NoMusic – The Norwegian Multi-Dialectal Slot and Intent Detection Corpus

Petter Mæhlum and **Yves Scherrer**
Language Technology Group
Department of Informatics
University of Oslo, Norway
pettemae@ifi.uio.no, yves.scherrer@ifi.uio.no

Abstract

This paper presents a new textual resource for Norwegian and its dialects. The NoMusic corpus contains Norwegian translations of the xSID dataset, an evaluation dataset for spoken language understanding (slot and intent detection). The translations cover Norwegian Bokmål, as well as eight dialects from three of the four major Norwegian dialect areas. To our knowledge, this is the first multi-parallel resource for written Norwegian dialects, and the first evaluation dataset for slot and intent detection focusing on non-standard Norwegian varieties. In this paper, we describe the annotation process and provide some analyses on the types of linguistic variation that can be found in the dataset.

1 Introduction

Over the last decades, various textual resources covering Norwegian dialects have been produced. This paper reports on the creation of yet another Norwegian dialect dataset which has some unique properties that set it apart from previous work.

As a starting point, we use the xSID corpus (van der Goot et al., 2021), which consists of natural prompts asked to digital assistants (e.g., *Is it going to rain today?*, *Change tomorrow morning's alarm to 6 am.*). A digital assistant will have to (a) recognize the *intent* of the prompt and (b) detect and classify the main *arguments*, also called *slots*, of the prompt. Solving these two tasks is commonly referred to as *spoken language understanding* (SLU) or *slot and intent detection* (SID).

The xSID corpus is already available in several low-resource and non-standard varieties (Aepli et al., 2023; Winkler et al., 2024) and consists of a text genre for which dialectal productions are natural. We have translated the English sentences of xSID into standard Norwegian Bokmål and into the dialects of eight native speakers of Norwegian who regularly write in these dialects. The slot and intent annotations were then semi-automatically transferred to the Norwegian translations.

The resulting dataset, which we call NoMusic (*NOrwegian MUlti-dialectal Slot and Intent detection Corpus*), has the following particularities compared to existing Norwegian dialect resources:

- It is a multi-parallel corpus, i.e., all translations have the same number of sentences with the same meanings.
- It is a natively written resource and does not consist of transcribed speech.
- It is openly available, as all the translations are created on purpose within the project.[1]

The corpus can be used for various purposes, both in dialectology and natural language processing, e.g.:

- to evaluate the robustness and cross-lingual and cross-lectal transfer capabilities of SLU systems, thanks to the slot and intent labels,
- to identify dialect-specific expressions,
- to investigate digital writing practices,
- to enable machine translation between different varieties of Norwegian.

In the following sections, we describe the data and the annotation process, and provide analyses of the observed linguistic variation.

2 Related Work

2.1 Dialect Corpora for Norwegian

The Norwegian language has two officially established written standards: Norwegian Bokmål and Norwegian Nynorsk. Bokmål is the more utilized of the two in terms of speakers, and is historically based on written Danish.

[1]The NoMusic dataset is integrated into the xSID repository https://github.com/mainlp/xsid, but it is also available on https://github.com/ltgoslo/NoMusic.

Proceedings of the Eleventh Workshop on NLP for Similar Languages, Varieties, and Dialects (VarDial 2024), pages 107–116

While there are cases of earlier dialectal writing, general acceptance of dialects in increasingly formal settings began in the 1970s (Bull et al., 2018, 235-238). Dialects are thus less stigmatized, even in writing, for example in social media.

Norwegian dialects have been researched both from dialectological and computational angles, and several textual resources have been created in recent years. Traditional dialectological corpora such as the Nordic Dialect Corpus (NDC, Johannessen et al., 2009)[2] or the LIA Norwegian Corpus (Hagen and Vangsnes, 2023)[3] typically consist of transcriptions of interviews conducted with a large number of informants. This setup does not lead to directly comparable texts because the different interviews will be of different lengths, cover different topics and contain different linguistic structures. Also, these transcriptions are typically made by trained annotators according to relatively strict guidelines; the resulting written representations are often quite different from "real-world" dialect writing, as they are meant to faithfully represent the spoken language, rather than the way users would write their own dialect in everyday communication. For example, NDC contains Bokmål glosses and phonemic spellings, but these do not necessarily match how the users of a particular dialect spell.

On the other hand, recent data collection efforts such as NorDial (Barnes et al., 2021, 2023) focused on identifying and annotating written dialect posts in social media. This does not address the problem of comparability, but even introduces other challenges: it is difficult to obtain a dense coverage of the different dialects used in Norway, and the resulting dataset may not be made publicly available due to copyright restrictions. It remains to be seen to what extent projects such as the Nordic Tweet Stream (NTS, Laitinen et al., 2018) provide a viable workaround to copyright and licensing questions.

2.2 Multi-Dialectal Corpora

A relatively common alternative strategy to create multi-dialectal corpora consists in asking dialect speakers to translate texts into their variety, either from the standard variety or from a third language like English.

The MADAR Corpus of Arabic Dialects (Bouamor et al., 2018) illustrates this approach: the authors use a fixed set of English sentences and have them translated by native Arabic dialect speakers into their variety. They use the Basic Travel Expressions Corpus (BTEC, Takezawa et al., 2007) as a starting point and obtain translations of 2000 sentences into 25 Arabic dialects.

The SwissDial corpus (Dogan-Schönberger et al., 2021) follows a similar strategy, resulting in 2500 sentences in 8 Swiss German dialects. The corpus contains both audio recordings and transcripts, making it suitable for speech processing applications. Moreover, the data is annotated on sentence level with topic and code-switching information.

In a related effort, the xSID corpus[4] (van der Goot et al., 2021; Aepli et al., 2023; Winkler et al., 2024) has been created to support the development of multilingual dialog systems. It consists of prompts to digital assistants and is annotated with intents and slots. The 800 prompts in xSID are originally in English and have been translated to 12 major languages and 4 low-resource varieties or dialects (as of version 0.5, with the latter being Bavarian German, South Tyrolean German, Swiss German, and Neapolitan). In contrast to the BTEC corpus used for MADAR, the xSID source data is freely available and provides additional sentence-level (intents) and chunk-level (slots) annotations for the SID task.

The DIALECT-COPA shared task held at VarDial 2024 (Chifu et al., 2024)[5] is based on a similar approach: it contains translations of the English causal commonsense reasoning corpus COPA (Roemmele et al., 2011; Ponti et al., 2020) into various South Slavic languages and dialects.

Most of the resources cited above are created by translation from (American) English. This can be problematic because the translators may not be sufficiently familiar with the North-American cultural references (music styles, holiday destinations, etc.) and/or linguistic expressions (e.g. date and time formats, imperial measurements) used in the original data. Furthermore, non-professional translators are prone to producing translationese, which can be perceived as unnatural and not representative of spontaneous dialect writing. We are aware of these limitations, but nevertheless find it the most practical and effective approach to create multi-dialectal annotated resources.

[2] https://tekstlab.uio.no/scandiasyn/
[3] https://tekstlab.uio.no/LIA/norsk/
[4] https://github.com/mainlp/xsid
[5] https://sites.google.com/view/vardial-2024/shared-tasks/dialect-copa

English	Set a reminder to go to the grocery store later
Danish	Sæt en påmindelse om at gå i supermarkedet senere
Bokmål	Sett på en påminnelse om å gå i butikken etterpå
A1	Minn mæ på at æ skal dra på butikken seinere.
A2	Sett enn påminnelse om å fære tel butikken seinar.
A3	Sett en alarm for å da te matbutikken seinere
A4	Sett en påminnelse om å gå te matbutikken seinar
A5	Sett en påminnelse for å gå t butikken seinar
A6	Sett en påminnelse om å stikke på butikken seinere.
A7	Sett på en påminnelse om å gå t butikken seinare
A8	Lag ein påminnelse om å gå på butikken seinere

Table 1: Examples of translations. The Danish translation is already part of xSID. The Norwegian dialect annotators are numbered *A1* to *A8* from North to South.

2.3 Spoken Language Understanding Datasets

The xSID corpus represents one of the few efforts to provide non-English datasets for the SLU/SID task. However, it only provides manually created validation and test sets. Training sets for non-English languages are available, but created automatically by machine translation. The only currently available SLU dataset that covers Norwegian is MASSIVE (FitzGerald et al., 2022). It provides training, validation and test sets for 51 languages, among which standard Norwegian Bokmål. The slot and intent label sets differ between xSID and MASSIVE, and we leave it to future work to investigate to what extent the two annotation standards can be harmonized meaningfully.

The NoMusic corpus is, to our knowledge, the first SID dataset that provides multiple alternative formulations of the same queries.[6] The alternatives show dialectal variation, but also different lexical and syntactic choices (see Section 4). This variety opens up new avenues for making both the training and the evaluation of SLU systems more robust.

3 Data and Annotation

The xSID corpus provides a development set of 300 sentences and a test set of 500 sentences. The NoMusic dataset consists of annotated translations of these sentences. It is produced in three phases:

1. Translate the English xSID sentences to standard Norwegian Bokmål and to the Norwegian dialects.
2. Annotate the Bokmål sentences with slots, using the English sentences as guides.
3. Annotate the dialectal sentences with slots, using the Bokmål sentences as guides.

The following sections describe these phases in detail.

3.1 Translation

We used the English xSID dataset as a starting point and produced translations to standard Norwegian Bokmål and to eight Norwegian dialects.[7] The dialect translations were made by university students who declared that they regularly write in their dialect.

The Bokmål translation was produced by one of the authors of the paper. While some dialects speakers normally use Nynorsk, the other written Norwegian norm, the choice of Bokmål is purely practical, and it is used as a means for more easily transferring the slot and intent labels, as well as functioning as a meta-language to which to compare the dialectal forms.

The translations were produced by editing .tsv files in a shared GitHub repository. The annotators had access to GitHub issues where they could discuss potential problems. An example sentence with all available translations is shown in Table 1.

3.2 Translation Guidelines

The translators were given simple instructions on how to translate, but were otherwise not controlled.

[6] The ITALIC dataset (Koudounas et al., 2023) provides audio files and transcripts of SLU prompts in various regional varieties of Italian, but it is only annotated with intents, not slots.

[7] Two additional dialect translations are in progress at the time of writing and will be added to the dataset when completed.

These guidelines mostly followed the ones from the xSID project, but deviated in some respects discussed here.

Time The xSID guidelines note that some languages that do not have pm/am equivalents might need to translate cases such as *7 pm* to *7 in the evening*. Our annotators were not given specific notes on these translations, but were generally asked to translate into natural written dialect. This has resulted in some variation. The 24-hour clock is widely used in Norway, but in the spoken language, the 12-hour clock is also used if the times are unambiguous. We see that three different strategies have been used by our annotators in these cases: 1) adding a temporal adverb (*om morran* 'in the morning', *ettermiddag* 'afternoon'), 2) leaving the time ambiguous, which often means directly translating the English time without pm or am. or 3) converting the time to the 24-hour clock (*4 pm* → *klokka 16* '16 o'clock'). At least 6 of the annotators convert to the 24 hour clock to some degree. There are also instances of confusion between am and pm in the translations, for example in one case 5pm was interpreted as 05:00 by one annotator.

Named Entities In the xSID guidelines it is noted that named entities are not to be translated, except for place names. While this has been the general tendency in our dataset, annotators were asked to translate the names of movies when an established Norwegian title exists, but otherwise not. There is also some confusion for certain named entities that contain translatable content, such as whether the *Theatres* part of *Cobb Theatres* should be translated or not. Some annotators have translated certain titles even in cases where there is no established Norwegian name.

Grammatical Mistakes Grammatical mistakes should be kept in the translations if possible, according to the xSID guidelines. We believe that this would have been difficult, as it is not obvious to decide how a certain mistake might map from one language to another. Our annotators were not specifically asked to keep mistakes from the English sentences. However, as discussed below, the informal nature of the writing has led to some spelling mistakes that are not reflections of the original English. It is difficult to distinguish between cases when deviations from normative writing are conscious representations of dialect, and when they are simply unintentional.

Capitalization and Punctuation Annotators were not asked to correct capitalization or punctuation, but were also not explicitly asked to ignore it; rather, they were asked to follow their usual dialect writing habits. As a result, we see different tendencies among annotators. Some diligently add it where needed, while some allow for variation in their translations. Table 1 is an example of this variation.

Abbreviations While there are generally few abbreviations, there are some spelling conventions that in the written language are similar to abbreviations, but that would not be detectable in the spoken language. The xSID guidelines discourage abbreviations that are not 'common in fluent discourse.' We see examples of abbreviations such as *min* 'minute', which might also be read in its abbreviatiated form, and we also commonly note the usage of shortened spelling conventions like writing *d* for Bokmål *det* 'it', or *t* for Bokmål *til* 'to', similar to the usage of *u* for *you* in English.

Avoiding Direct Translations The xSID guidelines point out that it is not necessary to directly translate certain things, exemplified by the ditransitive usage of *play*. We believe that this has been covered by asking the annotators to translate into natural-sounding dialectal Norwegian. Another example is the translation of the English polite marker *please*, which has been translated into a variety of ways in the data.

Possessive Determiners The xSID guidelines note that possessive determiners should be preserved and translated whenever possible, but the annotators were not explicitly asked to do this. Norwegian generally uses fewer possessive determiners than English. For example, four dialect and the Bokmål translations use a variation of *where I am now* or *here* to translate 'my current position': *her e e, her, der ej e no, her eg e nå, der jeg er nå*, perhaps due to a direct translation sounding a bit stilted.

3.3 Translator Demographics

Figure 1 shows the origin of the dialect translators (marked with *A1* to *A8*) in relation with the four major Norwegian dialect areas. It can be seen that three of the four main dialectal areas are represented in NoMusic, but that we lack translations from dialects representing Eastern Norwegian. This absence can be explained by there being less

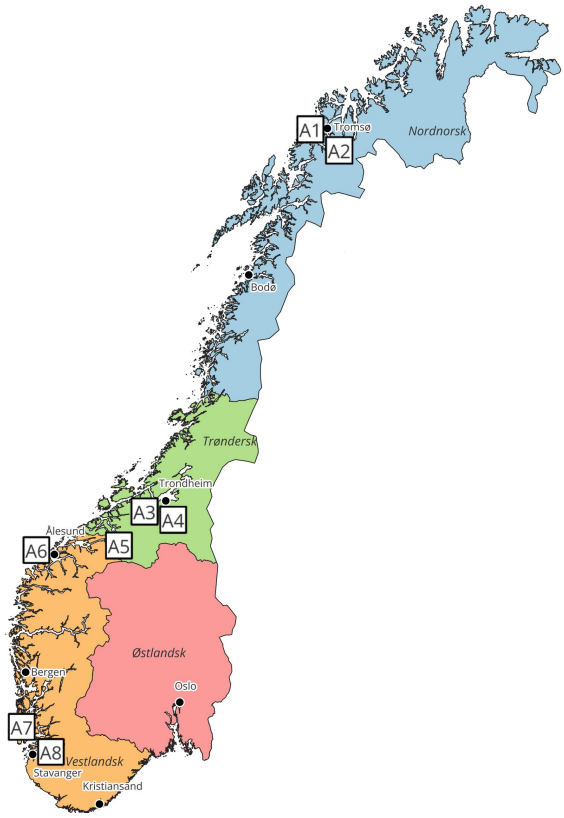

Figure 1: Map of Norway, with the four major dialect areas and the origins of the eight dialect annotators (*A1* to *A8*).

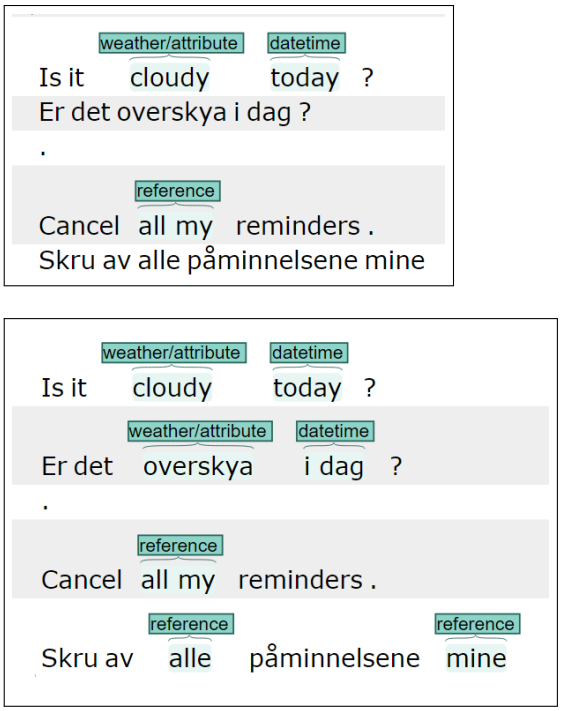

Figure 2: INCEpTION annotation interface showing the English-to-Norwegian annotation transfer. The upper part shows the initial state with pre-annotated English and unannotated Norwegian, the lower part shows the completed Norwegian annotations. Note the different number of labels.

perceived difference between the spoken language and the written language in Eastern Norway, as Bokmål is often associated with *Standard Østnorsk* 'Standard Eastern Norwegian', a commonly taught spoken variety.[8] Slåen (2022) describes written dialectal usage in the Northern reaches of the Eastern Norwegian dialectal area, but the tendency may be lower in and around Oslo.

As can be seen on the map, 2 translators speak Northern dialects, 3 central (Trøndersk), and 3 Western dialects. We had 6 female and 2 male translators. 6 translators were in the age range 20-24 and 2 in the range 25-29; all of them were university students on Bachelor's or Master's level.

3.4 Slot and Intent Annotations

Once the sentences are translated, they need to be labeled with slots and intents. Each sentence has a single intent, and the intent is not supposed to change across languages. Therefore, we automatically transfer the intent labels from English.

The slot labels are annotated manually in two steps, using the same procedure as for the original xSID corpus. In the first step we annotate the Bokmål version, using the annotated English sentence as a guide for each sentence. In the second step, the dialectal versions are annotated, using the already annotated Bokmål version as a guide.

We use the INCEpTION (Klie et al., 2018) platform for transferring the slot annotations. For the English-to-Bokmål step, we interleave annotated English sentences with their unannotated Bokmål translations. The annotation process is illustrated in Figure 2.[9] We note how the Norwegian syntax can lead to differences in the number of slot labels. In this case, the xSID guidelines state that consecutive reference labels specifically should be annotated as a single chunk, but as there are no discontinuous spans in the English data, we annotate them as two

[8] https://www.sprakradet.no/svardatabase/sporsmal-og-svar/oslodialekten/

[9] In order to upload the pre-annotated English sentences along with Bokmål, we merged the two and uploaded the resulting .txt file using the *plain text (one sentence per line)* setting. We then downloaded the UIMA CAS XMI file, which is INCEpTION's native format. Using the dkpro-cassis library (https://pypi.org/project/dkpro-cassis-tools/#description), we then added the English slot spans from the existing .conll files, and uploaded the resulting .XMI file. Annotations were added in a single token level layer.

Figure 3: Annotation of the dialect translations. Note how differences in spelling of *i dag* 'today' causes slight differences in labeling.

separate labels.

A similar process is used for the Bokmål-to-dialect annotation transfer step: the annotated Bokmål sentence is presented on top as a guide, with all dialectal translations following. See Figure 3 for an example.

4 Analysis

The dialectal translations differ in various respects from each other and from the standard version. In this section, we discuss different types of variation and their prevalence in the dataset, before briefly looking at how some of these features present themselves in the Nordic Dialect Corpus (NDC).

4.1 Variation in Translation

Unsurprisingly, the translations are largely similar in terms of word lengths and type-token ration, as reported in Table 2. We see that some annotators (A2, A6) have slightly longer sentences. The most striking difference is perhaps the lower number of types in English, but this could easily be attributed to the slightly higher morphological variation in Norwegian.

Annotator	Tokens	Types	Sent. length
A1	6200	1337	7.74
A2	6526	1360	8.15
A3	6282	1365	7.84
A4	6054	1346	7.56
A5	5955	1310	7.43
A6	6546	1350	8.17
A7	6004	1379	7.5
A8	6086	1366	7.6
Bokmål	6310	1392	7.88
English	6177	1245	7.71

Table 2: Tokens, types and average sentence lengths for the annotators, the Bokmål translations, and the original English.

4.2 Linguistic Variation

While there are many clear dialectal differences between the translators, that is not to say that all these differences are due to dialectal variation. For many sentences there are several possible translations, and there are also lexical or syntactic choices that do not necessarily have to be dialect-specific. For example, in Table 1, the verb 'to go' is expressed by *å gå, å dra, å fære* or *å stikke*, and 'grocery store' is translated by *butikken* or *matbutikken*. Before looking at dialectal features in the dataset, we discuss some more general features.

Spelling Annotators were asked to translate to their own dialect in a natural way. This has led to varying degrees of written expressions. In dialectal writing, the written forms naturally deviate from the established written norms, namely Bokmål or Nynorsk, but we would typically not expect deviations that cannot be explained by the dialectal features of the writer. We do see what we consider non-dialectal spelling deviations, or what would be spelling mistakes in a prescriptive setting. The frequency of these vary from annotator to annotator. In practice, this means that the corpus has some features of user-generated language that are not unique to dialectal writing.

Pronunciation Spelling One crucial difference between spoken dialect and written dialect is that not all words show indications of being associated with a dialect, and many words are left in their Bokmål or Nynorsk spelling, despite being pronounced differently from how most speakers would pronounce the normed spellings. In the NorDial

	A1	A2	A3	A4	A5	A6	A7	A8	NB
A1	568								
A2	239	612							
A3	319	235	573						
A4	317	224	313	609					
A5	310	233	294	334	589				
A6	312	228	273	276	265	570			
A7	280	194	258	291	281	277	582		
A8	255	192	223	264	246	270	315	632	
NB	313	217	271	292	316	301	304	287	675

Table 3: Lexical overlap between the dialect translations and the Bokmål (NB) translation. Words contained in the English dataset (mostly titles and names) are removed from the comparisons.

	Pearson's r	p-value
A1	−0.5329	0.1738
A2	−0.6885	0.0590
A3	−0.5516	0.1563
A4	−0.5792	0.1324
A5	−0.4962	0.2111
A6	−0.4454	0.2688
A7	−0.6159	0.1040
A8	−0.6069	0.1106

Table 4: Pearson correlation coefficients between lexical overlap and geographical distance.

corpus, the authors find that some sentences only contain a few words indicating dialect, although in spoken language all words would (Barnes et al., 2021). While most function words are written according to the pronunciation of a given dialect, many content words are not, despite obviously not following pronunciation rules. However, this varies from annotator to annotator in our dataset. An example is the word 'restaurant', whose spelling is kept in some cases by at least 6 annotators, while some use the spelling *resturant* or *resturang*. In the NDC, the spellings are *r[e/æ]s(s)t[u/o]ran(n)g(g)*.

Avoidance of Direct Translations As mentioned earlier, another source of variation is avoidance of direct translations, which can lead to syntactic and lexical variation. For example, when talking about weather predictions, it is quite common to use the auxiliary verb *skulle*, which indicates a planned action or a prediction, but it is in some cases also natural to use a more neutral feature with the composite auxiliary *komme til* 'will'. Both options are available in several dialects, and even the same user might alternative between these.

4.3 Lexical Overlap and Dialectal Features

We now examine the translations in terms of dialectal features and lexical overlap. Table 3 presents an overview of the lexical overlap between the translations. The diagonal shows the total number of types, reported in Table 2. We would expect annotators who come from dialectal areas in close proximity to exhibit higher overlap.

Table 4 shows the Pearson correlation coefficients between the lexical overlap (Table 3) and the geographical distances between the translators' origins. Correlations are computed for each annotator separately. The correlation coefficients indicate moderate to strong correlations,[10] but the p-values are too high to draw meaningful conclusions.

The clearest dialectal differences are observed in morphology. We will have a brief look at verbal, nominal and adjectival morphology, while acknowledging that this is only part of what constitutes dialectal variation in the dataset. Where attestations can be found, we look up corresponding forms in the NDC interface to inspect their distributions. Queries are done in Bokmål, and the reported phonological forms are compared to our dialectal writing.

Verbal Morphology One thing to observe in terms of verbal morphology is the infinitive. This is an oft-used dialectal feature, based on whether the dialect has infinitives (for consonant stem verbs) in -a, -e, -Ø (apocope) or a mix of these. For our annotators, we observe 5 patterns: infinitives ending in -e only (A1, A2 and A6), in -a only (A8), no ending (-Ø) (A5, A4), mixed -e and no ending (A3) and mixed -a and no ending (A7). Notably for A7 it seems like the apocope is only found in the verb *å vær*, but it is both consistent and frequent. According to the presentation of infinitives in Mæhlum and Røyneland (2023, p. 180), A1, A2 and A6 are all from typical e-infinitive areas, and A8 is from a typical a-infinitive area. A4 and A5 are theoretically both further south than the area typically associated with pure apocope. A3 is in the area for mixed infinitive, but A7's position in the South-West does not explain the form *å vær*. However, in the NDC, *vær* as an infinitive form is not infrequently observed in South and West Norway.

Nominal Morphology While there are not enough nouns to create a full overview of the writ-

[10]The correlations are negative because lexical similarity is compared with geographic distance.

ers' morphological systems, there is enough to give us indications. First of all, we get an impression of the gender system. Normally, both written Norwegian norms, Bokmål and Nynorsk, allow for a three-gender nominal inflection system, but to varying degrees. A three-gender system is obligatory in Nynorsk, while in Bokmål it is possible to conflate the masculine and feminine classes to a common gender (nor. *felleskjønn*). We see that all dialects mark feminine nouns to some extent, as all dialects use the feminine-specific singular definite marker -a (or -å) at least in some words (*boka 'the book', låta 'the tune', bogå 'the book'*), but not in all (*vermeldingen, vermeldinga* 'the weather forecast'). The use of the indefinite singular article *ei* 'a, an' is less frequent, as is also the case in Bokmål. While the masculine singular is invariably the same as in Bokmål, another difference between feminine and masculine nouns appears in the plural. Where some writers in our corpus have the same forms as in Bokmål for both genders (*filmer, stjerner* 'movies, stars'), we see that some writers have variant forms, which are still the same (*filma, stjerna*), while some make a distinction (*filma, stjerne; filmar, stjerner*). Some dialects have apocope in the plural definite (*filman*). Some annotators have the same forms for masculine and neuter nouns (*filma, minutta*), while others have the typical zero-ending that we also see in Nynorsk (*filma, minutt*).

Adjectival Morphology One notable feature for adjectives, is whether the neuter suffix -t is added to adjectives in -ig. This is done by the translator from Stavanger (A8), as in *tidligt* 'early'. This is confirmed to be a regional feature by the NDC, where corresponding forms are found in the area around Stavanger but not elsewhere in the country. We also see variation in the comparative forms, where three forms are found: *-ere (kaldere)* 'colder', *-are (kaldare)* and an apocopized version, *-ar (kaldar, kjørligar)* . While there are not many attestations of *kaldere*, we see that all attestations with the *-ere* ending are in Eastern Norway.

Lexicon While many words show clear dialectal influence, there are few cases where the annotators' lexical choices are markedly different from the standard language. One such example is the use of *bli å* lit. 'become to' as a future auxiliary.

Function Words Much of the variation seen between the translated material is in terms of function words: prepositions, pronouns, and determiners.

	I	ME	HOW	SOME	TO
A1	æ	mæ	kordan	nokka	til
A2	æ	mæ	kordan	nåkka	tel
A3	æ	mæ	koss	nokka	te
A4	æ	mæ	koss	nåkka	te
A5	e	me	koss	nokka	til
A6	ej	mej	kordan	nokke	til
A7	eg	meg	koss	noe	t(e)
A8	eg	meg	kordan	någe	te

Table 5: Selected pronouns and function words used by the different annotators.

In Table 5, we see five selected words that illustrate some of the variation between the translators. Looking at the pronominal variation, we get an idea of how distinctive some of these features are. The form *ej* (A6) 'I', is associated with an area between Ålesund and Bergen in the NDC, indicating that this is a quite distinguishing feature of A6's dialect. Otherwise it is only attested once close to Mo i Rana. Among the other words for *I*, both *æ* (A1-A4) and *e* (A5), are quite widespread in spoken Norwegian as reported in the NDC. The form *eg* (A7, A8) is more associated with the West, and is not found along the border to Sweden in the East. For the oblique forms, *mæ* (A1-A4) is quite widespread, except in the West and upper central areas, and *mej* (A6) is only registered in two locations: one on the Trøndersk/Vestnorsk border, and one in the Trøndersk area. The interrogative *Kordan* 'how' is mostly associated with Western and Northern Norwegian, while *koss*, also 'how', is associated with Southern, Central and upper Western Norway. The determiner *någe* 'some' is heavily associated with the Stavanger area, and is not found outside it except one attestation in Tromsø. *Noe* 'id.' is quite widespread, but not in the upper West. *Nokka* is associated with Trøndersk and Northern Norwegian, and a small cluster in the south in the NDC, while finally *nokke* is associated with the west.

4.4 Phonological Features

As the translators all report that they use dialectal writing in their daily lives, we see the translations as representative of at least some part of the written dialect of the area the translator represents, but this does not tell us to what degree the written language represents the spoken dialect of that area. However, some of these features can be inspected using NDC.

For example, a commonly used dialectal feature is the voicing of the ungeminated plosives /p/, /t/ and /k/ to /b/, /d/ and /g/. We see examples of this in our dataset in forms such as *søga* (Bokmål *søke*) 'search' and *bogå* (Bokmål *boka*) 'the book'. In NDC, forms of *søke* with voicing are only found in an area surrounding Kristiansand, while voiced forms of *boka* are found between Haugesund and Kristiansand.

5 Conclusion

We present a dataset of written dialectal Norwegian, which reflects various dialectal phenomena and is also annotated with slots and intents. The utterances are translations of the English validation and test sets of the xSID corpus (van der Goot et al., 2021).

Limitations

As discussed in Section 3.3, the geographical coverage of the translators is uneven, with Eastern dialects not represented at all in the corpus. This is due to linguistic factors, as discussed, and also to contingent factors related to the sample of qualified and interested students available during the project duration. We will consider extending the corpus if annotators from not yet covered areas become available.

Furthermore, as discussed in Section 2.2, the annotation workers are not professional translators and may find it difficult to produce natural and correct dialect writing in a translation setup. Moreover, certain cultural references and named entities may not be known well enough by our translators.

Finally, slot and intent detection models are typically applied to speech data in conjunction with an automatic speech recognition system. It could thus be useful to pair the dialectal transcripts with recorded speech. We currently do not offer speech recordings because our main goal was to create a resource for *written* dialectal Norwegian, but we may consider extending the dataset towards speech in the future.

Ethical Considerations

The translators were hired as student assistants and paid for the effective hours spent on the translation task (typically between 15 and 20 hours, not including slot annotation), according to the official salary schemes in use at the University of Oslo. The participation in the translation task was voluntary, and all translators agreed in writing that their productions may be publicly shared under the CC-BY-SA 4.0 licence.[11]

The English data used as source material is curated and does not contain any harmful content, to our knowledge.

Acknowledgements

The NoMusic project was granted funding from the TekstHub initiative at the University of Oslo.

We thank the many annotators who have contributed to this project: Andželika Andruškaite, Thomas Heim, Sanne Lima, Ada-Marie Sørensen Sneve, Elias Lynum Ringkjøb, Tonje Sandanger, Lilja Charlotte Storset, Ulrikke Strømsvold Tveit and Snorre Åldstedt.

We also thank Anders Næss Evensen for help with the system for uploading pre-annotated files to INCEpTION.

References

Noëmi Aepli, Çağrı Çöltekin, Rob Van Der Goot, Tommi Jauhiainen, Mourhaf Kazzaz, Nikola Ljubešić, Kai North, Barbara Plank, Yves Scherrer, and Marcos Zampieri. 2023. Findings of the VarDial evaluation campaign 2023. In *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*, pages 251–261, Dubrovnik, Croatia. Association for Computational Linguistics.

Jeremy Barnes, Petter Mæhlum, and Samia Touileb. 2021. NorDial: A preliminary corpus of written Norwegian dialect use. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 445–451, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.

Jeremy Barnes, Samia Touileb, Petter Mæhlum, and Pierre Lison. 2023. Identifying token-level dialectal features in social media. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 146–158, Tórshavn, Faroe Islands. University of Tartu Library.

Houda Bouamor, Nizar Habash, Mohammad Salameh, Wajdi Zaghouani, Owen Rambow, Dana Abdulrahim, Ossama Obeid, Salam Khalifa, Fadhl Eryani, Alexander Erdmann, and Kemal Oflazer. 2018. The MADAR Arabic dialect corpus and lexicon. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Tove Bull, Espen Karlsen, Eli Raanes, and Rolf Theil. 2018. *Norsk språkhistorie*, volume 3. Novus, Oslo.

[11] https://creativecommons.org/licenses/by-sa/4.0/

Adrian Chifu, Goran Glavaš, Radu Ionescu, Nikola Ljubešić, Aleksandra Miletić, Filip Miletić, Yves Scherrer, and Ivan Vulić. 2024. VarDial evaluation campaign 2024: Commonsense reasoning in dialects and multi-label similar language identification. In *Eleventh Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2024)*, Mexico City, Mexico. Association for Computational Linguistics.

Pelin Dogan-Schönberger, Julian Mäder, and Thomas Hofmann. 2021. SwissDial: Parallel multidialectal corpus of spoken Swiss German. *CoRR*, abs/2103.11401.

Jack FitzGerald, Christopher Hench, Charith Peris, Scott Mackie, Kay Rottmann, Ana Sanchez, Aaron Nash, Liam Urbach, Vishesh Kakarala, Richa Singh, Swetha Ranganath, Laurie Crist, Misha Britan, Wouter Leeuwis, Gokhan Tur, and Prem Natarajan. 2022. Massive: A 1m-example multilingual natural language understanding dataset with 51 typologically-diverse languages.

Kristin Hagen and Øystein A. Vangsnes. 2023. LIA-korpusa – eldre talemålsopptak for norsk og samisk gjort tilgjengelege. *Nordlyd*, 2(47):119–130.

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*, volume 4 of *NEALT Proceedings Series*.

Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. 2018. The inception platform: Machine-assisted and knowledge-oriented interactive annotation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 5–9. Association for Computational Linguistics. Event Title: The 27th International Conference on Computational Linguistics (COLING 2018).

Alkis Koudounas, Moreno La Quatra, Lorenzo Vaiani, Luca Colomba, Giuseppe Attanasio, Eliana Pastor, Luca Cagliero, and Elena Baralis. 2023. Italic: An italian intent classification dataset.

Mikko Laitinen, Jonas Lundberg, Magnus Levin, and Rafael Messias Martins. 2018. The Nordic Tweet Stream: A dynamic real-time monitor corpus of big and rich language data. In *Proceedings of the Digital Humanities in the Nordic Countries 3rd Conference*, pages 349–362.

Brit Mæhlum and Unn Røyneland. 2023. *Det norske dialektlandskapet: innføring i studiet av dialekter*. Cappelen Damm akademisk, Oslo.

Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. XCOPA: A multilingual dataset for causal commonsense reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2362–2376, Online. Association for Computational Linguistics.

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S Gordon. 2011. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *AAAI Spring Symposium on Logical Formalizations of Commonsense Reasoning*.

Mari Slåen. 2022. "jeg synes det er mest vanlig å skrive slik jeg prater" en sosiolingvistisk studie av elevers dialektale skriving og holdninger i gudbrandsdalen. Master's thesis, Norwegian University of Science and Technology.

Toshiyuki Takezawa, Genichiro Kikui, Masahide Mizushima, and Eiichiro Sumita. 2007. Multilingual spoken language corpus development for communication research. In *International Journal of Computational Linguistics & Chinese Language Processing, Volume 12, Number 3, September 2007: Special Issue on Invited Papers from ISCSLP 2006*, pages 303–324.

Rob van der Goot, Ibrahim Sharaf, Aizhan Imankulova, Ahmet Üstün, Marija Stepanović, Alan Ramponi, Siti Oryza Khairunnisa, Mamoru Komachi, and Barbara Plank. 2021. From masked language modeling to translation: Non-English auxiliary tasks improve zero-shot spoken language understanding. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2479–2497, Online. Association for Computational Linguistics.

Miriam Winkler, Virginija Juozapaityte, Rob van der Goot, and Barbara Plank. 2024. Slot and intent detection resources for Bavarian and Lithuanian: Assessing translations vs natural queries to digital assistants. In *Proceedings of The 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*. Association for Computational Linguistics.