiction	55K
Crawled and scraped text	Web & social media	2K

Table 1: Monolingual Bornholmsk data

In this paper, we outline efforts to digitise and capture Bornholmsk resources, and see what can be done with the scarce resources currently available, leading to embeddings, a part-of-speech tagger, and a prototype machine translation system.

2 Corpora

Bornholmsk digital text is generally absent. It has no data in the UD treebank, nor in CLARIN-DK, nor the LDC repository. Collection thus proceeded ad-hoc. Via the web, we compiled an informal corpus of texts including illustrative examples of the language (from e.g. Wikipedia pages), poems, song lyrics, social media comments, and stories. Additionally, some websites include small introductions to phrases in Bornholmsk for Danish speakers;[3] these serve multiple functions, providing sentences in the target language, as well as word:word translations, and finally acting as sentence-level parallel text data. In addition to material collected via the web, we use resources that have been digitised within the recently resumed *Bornholmsk Ordbog* (BO) dictionary project.[4]

A dictionary in digital format, primarily based upon Espersen et al. (1908), but supplied with various other lexicographic resources, has been compiled by Olav Terkelsen and is available from `http://onpweb.nfi.sc.ku.dk/espersen/index.html`. This material has not been used in this paper, but since the citations and phrases are translated into modern standard Danish, they represent a good candidate for future parallel text.

Other lexical resources have also been digitised, e.g. LærOrdb (1873), Adler (1856) and the glossary found in Skougaard (1804). Together with two very large, lexically ordered records of Bornholmsk,[5] primarily composed between 1923 and

[3]See e.g. Allan B. Hansen's gubbana.dk.

[4]For a description of this project, see Kjeldsen (2019).

[5]These records contain about twice as many lemmata as

1931 by the three original editors of BO, and the part of BO which was edited before work on the project came to a halt in the 1940s, these resources will be published as a fully searchable meta dictionary in August 2020. For this reason, apart from a smaller part of the edited part of BO which is used for training of the MT models (about 3000 sentence pairs), these sources have not been used in the present project.

Some prose and poems have been digitised, namely three longer prose texts written by Otto J. Lund (*Mâgårsfolken*, Lund 1935b, *Enj Galnerojs*, Lund 1935a, and *Brâfolk å Stommene*, Lund 1941), a number of poems by the same author, *Lyngblomster* (Lund, 1930), as well as a collection of folk stories published by J. P. Kuhre in 1938 under the title *Borrinjholmska Sansâger* has been used. The latter text collection is of special value: it is in many respects the best written representative of canonical Bornholmsk, the orthography used is unusually consistent and each story is translated to (somewhat old fashioned) standard Danish, more or less sentence by sentence. Although not identical, the orthographic principles used by Kuhre are very similar to those used in the BO dictionary project.

A data statement (Bender and Friedman, 2018) for these resources is given in the appendices. The data used in and produced by the dictionary project will be published under CC BY-SA.

3 Embeddings and Alignment

Given some text in Bornholmsk, we attempted to induce distributional word embeddings. For this, we chose FastText (Bojanowski et al., 2017). As Bornholmsk is a low-resource language, it is important to be able to connect it to other languages easily. Standard FastText embeddings are available for many languages. FastText supports subword embeddings, which are likely to be useful in a language like Bornholmsk that has a relatively small alphabet, and also have some chance of compensating for the high data sparsity.

Embeddings are induced with 300 dimensions, in order to be compatible with the public Common Crawl-based FastText embeddings. Having induced these embeddings for Bornholmsk $e_{bornholmsk}$, they are then aligned into the embedded space of Danish from FastText e_{danish}. We try three alignment methods: (1) unsupervised alignment, where matching surface forms are used as anchor points for the two embedded spaces; (2) alignment augmented with the 1:1 word dictionaries captured earlier, where these translations are used as anchor points; (3) a mixed alignment, using both unsupervised and supervised points. Dictionary words missing from just one language are inserted into the dictionary using the embedding of anchor point in the other language, post-alignment. We choose Danish (e_{danish}) as the target space for Bornholmsk as the two languages are likely to have some lexical overlap, and there is vastly more data for Danish.

Espersen's dictionary.

da-bo 'hvid'	**bo-da 'vid'**	**da-bo 'morgen'**	**bo-da 'mârn'**
vid	hvid	mârn	morgen
vidd	sort	Imârn	aften
vid-	rød	mârnmål	eftermiddag
vidt	gul	mârnijn	majmorgen
vida	hvidfarvet	mârna	formiddag

Table 2: Closest words after supervised alignment

To align vectors, a transformation is built from the singular value decomposition of the product of the target space and the transpose of the source space (Smith et al., 2017). This orthogonal transformation aligns the source language to the target, thus mapping Bornholmsk embeddings into e_{danish}. A test set of 10 % of the bilingual mappings was held out for evaluation. In this case, the mean similarity was 0.3469 for unsupervised (i.e. lexical match) anchoring, 0.4238 for supervised anchoring over translated word pairs, and 0.3959 for the union of unsupervised and supervised anchor pairs. We can see that while the unsupervised alignment is helpful, when supervised pairs are available, it detracts from performance. Table 2 shows closest pairs for sample words.

4 Part-of-speech Tagging

Because there is no part-of-speech (PoS)-tagged data, we must look to resources from other languages. Using aligned embeddings, it is possible to train a PoS tagger for one language l_{source} where the words are represented in embeddings space e. By mapping words in sentences in a target language l_{target} into e, these sentences can be posed to the tagger as if they were in l_{source}. This requires that embeddings for both languages, e_{source} and e_{target}, are aligned to the general em-

beddings space e. There is also an assumption that l_{source} and l_{target} will be sufficiently distributionally and grammatically similar.

One is more likely to encounter new words during tagging when training data is limited, so a PoS tagger that tolerates previously-unseen words is preferable. The `structbilty` tagger[6] uses a bidirectional LSTM with language modelling as auxiliary loss function and achieves good accuracy on unknown words (Plank et al., 2016).

The source language evaluated is Danish and training and validation data is taken from the Universal Dependencies corpus (Nivre et al., 2016). Sans PoS-annotated Bornholmsk, we give example tagged sentences. Many structures and words picked up correctly, despite absent training data and a very small monolingual dataset for embedding induction. However, basic structures are occasionally missing (cf. #3).

1) *Hanj/PROPN fijk/VERB dask/NOUN på/ADP sinj/ADJ luzagâda/NOUN*

2) *de/PRON ska/VERB varra/X så/ADV galed/ADJ ,/PUNCT sa/SCONJ de/PRON ammar/VERB ijkkje/ADV ./PUNCT*

2) *Hon/PROPN ve/X hâ/X ham/PRON som/ADP kjærest/NOUN*

5 Danish-Bornholmsk Translation

Despite the low-resource situation, there is some useful data for developing Bornholmsk-Danish translation. These vary in term: Full translations of a few songs and poems can be found, which are parallel line-by-line. Snippets of words giving example uses in various informal 1:1 word-level dictionaries are also available – as well as the word mappings themselves.

We used Kuhre's folk stories as parallel Danish-Bornholmsk text. Further, we used entries from the nascent *Bornholmsk Ordbog*, which includes a number of genuine examples of how the language might be used. Noisier and non-canonical web data were included, to improve vocabulary coverage. The monolingual corpora is the basis for word embeddings, in this case with GloVe (Pennington et al., 2014) in 50 dimensions.

The Kuhre text is in an older form of Danish, some spelling reforms ago. Specifically, vowels are annotated differently (aa and ee vs. å and é), and nouns have a capital initial. This data is copied with case removed, and with the vowels converted to the modern format, so that the resulting model is not too surprised by modern Danish.

The *Bornholmsk Ordbog* is a work in progress, i.a. containing usage examples such as:

<bællana hadde âgebakka hærudanforr i vijnters {børnene havde kælkebakke herudenfor (huset) i vinters}>

These are converted into plaintext and used as supporting parallel examples. Table 3 gives an overview of the parallel text used.

5.1 Experimental Setup

We trained a translation model with OpenNMT (Klein et al., 2017) using all parallel text. The Bornholmsk side of this was combined with the Bornholmsk monolingual texts to build a language model and embeddings. Test and validation data were both 500 pairs taken from the input data. Parameters included: Glorot initialization, locked to the encoding vectors, dropout at 0.4, an average decay of 1e-4, and validation every 4000 steps.

5.2 Pilot Results

The translation performed reasonably, given the very small training data size. Examples:

Danish: *der stod en lys sky på en mørk baggrund .*
Output: *dær sto en art sjy på ejn âzstæl .*
Reference: *dær sto et lyst sjy på ejn morkjer bâggrujnn .*

Danish: *Vil du have lidt brød*
Output: *Vil du hâ lid brø*
Reference: *Ve du hâ lid brø*

Danish: *bliver der så at de alle kan komme op og køre?*
Output: *bler dær så a di ajle ver opp å kjöra ?*
Reference: *bler dær sa di ajle kajn komma opp å âga ?*

Danish: *hesten satte bagkoden så hårdt i stenen , at der er mærke efter det endnu .*
Output: *hæstijn satte bâgkodan så hårt i stenijn , at dær e mærke ætte dæjn len .*
Reference: *hæstijn satte bâgkodan så hårt i stenijn , at dær e mærke ætte 'd inu .*

Danish: *Hvor står mit klapbord*
Output: *Vor får minj dâuestola*
Reference: *Vor står mit flojbor*

Due to the relatively small size of the datasets involved today, we do not report an evaluation metric score. However, we do provide a qualitative evaluation with examples.

There are many unknown words, also in the reference data. These words are mostly a factor of the limited corpus size, and will become rarer as the amount of digitised Bornholmsk increases.

[6] `https://github.com/bplank/bilstm-aux`.

[7] These entries contain optional terms that are both expanded & omitted to create additional training data.

Name	Genre	Sentence pairs	Tokens (da)	Tokens (da-bornholm)
Borrinjholmska Sansåger (Kuhre)	Fiction	1K	29.1K	30.2K
Bornholmsk Ordbog: citations	Dictionary	3.6K[7]	41.9K	34.2K
Gubbana.dk	Web data	1.4K	8.3K	8.0K
Web scrape	Web data	<1K	<1K	<1K

Table 3: Parallel corpora for Bornholmsk-Danish

Another result of the low volume of training data is that there are few points in the embeddings. This means that the nearest neighbour may actually have a quite different meaning to the target. Thus, qualitatively, when the model gets things wrong, it will tend to get them really wrong.

Despite its small size and different training genre, the model was also capable of producing tolerable output given colloquial modern Danish. Some Danish spellings creep through and the phrasing is imprecise, but nevertheless, a rough mapping is available between the two languages:

Danish: *Hej , hvad hedder du ?*
Model output: *Te , va heder du ?*
Correct Bornholmsk: *Hai , va heder du ?*

Danish: *Det er Mads og han er en god dreng.*
Model output: *Ded e slæføre å hajn e en go majn.*
Correct Bornholmsk: *De(d) e Mads å hajn e ejn goer horra.*

Danish: *København er en af de større byer – faktisk den største .*
Model output: *København e en majed råganat !*
Correct Bornholmsk: *Kjøvvenhawn e ejn å di storre byana – fakta dæjn storsta .*

6 Related Work

There is no former work that we are aware of on NLP for Bornholmsk. The closest resource is an openly-available toolkit for Danish, DKIE (Derczynski et al., 2014), which is designed for the GATE platform (Cunningham et al., 2012), though even for Danish work is scarce (Kirkedal et al., 2019). Written Bornholmsk corpora are also rare; these exist almost entirely in smaller collections, some of which have been built with great care.

Two other Scandinavian tongues as small as Bornholmsk have had quite different stories. Faroese (ISO639: `fao`; BCP-47: `fo-FO`) is spoken by about 72000 people, many of whom live in the Faroes; it has a fairly long written tradition and is actively published in. It has some NLP visibility, being present in the Universal Dependencies treebanks, and a steady if slow stream of NLP research includes the language (e.g. Richter et al. (2018)). In contrast, Scandoromani (ISO639: `rmg`/`rmu`) has many fewer speakers than Bornholmsk; its original grammar has been overtaken by that of the dominant languages in the regions where it is spoken and is thus lost. There are nevertheless efforts to document the remnants of this tongue (Carling et al., 2014).

No machine translation is available for Scandoromani or Faroese. The Faroes built an innovative solution to this where phrases to be translated are distributed to citizens, who film themselves saying the translation, making essentially a translation memory (Kay, 1997) for Faroese.[8]

7 Conclusion

This work introduced resources and tools for doing natural language processing for Bornholmsk, an endangered Nordic language. Contributions included corpus creation, corpus collection, basic NLP resources, and a pilot translation model. The corpora are licensed separately; the NLP embeddings and models are available openly via ITU Copenhagen's NLP group page, `https://nlp.itu.dk/resources/`, and the public domain texts are available from this paper's authors. Future work should focus on digitising more text (incl. lexicographic resources); on making the best use possible out of the available corpora; on tuning models to perform better on the existing data; on increasing awareness around Bornholmsk; on helping learn Bornholmsk; and on making it possible for Bornholmsk-speakers to work digitally in Bornholmsk instead of Danish.

Acknowledgments

This research was carried out with support from the *Bornholmsk Ordbog* project sponsored by Sparekassen Bornholms Fond, Brødrene E., S. & A. Larsens Legat, Kunst- og Kulturhistorisk Råd under Bornholms Regionskommune and Bornholms Brandforsikring, and with thanks to the NEIC/NordForsk NLPL project. Titan X GPUs used in this research were donated by the NVIDIA Corporation. We are grateful to those who have gathered and published resources on Bornholmsk. Thanks to Emily M. Bender for helpful feedback.

[8] See https://www.faroeislandstranslate.com/ .

References

A. P. Adler. 1856. *Pröve paa et bornholmsk Dialekt-Lexikon, 1. og 2. Samling*. C. A. Reitzels Bo og Arvinger, København.

M. Baumann-Larsen. 1973. The function of dialects in the religious life of the bornholm inhabitants. In *Zur Theorie der Religion/Sociological Theories of Religion*, pages 236–242. Springer.

Emily M. Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Gerd Carling, Lenny Lindell, and Gilbert Ambrazaitis. 2014. *Scandoromani: Remnants of a mixed language*. Brill.

Hamish Cunningham, Diana Maynard, Kalina Bontcheva, Valentin Tablan, Niraj Aswani, Ian Roberts, Genevieve Gorrell, Adam Funk, Angus Roberts, Danica Damljanovic, Thomas Heitz, Mark A. Greenwood, Horacio Saggion, Johann Petrak, Yaoyong Li, Wim Peters, Leon Derczynski, et al. 2012. *Developing Language Processing Components with GATE Version 8 (a User Guide)*. University of Sheffield.

Leon Derczynski, C. Vilhelmsen, and Kenneth S Bøgh. 2014. DKIE: Open source information extraction for Danish. In *Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 61–64.

Johan Christian Subcleff Espersen, Vilhelm Thomsen, Ludvig Frands Adalbert Wimmer, and Viggo Holm. 1908. *Bornholmsk Ordbog*. Bianco Lunos Bogtrykkeri.

Martin Kay. 1997. The proper place of men and machines in language translation. *machine translation*, 12(1-2):3–23.

Andreas Kirkedal, Barbara Plank, Leon Derczynski, and Natalie Schluter. 2019. The Lacunae of Danish Natural Language Processing. In *Proceedings of the Nordic Conference on Computational Linguistics (NODALIDA)*. Northern European Association for Language Technology.

Alex Speed Kjeldsen. 2019. Bornholmsk Ordbog, version 2.0, forthcoming. *Mål og Mæle*, 40. årgang:22–31. [expected medio 2019].

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M. Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In *Proc. ACL*.

H. A. Koefoed. 1944. Sproget – Bornholmsk. In Hans Hjorth, editor, *Bornholmernes Land*, volume 2, pages 267–282. Bornholms Tidendes Forlag, Rønne.

H. A. Koefoed. 1969. Folkemål. In Bent Rying, editor, *Gyldendals egnsbeskrivelser. Bornholm*, pages 99–109. Gyldendal, København.

J. P. Kuhre. 1938. *Borrinjholmska Sansâger. Bornholmske folkeæventyr og dyrefabler*, volume 45 of *Danmarks Folkeminder*. Schønberg, København.

LærOrdb. 1873. *Bornholmsk Ordbog. Udgivet af Lærere*. Colbergs Boghandel, Rønne.

Anne Larsen. 2019. Dialekt på tværs af steder og generationer. Available online at https://dialekt.ku.dk/maanedens_emne/dialekt-i-periferien/.

Otto J. Lund. 1930. *Lyngblomster. Borrinjholmska Dækt*. Henry Andersen, Aakirkeby.

Otto J. Lund. 1935a. Enj Galnerojs. In *Vår Larkan ryggar. To borrinjholmska Fortællinjer*, pages 69–139. Eget Forlag, Aakirkeby.

Otto J. Lund. 1935b. Mâgårsfolken. In *Vår Larkan ryggar. To borrinjholmska Fortællinjer*, pages 5–68. Eget Forlag, Aakirkeby.

Otto J. Lund. 1941. *Brâfolk å Stommene, Fortælling*. Eget Forlag, Aakirkeby.

Peter Møller. 1918. *Det bornholmske Sprog*. Fritz Sørensens Boghandels Forlag, Rønne.

Joakim Nivre, Marie-Catherine De Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, et al. 2016. Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC)*, pages 1659–1666.

Karen Margrethe Pedersen. 2009. Bornholmsk dialektsyntaks. In *I mund og bog*, pages 249–262. Museum Tusculanum Press.

Karen Margrethe Pedersen. 2013. Refleksivt *sig/dem* – varianter gennem 800 år. *Danske talesprog*, 17:1–37.

Karen Margrethe Pedersen. 2018. Bornholmsk dialekt – i historisk og geografisk belysning. *Mål og Mæle*, 39. årgang:12–17.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Barbara Plank, Anders Søgaard, and Yoav Goldberg. 2016. Multilingual Part-of-Speech Tagging with Bidirectional Long Short-Term Memory Models and Auxiliary Loss. In *Proc. ACL*.

John Prince. 1924. The Danish Dialect of Bornholm. *Proceedings of the American Philosophical Society*, 63:190–207.

Caitlin Richter, Matthew Wickes, Deniz Beser, and Mitch Marcus. 2018. Low-resource post processing of noisy OCR output for historical corpus digitisation. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC)*, Miyazaki, Japan. European Languages Resources Association (ELRA).

Aage Rohmann. 1928. Det bornholmske sprog. *Bornholmske Samlinger*, 1. Række, 19:153–166.

Peder Nikolai Skougaard. 1804. *Beskrivelse over Bornholm, 1. Del*. København.

Samuel L. Smith, David H. P. Turban, Steven Hamblin, and Nils Y. Hammerla. 2017. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. In *Proc. ICLR (conference track)*.

H. P. Sonne. 1957. Sproget. In R. Nielsen and Th. Sørensen, editors, *Bogen om Bornholm*, pages 520–544. Danskernes Forlag, Aabenraa.

Appendix 1: Data Statement

Curation rationale Collection of Bornholmsk documents and parallel texts from speakers who have had Bornholmsk as their (dominant) L1.

Language variety BCP-47: `da-DK-bornholm`

Speaker demographic

- Speakers of Bornholmsk
- Age: mostly 60+
- Gender: male and female.
- Race/ethnicity: mostly of Scandinavian descent.
- Native language: Danish (Bornholmsk).
- Socioeconomic status: various.
- Different speakers represented: unknown.
- Presence of disordered speech: Generally not prevalent.

Annotator demographic

- Age: 30+
- Gender: male and female.
- Race/ethnicity: white northern European.
- Native language: Danish (Bornholmsk).
- Socioeconomic status: unknown.

Speech situation Literary works, with some ad-hoc collections and samples of the language.

Text characteristics Mostly literary works.

Provenance Original authors are credited in this work.