

Slot and Intent Detection Resources for Bavarian and Lithuanian: Assessing Translations vs Natural Queries to Digital Assistants

Miriam Winkler[▲] Virginija Juozapaityte[Ⓢ]
Rob van der Goot[Ⓢ] Barbara Plank^{▲ⓈⓈ}

[▲] MaiNLP lab, Center for Information and Language Processing, LMU Munich, Germany

[Ⓢ] Munich Center for Machine Learning (MCML), Munich, Germany

[Ⓢ] Department of Computer Science, IT University of Copenhagen, Denmark

[▲]mi.winkler@campus.lmu.de [▲]b.plank@lmu.de [Ⓢ]robv@itu.dk

Abstract

Digital assistants perform well in high-resource languages like English, where tasks like slot and intent detection (SID) are well-supported. Many recent SID datasets start including multiple language varieties. However, it is unclear how realistic these translated datasets are. Therefore, we extend one such dataset, namely xSID-0.4 (van der Goot et al., 2021a; Aepli et al., 2023), to include two underrepresented languages: Bavarian, a German dialect, and Lithuanian, a Baltic language. Both language variants have limited speaker populations and are often not included in multilingual projects. In addition to translations we provide “natural” queries to digital assistants generated by native speakers. We further include utterances from another dataset for Bavarian to build the richest SID dataset available today for a low-resource dialect without standard orthography. We then set out to evaluate models trained on English in a zero-shot scenario on our target language variants. Our evaluation reveals that translated data can produce overly optimistic scores. However, the error patterns in translated and natural datasets are highly similar. Cross-dataset experiments demonstrate that data collection methods influence performance, with scores lower than those achieved with single-dataset translations. This work contributes to enhancing SID datasets for underrepresented languages, yielding NaLiBaSID, a new evaluation dataset for Bavarian and Lithuanian.

Keywords: Less-Resourced Languages, Dialects, Multilinguality, Corpus (Creation, Annotation, etc.)

1. Introduction

In the growing landscape of natural language understanding (NLU), it is important to not leave low-resource languages behind—since there exists an immense linguistic diversity that spans the globe. Many native speakers of such language variants may have the desire to use language technology like virtual assistants in their mother tongue.¹ Therefore, this paper delves into the realms of Slot and Intent Detection (SID), illustrated in Figure 1. We focus on two languages—motivated by access to native speakers—Bavarian (Central Bavarian (Vergeiner and Bülow, 2023)) and Lithuanian. Bavarian, a standard German dialect, comes from the heart of Europe, while Lithuanian, a Baltic language, finds its stronghold in the northeastern corners of the continent.

Both language variants have limited speaker populations, which are often excluded from multilingual projects. For Bavarian, Rowley (2011) estimates the speaker population to around 11M, while Ethnologue (Eberhard et al., 2023) cites over 14M. For Lithuanian, Ethnologue estimates only approximately 2.7M speakers. Compared to languages frequently used in NLP such as English (379M native speakers), German (75M na-

¹As for example found in a survey of dialect speakers by Blaschke et al. (2024).




English 	What will the weather be in New York city this week (de-ba) Wia werds Weda in New York City dera Woch (lt) Koks oras bus šių savaitę Niujorko mieste
Bavarian 	Wos is grod für a weda in äding ? (en) What is the weather like in Altötting right now ?
Lithuanian 	Pateikt rytojaus orų prognozę Vilniuje . (en) Give tomorrow 's weather forecast for Vilnius .

Figure 1: Excerpts of the *de-ba* and *lt* translation datasets and natural *nat:de-ba* and *nat:lt* datasets. Slots: location and datetime. *Italic*: glossary. *Image Source*: wikipedia.org

tive speakers) or Spanish (485M native speakers) (speaker numbers according to Ethnologue), these two languages are less-resource languages.

The reason for studying these languages in a unified framework stems from the desire to compare a novel data collection approach that differs from previous practices in the field and from a fundamental question: *How well does slot and intent detection work on translated vs natural queries to digital assistants?* The limited availability of linguistic resources and research regarding these languages presents a compelling challenge and opportunity.

We provide novel datasets and zero-shot transfer results, providing a new perspective on cross-

lingual research and the robustness of SID for two language varieties. Through this, we aim to not only advance our understanding of SID but also contribute to enhance linguistic inclusivity.

Our Contributions:

- We present *NaLiBaSID*² (**N**atural **L**ithuanian and **B**avarian **S**ID), which contains new slot and intent detection evaluation datasets for Bavarian (*de-ba*) and Lithuanian (*lt*), created by manual translation and applying the translation and annotation schemes by xSID (van der Goot et al., 2021a).
- Additionally, in *NaLiBaSID* we collect natural datasets of utterances from native speakers: *nat:de-ba* for Bavarian and *nat:lt* for Lithuanian, to be able to evaluate on more realistic data.
- For Bavarian, *NaLiBaSID* further contains translations a part of the large MASSIVE (FitzGerald et al., 2022) dataset (*MAS:de-ba*) to Bavarian to evaluate the effect of transferring to a low-resource language without orthography in a cross-datasets setting.
- We evaluate the performance of cross-lingual language models on our translated and native data, to gauge the effect of having natural utterances versus translations for SID.

2. Related Work

There is a large body of work on SID, a task also referred to as task-oriented dialogue or spoken language understanding, e.g. Schuster et al. (2019); Xu et al. (2020); van der Goot et al. (2021a). Common use cases for such task-oriented dialogue systems as found in popular digital assistants are for example entertainment (music, videos), booking different kinds of establishments or looking up specific information. Slot detection and intent classification are the two sub tasks of NLU that are usually performed in a joint setup. Additionally, dialect systems often include dialogue state tracking, dialogue policy management and response generation tasks (Razumovskaia et al., 2022). The development of multilingual NLU systems, however, is greatly limited by the low availability of suitable data for less-represented languages. Most research on multilingual SID therefore focuses on transfer from high-resource languages via multilingual language representations.

²Available at: <https://github.com/mainlp/NaLiBaSID>

In the next part we focus on available datasets, while we refer the reader to a recent comprehensive survey by Razumovskaia et al. (2022) for details on methods.

2.1. Source Datasets

Our data is based on two different source datasets: xSID-0.4 (van der Goot et al., 2021a; Aepli et al., 2023) and MASSIVE (FitzGerald et al., 2022).

xSID-0.4 The basis for the translation was the English xSID (van der Goot et al., 2021a) data, which includes utterances from the Snips (Coucke et al., 2018) and Facebook (Schuster et al., 2019) datasets. xSID 0.4 is a cross-lingual evaluation dataset that covers 15 different languages, among these, South Tyrolean, which is also a southern Austro-Bavarian dialect. It consists of 800 utterances per language split into a development and test part and a large English training split (43,605 utterances). The sentences are annotated with a set of 16 intents and 34 slot labels.

Most utterances are distributed evenly across all intents. However, the “weather/find” intent has remarkably more utterances, while “snooze_alarm”, “time_left_on_alarm” and “modify_alarm” occurred only a few times. The exact intent distribution is a part of Table 2 and applies also to both the Lithuanian and Bavarian xSID dataset translations.

MASSIVE The MASSIVE 1.1 dataset by FitzGerald et al. (2022) is a large collection of multi-lingual SID data covering 52 languages that was collected from speaker utterances to virtual assistants. It is larger than xSID and consists of 859,092 sentences. For each language variation, the data set of 16,521 sentences per language is split into 2,974 test, 2,033 dev and 11,514 train utterances. The annotations include 60 different intents and 55 slot labels.

There is a difference in quality between xSID and MASSIVE: In contrast to xSID, where the data was manually translated by the authors themselves (van der Goot et al., 2021a), the data for MASSIVE was collected and translated from and by Amazon MTurk crowd workers which could have yielded to lower quality data when comparing to xSID. Each sentence was processed by two workers in two translation steps. Three other workers then judged the quality of the translations and annotations in the following quality assurance step.

2.2. Other Multilingual Datasets

In recent years, several multilingual dataset were proposed. They add diversity to the already available datasets. One of the biggest multilingual

Dataset	Translation Src	Native	Intents	Slots	# sents
<i>de-ba</i>	xSID	–	16	34	800
<i>lt</i>	xSID	–	16	34	800
<i>MAS:de-ba</i>	MASSIVE	–	14	27	2,021
<i>xMAS:de-ba</i>	MASSIVE+xSID	–	16	34	2,821
<i>nat:de-ba</i>	n/a	collected	16	26	315
<i>nat:lt</i>	n/a	collected	16	30	327

Table 1: Overview of all data in *NaLiBaSID* with an overview of the sources for translation vs native data collections.

datasets is MultiAtis++ (Xu et al., 2020). It is an extension of the Multilingual Atis dataset by Upadhyay et al. (2018) that adds six new languages to the three existing ones. This brings up the total number of available languages to nine. MultiAtis++ is a big step from the preceding multilingual NLU datasets like the Facebook dataset (Schuster et al., 2019) (English, Spanish, Thai) and the Multilingual Atis dataset by Upadhyay et al. (2018) (English, Hindi, Turkish) as mentioned before that only consist of three languages each. For a more complete overview of datasets, we refer to van der Goot et al. (2021a); Aepli et al. (2023), FitzGerald et al. (2022) and the SLU dataset surveys by Qin et al. (2021) and Razumovskaia et al. (2022).

2.3. Translationese

Translationese (Toury, 2012; Zhang and Toral, 2019; Laviosa, 1997) is a concept observed in machine translation that states that a trace of the source language remains in the sentence even after translation. This phenomenon occurs in all tasks that involve translations. Translationese also simplifies the sentences and the translations are usually written in a less complicated structure than the original, which could make it easier for some systems or task types to deal with the translated sentence (Bizzoni et al., 2020).

In this paper, we study the effect of translationese on slot and intent detection by comparing it against natural queries. It should be noted that another aspect that affects performance is that people ask different things in different parts of the world; however teasing these two aspects apart is non-trivial.

3. *NaLiBaSID* Data

The datasets we present are the xSID (van der Goot et al., 2021a) translations *de-ba* for Bavarian and *lt* for Lithuanian. Additionally, we translate a part of the MASSIVE (FitzGerald et al., 2022) data into Bavarian in the *MAS:de-ba* dataset. Together, the xSID and MASSIVE translations form the larger cross-datasets Bavarian dataset *xMAS:de-ba*. We evaluate both separate dataset

splits and the combined *xMAS:de-ba*. Furthermore, we present the natural datasets for the two languages: *nat:de-ba* and *nat:lt*. The acquisition, translation and annotation processes are explained in this Section.

All datasets, unless stated otherwise, follow the original xSID test and development set split of 60-40%. Both the Bavarian and Lithuanian translations are based on the translation guidelines by van der Goot et al. (2021a, Appendix F), recently also applied to extensions for Swiss German and Neapolitan (Aepli et al., 2023). For both language variants, we translate from the English original data portion. The annotations are done according to the annotation guidelines of the xSID dataset (van der Goot et al., 2021a, Appendix G).

An overview over the dataset statistics for the new evaluation sets is given in Table 1.

3.1. Translated Datasets

3.1.1. xSID Bavarian and Lithuanian Datasets

Translation and annotation is done by the authors of this paper; we opted for this option instead of crowd-sourcing, as we have more control, crowd-sourcing being not available for Bavarian/Lithuanian, and the fact that we can ensure higher quality translation and annotations are done by motivated native speakers. Furthermore, the utterances are relatively simple, and agreement for intent detection was shown to be high for untrained, motivated native speakers (0.924 Fleiss Kappa) as found in van der Goot et al. (2021a). We consider the good quality we achieved as the strength of our datasets. However, since the annotations and translations were carried out by the same people, regional and personal biases may be present in the datasets.

We have translated the utterances of the English xSID-0.4 (van der Goot et al., 2021a; Aepli et al., 2023) dataset in such a way that the Bavarian and Lithuanian versions (*de-ba* and *lt*) sound as fluid and natural as possible while trying to stay close the meaning of the original sentence (example in Figure 1). For the annotations, we have adopted the intents and slots of the xSID dataset as they

were used by [van der Goot et al. \(2021a\)](#). The sizes of the xSID translation datasets are the same for both languages and can be seen in Table 1. The test-dev split distributes the sentences into a test set of 500 and a development set of 300 sentences.

Translation While translating the Bavarian data, we encountered a difficulty that stems from the lack of a standard orthography for the dialect: There are many different ways to write a word and the variants differ according to region and the individual native speaker. The translated Bavarian data therefore only contains the personally preferred spellings from one of the authors and not multiple variations. We go into more detail on this in Section 3.2. Further challenges for Bavarian translation are numbers and time indications, which are accumulated in Appendix A.1.

There are also noteworthy key points that had an influence on the Lithuanian translations. It is particularly worthy to mention that there is no fixed word order in the Lithuanian language. Instead, the word functions are marked by cases. An example for this can be seen in Table 3. For this reason, the words and resulting annotations may appear in a different order than in the original sentence.

This is also interesting in terms of name mentions. For correct usage of names in the Lithuanian language, they need to be transcribed to the Lithuanian alphabet and an ending that allows the application of cases must be attached. There are some modern English names that occur in the English data, which are not translated into Lithuanian in order to match the other xSID datasets and follow the translation guidelines by [van der Goot et al. \(2021a\)](#). It is not grammatically correct yet often used in Lithuanian. More details on the adjustments, challenges and possible word orders are presented in Appendix A.2.

3.1.2. MASSIVE Bavarian Dataset

To explore the model performance in a cross-datasets setup, we additionally translated a portion of the MASSIVE ([FitzGerald et al., 2022](#)) data.

In total, 2,021 utterances were extracted across the MASSIVE test (366), dev (256) and train (1,399) sets. The final distribution of the intents in our *MAS:de-ba* dataset is given in Table 2. We provide details on the intent mapping between the xSID and MASSIVE data sources and further re-annotation information in appendix B. The test set contains 1,212 sentences, the development set 809 sentences.

We only translated a selection of the total MASSIVE sentences because only few of the MASSIVE intents overlap well with the xSID intents, and

Intent	MASSIVE	xSID
PlayMusic	588	63
weather/find	439	202
set_alarm	225	53
show_alarms	176	29
set_reminder	171	50
show_reminders	169	31
cancel_alarm	104	55
SearchScreeningEvent	59	60
AddToPlaylist	36	53
BookRestaurant	27	69
cancel_reminder	14	26
SearchCreativeWork	6	52
modify_alarm	4	1
snooze_alarm	3	5
RateBook	0	47
time_left_on_alarm	0	4

Table 2: Distribution over intents in annotated (sub)parts of MASSIVE and xSID.

MASSIVE is highly skewed towards the “PlayMusic” intent. If all MASSIVE “PlayMusic” utterances were translated, 36% of utterances would belong to this intent. MASSIVE also covers a much wider range of topics per intent, so we manually sorted out some sentences because they either fitted another xSID intent better or did not fit at all into an intent category of the xSID label set. Afterwards, we re-annotated the completely fitting sentences following the annotation guidelines of [van der Goot et al. \(2021a, Appendix G\)](#) for this process.

A part of the *MAS:de-ba* utterances is similar to the xSID sentences. However, some intents have varying levels of complexity across the two datasets. To name some examples, we show sentences in Figure 2. The “reminder” intents of the MASSIVE dataset often had a much higher level of complexity and detail in their sentences, whereas the “weather/find” utterances had less variety than the xSID counterparts.

3.1.3. Combined Bavarian Dataset

We additionally provide the combination of the separate xSID and MASSIVE splits as a single dataset: *xMAS:de-ba*. With this, we create the largest dataset for a low-resource language variant without orthography.

The combination brings up the size to 2,821 sentences in total (800 xSID; 2,021 MASSIVE; cf. Table 2). The test set consists of 1,694 and the development set of 1,127 sentences.

In Section 5, we provide results on *xMAS:de-ba* in Tables 4 and 5. We use this dataset to show the cross-dataset evaluation performance in a combined dataset. We expect the results to show a mean performance of the individual results of *de-*

“reminder/set_reminder”	
MASSIVE	remind me to call my daughter first thing on thursday to wish her happy anniversary
xSID	Remind me to call my sister tomorrow

“weather/find”	
MASSIVE	any signs of rain
xSID	how many inches of rain are we suppose to get tomorrow?

Figure 2: Example sentences illustrating the varying levels of complexity and detail between the MASSIVE and xSID sentences for the “reminder/set_reminder” and “weather/find” intents.

ba and *MAS:de-ba*, respectively.

This cross-dataset collection *xMAS:de-ba* exists only for Bavarian because of resource limitations and the Lithuanian translation remains an aspect for future work.

3.2. Natural Datasets

The natural datasets for Bavarian (*nat:de-ba*) and Lithuanian (*nat:lt*) are both collected from native speakers of the respective languages to investigate the effect of naturalness as opposed to the translationese of translated datasets.

For Bavarian, we acquired 315 sentences, for Lithuanian 327 sentences, therefore nearly an equal amount. Examples for the natural Lithuanian data can be seen in Table 3. The acquisition of the sentences was designed in such a way that for each xSID intent, participants were asked to provide written utterances, imagining they give commands to a digital assistant. Like this, we have sentences for all 16 xSID intents in each natural dataset. The exact acquisition methods are highlighted individually for each language in the following. For the slots, we manually annotated all natural data. However, the sentences did not contain every slot that appeared in the xSID set (34). The final Bavarian sentences contain 26 slot labels, whereas in the Lithuanian data 30 slot types occurred. Those statistics can be seen in Table 1.

Acquisition for Bavarian We collected utterances from native speakers by providing them with a questionnaire, which indicated that they contribute to a research project for Bavarian. An example question that was used in the form can be seen in Figure 3.

The 21 participants were presented with 16 different scenarios in which they were asked to imagine giving commands or asking questions to a digital assistant. The collected sentences are left intact as they were received and were not corrected or altered in any way. All participants originate from the Upper Bavaria region between Munich and the Austrian border, belonging to the Central Bavarian language region.

The wide region from which the anonymous participants come is responsible for a peculiarity of the Bavarian dialect that appeared in the natural Bavarian data: As there is no standardised orthography for the dialect, different spellings of the same words occurred. While this also applies to the translated sentences, there are less variations because it was manually done by one person, one of the authors. The spelling variations are much more common and relevant in the natural data as the sentences were produced by different people with different preferences. To exemplify this, consider the word German word “stellen” (to set). The verb is referred to with 4 different writing variations: “stei”, “stö”, “steu”, “stoi”.

Acquisition for Lithuanian The sentences for the Lithuanian natural data were collected from 19 participants with the help of a questionnaire (cf. Figure 3). The utterances are again commands for various tasks that match the xSID intents. However, digital assistants are not very common in Lithuania which led to the production of unsuitable sentences that were too abstract or did not fit the intent. These utterances were removed from the dataset.

In addition, two types of mistakes were corrected by one of the authors. One the one hand, mistyped words that were clearly identifiable were corrected and on the other hand, words that did not use the Lithuanian alphabet were transcribed. Some Lithuanian letters do not exist in the English alphabet and are therefore often exchanged for similar Latin letters.

4. Experimental Setup

Following van der Goot et al. (2021a), we evaluate a zero-shot scenario and train a model on the xSID English training data. We test the performance of the model on our Bavarian and Lithuanian datasets. We chose this setup for our experiments due to the lack of Bavarian and Lithuanian resources for training such models on our target languages. This setup allows us to investigate

Language	Sentence	Word Order	Comment
English	show set alarms	V-A-O	English translation
Lithuanian	parodyti žadintuvus nustatytus	V-O-A	original sentence
	rodyti nustatytus žadintuvus	V-A-O	Verb in infinitive without prefix
	nustatytus žadintuvus parodyti	A-O-V	Verb in infinitive
	žadintuvus nustatytus rodyti	O-A-V	Verb in infinitive form without prefix

Table 3: Lithuanian sentences from nat:lt with different word orders. V = Verb, A = Adjective, O = Object.

Bavarian Google Forms form:

Stell dir vor...

Dass du einen Wecker einstellen willst.

Beispiel: Stei ma an Wecka auf 8 Uhr.

Lithuanian Questionnaire:

Jums reikia nurodyti / liepti skaitmeniniam asistentui balsu tvarkyti veiksmus susijusius su žadintuvu. Kaip jus nurodytumete atlikti atlikti šias užduotis:

a) Atšaukti žadintuvą

b) Nustatyti žadintuvą

c) Pakeisti žadintuvo laiką

Figure 3: Example questions extracted from the Bavarian and Lithuanian questionnaires.

the transfer capabilities from English to the low-resource languages.

We use the MaChAmp toolkit (van der Goot et al., 2021b) v0.4 for all our experiments with default hyperparameters. We implement slot and intent detection with a shared encoder. Each task then has its own decoder. For intent detection we classify the special start token, for intent detection we use IOB labels on the word level. The pre-trained language model we used for the baseline model is mBERT (Devlin et al., 2019) following van der Goot et al. (2021a). mBERT includes both Bavarian and Lithuanian in the pre-training data. For evaluation, the metrics used are accuracy for intent classification and span F1-scores to evaluate slot detection. For both metrics, we use the implementation of van der Goot et al. (2021a).

We train over three different random seeds and present the average results and standard deviations on the development data in Table 4. Likewise, the average test data results can be seen in Table 5. The full result tables with the scores per seed for both the development and test datasets can be seen in the Appendix in Tables 10 and 11, respectively.

The English and standard German xSID (van der Goot et al., 2021a) datasets are the base for the comparison of the performance on our Bavarian sets. We compare our results to the results of on the xSID dataset, but we want to note that we use a newer version of MaChAmp

and the results are slightly higher than the original experiment scores (van der Goot et al., 2021a).

5. Results

The results on the development data can be seen in Table 4 and those on the test data in Table 5. Below, we provide an evaluation of the performance of the intent classification and slot detection tasks.

5.1. Development Data Results

Intent Accuracy The task of intent classification is easy for standard languages. This can be seen when predicting on the English xSID data. There, the models achieve 100% over all three random seeds. Therefore, we consider the task on English data as solved. This is not the case for standard German, where intent accuracy is 75.7%.

When comparing standard German to Bavarian, however, we observe that intent accuracy drops from 75.7% to 66.6% and 51.3% on the translated Bavarian datasets *de-ba* and *xMAS:de-ba*, respectively. Overall despite of the drop in performance, the model performs intent classification quite well on this low-resource dialect. It is interesting, however, that the score of *MAS:de-ba* lies below *de-ba*. The combined *xMAS:de-ba* score lies between the *de-ba* and *MAS:de-ba* results and shows that the MASSIVE portion of the data has a negative impact on the cross-dataset results. Therefore we can derive the hypothesis that the method of collecting the data is very important as the sentences may differ from each other (Figure 2), and we should focus more on cross-dataset experiments, so that we can ensure our models are robust.

We hypothesise that the better performance on *de-ba* stems from the usage of xSID training data for the model training, meaning that the xSID evaluation sentences are more similar to the training data. Therefore, other sentences are more distant to the model and harder to predict on.

The Lithuanian intent accuracy moves within the same lines as the Bavarian data with an accuracy of 59.4% and lies between the scores of *xMAS:de-ba* and *de-ba*. We expected a similar performance, as both languages are included in the mBERT (Devlin et al., 2019) training data.

	Intent Class.	Slot Det.
Language	Accuracy	Slot F1 Score
<i>xSID:en</i>	100.0 \pm 0.0	97.2 \pm 0.1
<i>xSID:de</i>	75.7 \pm 4.0	73.0 \pm 0.7
<i>de-ba</i>	66.6 \pm 3.7	46.5 \pm 3.7
<i>MAS:de-ba</i>	44.2 \pm 4.7	18.6 \pm 1.2
<i>xMAS:de-ba</i>	51.3 \pm 4.0	28.8 \pm 2.3
<i>lt</i>	59.4 \pm 4.3	47.1 \pm 2.8

Table 4: Development data results and standard deviations on English, German, Bavarian and Lithuanian. Average scores across three random seeds.

A closer analysis of the results is provided in Section 6.

Slot F1 Scores Our results confirm that slot detection is a more difficult task than intent classification. This is reflected in the low *xSID:en* and *xSID:de* scores, and the even lower scores of our low-resource variants (Table 4, lower part).

The low F1 scores of 18.6% on *MAS:de-ba* and the resulting 28.8% on *xMAS:de-ba* are especially noteworthy. The prediction in the cross-dataset setup does not work as well as on *de-ba* where the evaluation reached 46.5% for the reasons already explained on the intent accuracy results.

A similar score can be observed for the Lithuanian data. The translated *lt* data reached a F1 score of 47.1%. Interestingly, this score is a bit higher than the score of the translated Bavarian *MAS:de-ba* dataset which underlines the low cross-dataset performance.

To investigate whether the detection of the borders of the slots or the labelling itself is the bottleneck, we provide a further analysis in Section 6.2.

5.2. Test Data Results

Intent Accuracy The intent accuracy scores on the test datasets barely differ from the results on the development data which confirms our findings on the development data. The *xSID*-only data *de-ba* still performs better than *MAS:de-ba* and *xMAS:de-ba*. As before, *MAS:de-ba* lowers the cross-dataset experiment results of *xMAS:de-ba*. The scores for Lithuanian *lt* again lie in between those values.

Most importantly, we observe that the natural datasets’ performances are lower than the translated datasets. This shows that the natural data is harder for the model to predict on, and translations provide a slightly more optimistic evaluation score. We hypothesise that the higher results on translated queries are due to the influence of “translationese” on the translated data (Section 2.3).

	Intent Class.	Slot Det.
Language	Accuracy	Slot F1 Score
<i>xSID:en</i>	99.2 \pm 0.2	95.4 \pm 0.5
<i>xSID:de</i>	74.5 \pm 3.9	69.6 \pm 0.8
<i>de-ba</i>	67.3 \pm 4.3	44.8 \pm 2.2
<i>MAS:de-ba</i>	45.4 \pm 4.8	18.6 \pm 0.9
<i>xMAS:de-ba</i>	51.1 \pm 4.9	27.1 \pm 1.4
<i>lt</i>	58.9 \pm 3.6	43.3 \pm 2.3
<i>nat:de-ba</i>	38.0 \pm 5.2	25.0 \pm 1.6
<i>nat:lt</i>	43.2 \pm 5.5	31.6 \pm 2.3

Table 5: Test data results and standard deviations. Average scores across three random seeds.

Translationese makes the translated datasets better known to the model than the natural sentences due to English traces in the sentences that the native utterances miss completely.

Slot F1 Scores Just like with the development data, the slot f1 scores are lower than the slot detection results on *xSID:en* and *xSID:de*.

It is especially noteworthy that even though the performance on *nat:lt* is low, it is still higher than the score of *MAS:de-ba* and the combined *xMAS:de-ba* dataset. Both of them should perform better according due to the translationese influence. This underlines the low performance of the cross-dataset experiments even further and hints that the MASSIVE utterances are challenging for the model that was trained only on *xSID* data.

The trends for the Lithuanian data corresponds to the Bavarian data. The natural dataset *nat:de-ba* performs worse than the translated data which leads us back to the influence of translationese that improves the performance on the translated data. The highest Bavarian slot F1 score was again reached on the *xSID*-only *de-ba*.

6. Analysis

To identify in which cases the model fails, we conduct several in-depth analyses.

6.1. Confusion Matrices for Intent Accuracy

***xSID* Datasets** For the analysis of the error types made on each of our proposed datasets for intents, we use confusion matrices of the best run (random seed 0212), shown in Figure 4. The results on the translated *xSID* data (*de-ba* and *lt*) show that there is a large disparity on the difficulty of the true labels. The “snooze_alarm” and “cancel_alarm” intents are difficult and “PlayMusic” and “weather/find” are heavily overpredicted for both languages.

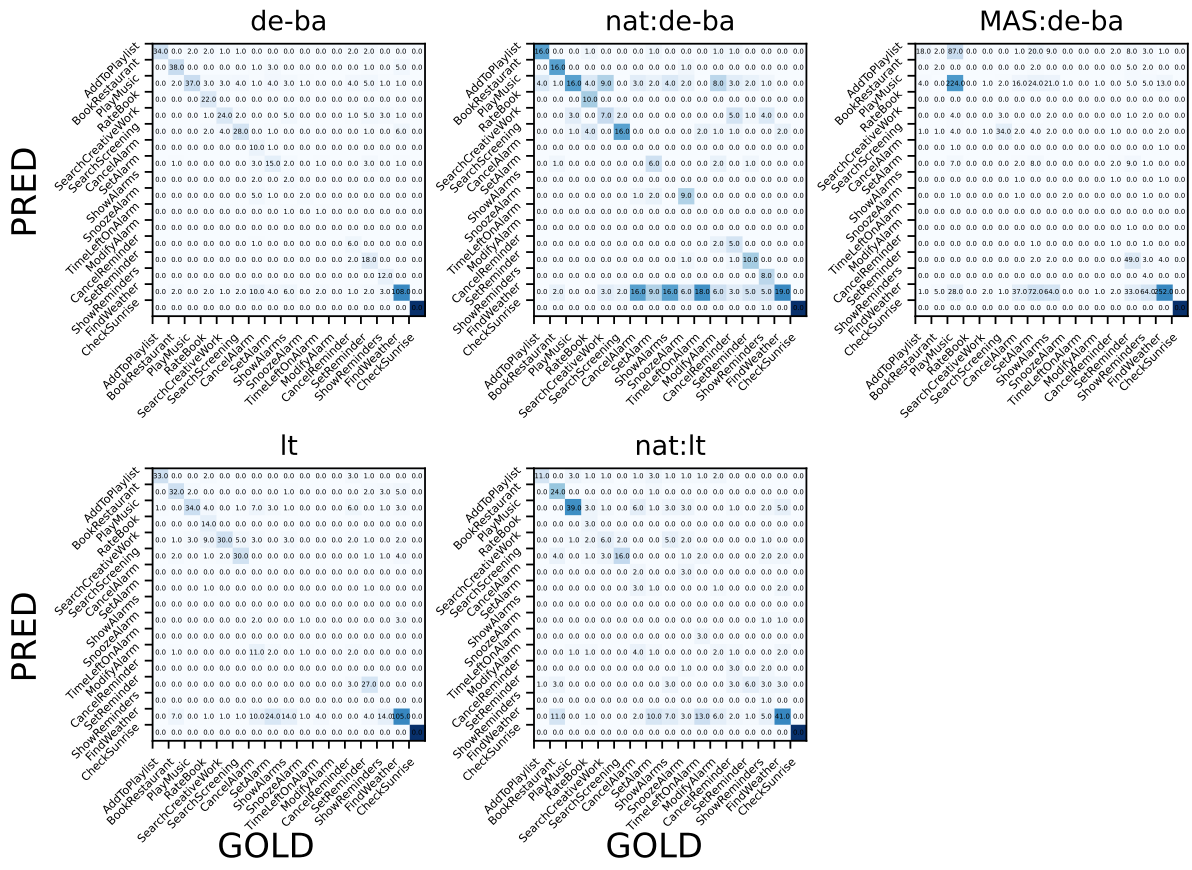


Figure 4: Confusion Matrices for the test sets.

While the trends are mostly the same, there are also differences between the two language varieties: In *It* the “weather/find” label is wrongly predicted for a wider variety of true labels, and “SearchCreativeWork” is harder in the Bavarian dataset. Also, in the Lithuanian data, many of the difficult intents are never found (0.00 on the diagonal). The MASSIVE data (*MAS:de-ba*) shows the same tendencies as the xSID data. However, as there is a higher number of “PlayMusic” utterances in the MASSIVE data, other intents like “AddToPlaylist” tend to be predicted wrongly as “PlayMusic” which creates an overprediction of the intent, even though the majority of the intent’s predictions are correct.

Natural Datasets Interestingly, the results on the translated and natural datasets show very similar trends: for example, both *de-ba* and *nat:de-ba* the “PlayMusic”, “SearchCreativeWork”, and “weather/find” intents are overpredicted; although there are more of the same errors in *nat:de-ba*. For the results on Lithuanian, we also see a lot of correspondence in the type of errors, but for the natural data (*nat:It*), there are also more labels that are consistently not found, like “cancel_alarm”, “show_alarms”, and “snooze_alarm”.

Deviations Analysis The high standard deviations between the three random seeds (cf. Table 5) lead us to examine the confusion matrices of the best and worst seed for each dataset to investigate possible causes for the performance differences. The models trained on English only has unstable performance on the other languages’ datasets on specific labels. Even though the languages are not related, the specific mispredicted intents overlap between the languages. *de-ba* struggles with “BookRestaurant”, “SearchCreativeWork” and “weather/find”. *MAS:de-ba* has difficulties in classifying “AddToPlaylist”. *It* tends to overpredict “PlayMusic”. These trends also apply to the natural datasets *nat:de-ba* and *nat:It*. This shows that there are just a few labels that lead to the large stdev values we observe. The performance on the other intents is more stable.

6.2. Unlabeled and Loose F1 Scores for Slot Detection

We next aim to gain a deeper understanding of slot prediction performance. To do so, we calculate the unlabeled and loose F1 scores (van der Goot et al., 2021a) for our datasets. For that, we use the average results over three random seeds

Language	Strict F1	Unlab.F1	Loose F1
<i>MAS:de-ba</i>	18.6 ±0.8	41.6 ±2.1	35.3 ±1.9
<i>de-ba</i>	45.0 ±2.3	66.2 ±2.0	58.1 ±2.9
<i>lt</i>	43.4 ±2.4	68.7 ±2.9	54.3 ±2.7
<i>nat:de-ba</i>	26.6 ±0.7	55.3 ±3.6	41.9 ±1.6
<i>nat:lt</i>	34.2 ±3.7	54.6 ±3.0	45.4 ±3.5

Table 6: Slot detection results on our datasets. Averages over 3 random seeds + standard deviations. Unlab. (Unlabeled) F1.

and the standard deviations. Unlabeled F1 is a score that indicates the model’s ability to identify slots, regardless of whether it correctly assigns a label to the slot. The loose F1 score confirms whether correctly labeled slots are identified with non-exact boundaries. A high unlabeled F1 score indicates that the model assigns incorrect labels to slots whereas a high loose F1 score suggests that the baseline model correctly predicts the label but is inaccurate with slot boundaries.

Table 6 reports the results for all three metrics. In general, the unlabeled and loose F1 are both substantially higher compared to the strict span-F1 score. This indicates that there are many cases where the model produces a partially correct prediction. This is especially true for the natural datasets; the differences on the strict F1 scores are striking. The differences on loose F1 are much smaller, and on unlabeled F1 even smaller. Therefore, we hypothesise that transferring to sentences produced by native speakers mostly leads to partially correct predictions instead of completely wrong ones, which strict F1 evaluation misses. Across datasets, the unlabeled F1 is higher compared to the loose F1, showing that the larger problem for the model is to find the correct label instead of finding the exact span.

If we compare the results on the two language variants to each other, we can see that the difference between unlabeled and loose F1 are largest for the Lithuanian datasets, meaning that the problem of finding the exact span is easier for the model. For German the largest difference between unlabeled and loose is for the native data.

7. Conclusions

We have contributed *NaLiBaSID*, a new benchmark that encompasses various datasets for SID for two low-resource language varieties: Bavarian and Lithuanian. We provide data directly translated and annotated from a cross-lingual benchmark (xSID), data for cross-dataset evaluation (MASSIVE + xSID), and natural data generated from native speakers. Our results show that the translated datasets lead to overoptimistic es-

timates of performance. However, it is interesting that the types of errors made on the translated and the natural data are highly similar. The cross-dataset results show that results are already a lot lower when data is collected with a different procedure. With our work, we aim to contribute to supporting linguistic diversity of NLU applications.

Acknowledgements

We thank the anonymous reviewers as well as the members of the MaiNLP research lab for their constructive feedback. We also thank our participants that provided the native queries. This work is supported by ERC Consolidator Grant DIALECT no. 101043235.

8. Bibliographical 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 *Proceedings of the Tenth Workshop on NLP for Similar Languages, Varieties and Dialects*, Dubrovnik, Croatia. Association for Computational Linguistics.
- Emanuele Bastianelli, Andrea Vanzo, Pawel Swietojanski, and Verena Rieser. 2020. *SLURP: A spoken language understanding resource package*. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7252–7262, Online. Association for Computational Linguistics.
- 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.
- Yuri Bizzoni, Tom S Juzek, Cristina España-Bonet, Koel Dutta Chowdhury, Josef van Genabith, and Elke Teich. 2020. *How human is machine translationese? comparing human and machine translations of text and speech*. In *International Workshop on Spoken Language Translation*.
- Verena Blaschke, Christoph Purschke, Hinrich Schütze, and Barbara Plank. 2024. *What do dialect speakers want? a survey of attitudes towards language technology for german dialects*.
- Alice Coucke, Alaa Saade, Adrien Ball, Théodore Bluche, Alexandre Caulier, David Leroy, Clément Doumouro, Thibault Gisselbrecht,

- Francesco Caltagirone, Thibaut Lavril, Maël Primet, and Joseph Dureau. 2018. [Snips voice platform: an embedded spoken language understanding system for private-by-design voice interfaces](#).
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- David M. Eberhard, Gary F. Simons, and Charles D. Fennig. 2023. [Ethnologue: Languages of the world](#).
- 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](#).
- Sara Laviosa. 1997. [How comparable can 'comparable corpora' be?](#) *Target-international Journal of Translation Studies*, 9:289–319.
- Libo Qin, Tianbao Xie, Wanxiang Che, and Ting Liu. 2021. [A survey on spoken language understanding: Recent advances and new frontiers](#).
- Evgeniia Razumovskaia, Goran Glavaš, Olga Majewska, Edoardo M. Ponti, Anna Korhonen, and Ivan Vulić. 2022. [Crossing the conversational chasm: A primer on natural language processing for multilingual task-oriented dialogue systems](#). *Journal of Artificial Intelligence Research* 74, 74.
- Anthony R. Rowley. 2011. [Bavarian: Successful dialect or failed language?](#) In Joshua Fishman and Ofelia Garcia, editors, *Handbook of Language and Ethnic Identity*, volume 2 (The Success-Failure Continuum in Language and Ethnic Identity Efforts), pages 299–309. Oxford University Press.
- Sebastian Schuster, Sonal Gupta, Rushin Shah, and Mike Lewis. 2019. [Cross-lingual transfer learning for multilingual task oriented dialog](#).
- Gideon Toury. 2012. *Descriptive Translation Studies – and beyond*. John Benjamins Publishing Company.
- Shyam Upadhyay, Manaal Faruqui, Gokhan Tür, Hakkani-Tür Dilek, and Larry Heck. 2018. [\(almost\) zero-shot cross-lingual spoken language understanding](#). In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6034–6038.
- Rob van der Goot, Ibrahim Sharaf, Aizhan Imankulova, Ahmet Üstün, Marija Stepanović, Alan Ramponi, Siti Oryza Khairunnisa, Mamoru Komachi, and Barbara Plank. 2021a. [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.
- Rob van der Goot, Ahmet Üstün, Alan Ramponi, Ibrahim Sharaf, and Barbara Plank. 2021b. [Massive choice, ample tasks \(MaChAmp\): A toolkit for multi-task learning in NLP](#). In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 176–197, Online. Association for Computational Linguistics.
- Philip C. Vergeiner and Lars Bülow. 2023. [Geolinguistic structures of dialect phonology in the german-speaking alpine region: A dialectometric approach using crowdsourcing data](#). *Open Linguistics*, 9(1):20220252.
- Weijia Xu, Batool Haider, and Saab Mansour. 2020. [End-to-end slot alignment and recognition for cross-lingual NLU](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5052–5063, Online. Association for Computational Linguistics.
- Mike Zhang and Antonio Toral. 2019. [The effect of translationese in machine translation test sets](#). In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 73–81, Florence, Italy. Association for Computational Linguistics.

Appendix

A. Translation Challenges

The translation was carried out following the translation guidelines of [van der Goot et al. \(2021a, Appendix F\)](#). However, the Bavarian and Lithuanian languages posed some challenges which are further examined here.

A.1. Bavarian

1. Numbers mainly appear in the xSID data set as numerals and sometimes as written out forms. We translate them as a mixture of digits and words. This is due to the fact that the Bavarian dialect has dialect pronunciations of the numbers that we want to reflect in the data.
2. Time mentions in English are usually accompanied by “am” and “pm”. As there is no corresponding fixed expression in German, we translate “am” as “in da fria” (“in the morning”) and “pm” as “nammiddog”/“nammidog” (“afternoon”) or “aufnacht”/“auf Nacht” (“in the evening/night”) depending on the time of day to reflect these time expressions. We want to note, however, that in spoken language, these clarifications of time of day are usually omitted.
3. We did not translate names or named entities into a German variation.

A.2. Lithuanian

1. Geographical locations such as cities, countries, etc. have to be translated to the Lithuanian language, since the names of locations have cases and need to match the rest of the sentence. For this, we use the dictionary of the Lithuanian language commission³ for confirmed translation of locations.
2. The numbers have cases in Lithuanian, too. In written language, numbers are always written as digits and Lithuanian speakers intuitively adjust the case. We transcribe part of the numbers to words to keep variety between digits and numeral words and to portray the different cases in our data.
3. While translating, we tried to keep grammatical mistakes in similar form. But it was difficult as nouns and verbs often have different structures so it was not always possible to find ways to replicate the errors from the English sentence in Lithuanian language.

4. No name transcription was done on translations.
5. Sometimes words like book, movie or game are added, even if it is not in the original sentence. This allows for application of cases when the name of the medium is not translated.
6. Sentence modifications were created to see how many words need to be replaced for the baseline model to correctly predict the intent label. Table 7 shows the small sample of a dataset created from 1 sentence. The “word order” column shows the word order in each sentence, here: V- verb, O - object, A - adjective. In block 2, adjective is replaced with the possessive pronoun ‘my’. In block 3, some Lithuanian words are replaced with English. All of these sentences in block 3 would never be used in Lithuanian spoken language.

B. MASSIVE Re-Annotation

The intents of the MASSIVE ([FitzGerald et al., 2022](#)) dataset differ from the fixed set of 16 intents that was used in the annotation of the xSID ([van der Goot et al., 2021a](#)) dataset.

For that reason, the utterances had to be manually inspected and aligned to the xSID set. The mapping is portrayed alongside the frequencies in Table 8. For all xSID intents except for “weather/find”, parts of several MASSIVE intents could be subordinated to each intent.

The MASSIVE slot labels were also replaced with the xSID slot labels as can be reviewed in Table 9. The slot renaming process was performed together with relabelling, as there are multiple slot types in xSID that do not appear in MASSIVE but hold valuable information, e.g. “reference”.

C. Full Experiment Results

Tables 10 and 11 show the full experiment results on all datasets across the three different random seeds we trained. Table 10 shows all intent accuracy and slot-f1 scores for the development data, whereas table 11 presents the full results on the test dataset splits.

³<https://pasaulio-varnai.vlkk.lt/>

Sentence	Predicted intent	Predicted slot	Word order	Comment
parodyti [žadintuvus nustatytus]	SearchCreativeWork	object_name	V-O-A	original sentence
parodyti (žadintuvus nustatytus)	SearchCreativeWork	object_name	V-O-A	original sentence
show set alarms	alarm/show_alarms		V-A-O	English translation. Set- adjective
parodyti (žadintuvus nustatytus)	weather/find	object_name	V-O-A	Verb in correct infinitive
(parodyk žadintuvus nustatytus)	weather/find	object_name	V-O-A	Verb in imperative
parodyti (nustatytus žadintuvus)	weather/find	object_name	V-A-O	Verb in shortened infinitive
parodyti nustatytus žadintuvus	weather/find		V-A-O	Verb in full infinitive form
(parodyk nustatytus žadintuvus)	weather/find	object_name	V-A-O	Verb in imperative
rodyt žadintuvus nustatytus	weather/find		V-O-A	Verb in shortened infinitive without prefix
rodyti žadintuvus nustatytus	weather/find		V-O-A	Verb in correct infinitive form without prefix
rodyti (žadintuvus nustatytus)	SearchCreativeWork	object_name	V-O-A	Verb in imperative form without prefix
rodyti nustatytus žadintuvus	weather/find		V-A-O	Verb in shortened infinitive without prefix
rodyk nustatytus žadintuvus	weather/find		V-A-O	Verb in correct infinitive without prefix
nustatytus (žadintuvus parodyti)	weather/find	object_name	A-O-V	Verb in imperative without prefix
nustatytus žadintuvus parodyti	weather/find		A-O-V	Verb in shortened infinitive
(nustatytus žadintuvus) (parodyk)	weather/find	object_name; location	A-O-V	Verb in imperative
(nustatytus žadintuvus) rodyt	weather/find	object_name	A-O-V	Verb in shortened infinitive without prefix
nustatytus žadintuvus rodyti	weather/find		A-O-V	Verb in imperative without prefix
(nustatytus žadintuvus) (rodyk)	weather/find	object_name; location	A-O-V	Verb in imperative without prefix
(žadintuvus nustatytus) parodyti	weather/find	object_name	O-A-V	Verb in shortened infinitive
(žadintuvus nustatytus) parodyti	weather/find	object_name	O-A-V	Verb in correct infinitive form
(žadintuvus nustatytus) parodyk	SearchCreativeWork	object_name	O-A-V	Verb in imperative
(žadintuvus nustatytus) rodyt	weather/find	object_name	O-A-V	Verb in shortened infinitive without prefix
(žadintuvus nustatytus) rodyti	weather/find	object_name	O-A-V	Verb in correct infinitive form without prefix
(žadintuvus nustatytus) rodyk)	SearchCreativeWork	object_name	O-A-V	Verb in imperative without prefix
nustatytas (eng.set) → mano (eng. my)				
parodyti (žadintuvus mano)	SearchCreativeWork	object_name	V-O-A	Verb in shortened infinitive
parodyti (žadintuvus mano)	SearchCreativeWork	object_name	V-O-A	Verb in correct infinitive
(parodyk žadintuvus mano)	SearchCreativeWork	artist	V-O-A	Verb in imperative
parodyti (mano žadintuvus)	SearchCreativeWork	object_name	V-A-O	Verb in shortened infinitive
parodyti (mano žadintuvus)	SearchCreativeWork	artist	V-A-O	Verb in correct infinitive
(parodyk mano žadintuvus)	PlayMusic	artist	V-A-O	Verb in imperative
it word → eng word				
show (nustatytus žadintuvus)	SearchCreativeWork	object_name	V-A-O	V:it → eng
show nustatytus alarms	alarm/show_alarms		V-A-O	V and O:it → eng
rodyti nustatytus alarms	weather/find		V-A-O	O:it → eng, V - infinity form, no prefix
rodyk nustatytus alarms	alarm/cancel_alarm		V-A-O	O:it → eng, V - imperative form, no prefix
parodyk nustatytus alarms	alarm/modify_alarm		V-A-O	O:it → eng, V - imperative form

Table 7: Different forms of nat:It dataset's id:47 sentence. Note 1: shortened infinitive is used in spoken language and comes from one of dialects. Note 2: the 3rd block replaces some Lithuanian words to English, and those sentences are incorrect. It tests how many words need to replace for baseline model to correctly identify intent.

xSID Intent	MASSIVE Intent	# sents	# total
PlayMusic	play_music	506	588
	play_radio	79	
	music_likeness	2	
	calendar_set	1	
weather/find	weather_query	439	439
alarm/set_alarm	alarm_set	223	225
	calendar_set	2	
alarm/show_alarms	alarm_query	172	176
	alarm_set	3	
	calendar_query	1	
reminder/show_reminders	calendar_query	166	169
	alarm_query	2	
	calender_set	1	
reminder/set_reminder	calendar_set	161	171
	alarm_set	7	
	calendar_query	2	
	play_music	1	
alarm/cancel_alarm	alarm_remove	104	104
SearchScreeningEvent	recommendation_movies	48	59
	recommendation_events	6	
	calendar_query	4	
	takeaway_order	1	
AddToPlaylist	music_likeness	35	36
	play_music	1	
BookRestaurant	recommendation_locations	25	27
	takeaway_query	2	
reminder/cancel_reminder	calendar_remove	8	14
	alarm_remove	6	
SearchCreativeWork	music_query	2	6
	recommendation_movies	4	
alarm/modify_alarm	alarm_remove	2	4
	alarm_set	2	
alarm/snooze_alarm	alarm_remove	2	3
	alarm_set	1	
RateBook	—	0	0
alarm/time_left_on_alarm	—	0	0
			2021

Table 8: Mapping of the MASSIVE intents to the xSID intents for the MASSIVE utterances that are part of xMAS:de-ba.

xSID slot	MASSIVE slot	# slots
datetime	date	1001
reminder/todo	time meal_type timeofday event_name	350
reference	—	350
location	place_name	196
weather/attribute	weather_descriptor	191
artist	artist_name	182
genre	music_genre	141
track	song_name	130
music_item	media_type music_album	99
movie_type	movie_type	38
playlist	playlist_name	36
service	media_type app_name	34
condition_description	weather_descriptor	28
condition_temperature	weather_descriptor	27
object_location_type	—	22
sort	—	21
movie_name	movie_name	9
restaurant_type	business_type	9
object_type	—	8
served_dish	food_type	8
object_select	—	6
rating_value	—	3
restaurant_name	business_name	2
rating_unit	—	1
party_size_number	—	1

Table 9: Mapping of the MASSIVE slot labels to the xSID slot labels for the MASSIVE utterances that are part of *xMAS:de-ba*

Dev Data		mBERT-sid		
		s8446	s0212	s2301
Intent Classification	<i>xSID:en</i>	100	100 <i>100</i>	100
	<i>xSID:de</i>	72.0	80.0 <i>75.7</i>	75.0
	<i>de-ba</i>	68.3	69.0 <i>66.6</i>	62.3
	<i>MAS:de-ba</i>	46.9	51.5 <i>47.0</i>	42.5
	<i>xMAS:de-ba</i>	53.1	54.1 <i>51.3</i>	46.8
	<i>It</i>	60.7	54.7 <i>59.4</i>	63.0
	Slot Detection	<i>xSID:en</i>	97.3	97.1 <i>97.2</i>
<i>xSID:de</i>		72.5	72.8 <i>73.0</i>	73.8
<i>de-ba</i>		50,1	46,6 <i>46,5</i>	42,7
<i>MAS:de-ba</i>		19.8	18.3 <i>18.5</i>	17.4
<i>xMAS:de-ba</i>		31.4	27.3 <i>28.8</i>	27.6
<i>It</i>		48,4	43,9 <i>47,1</i>	48,9

Table 10: Development data results across all three different random seeds and the *average scores*.

Test Data		mBERT-sid		
		s8446	s0212	s2301
Intent Classification	<i>xSID:en</i>	99.4	99.2 99.2	99.0
	<i>xSID:de</i>	73.2	78.8 74.5	71.4
	<i>de-ba</i>	67.8	71.4 67.3	62.8
	<i>MAS:de-ba</i>	51.5	49.0 48.0	43.4
	<i>xMAS:de-ba</i>	53.0	54.7 51.1	45.5
	<i>lt</i>	60.8	54.8 58.9	61.2
	<i>nat:de-ba</i>	36.6	43.8 38.0	33.7
	<i>nat:lt</i>	48.3	37.3 43.2	44.0
Slot Detection	<i>xSID:en</i>	95.6	94.8 95.4	95.8
	<i>xSID:de</i>	68.8	69.5 69.6	70.4
	<i>de-ba</i>	46.9	44.8 44.8	42.6
	<i>MAS:de-ba</i>	19.5	17.8 18.6	18.5
	<i>xMAS:de-ba</i>	28.4	27.4 27.1	25.6
	<i>lt</i>	44.3	40.7 43.3	45.1
	<i>nat:de-ba</i>	23.3	26.4 25.0	25.4
	<i>nat:lt</i>	32.0	29.0 31.6	33.6

Table 11: Test data results across all three different random seeds and the *average scores*. The natural dataset are not split into development and test sets and are instead used in their entirety for the test data experiments.

D. NaLiBaSID Data Statement

Data statement for *NaLiBaSID* following [Bender and Friedman \(2018\)](#):

A. CURATION RATIONALE Dataset for use in slot and intent detection tasks for digital assistants.

- *de-ba, lt*: We translated from the xSID dataset ([van der Goot et al., 2021a](#)) and expand the language availability of this dataset.
- *MAS:de-ba*: We translated a sample from the MASSIVE dataset ([FitzGerald et al., 2022](#)).
- *nat:de-ba, nat:lt*: We collected a small dataset of natural utterances written down by native speakers of the respective language.

B. LANGUAGE VARIETY

- *de-ba, MAS:de-ba*: The translation was carried out by a native Bavarian speaker, one of the authors.
- *lt*: The translation was carried out by a native Lithuanian speaker, one of the authors.
- *nat:de-ba*: The sentences were produced by native Bavarian speakers stemming from the southern Bavarian region in Germany and the border area of Austria, therefore the dialect can be classified as Central Bavarian (also reaching into the transition area between Central and Southern Bavarian) ([Vergeiner and Bülow, 2023](#)).
- *nat:lt*: The data was created by native Lithuanian speakers stemming mostly from the area of the capital Vilnius.

Bavarian corresponds to the iso 639-3 code “bar”, Lithuanian to “lit”.

C. SPEAKER DEMOGRAPHIC

- *de-ba, MAS:de-ba, lt*: The demographics of the crowd workers that generated the original data are unknown.
- *nat:de-ba*: The survey asking the participants to produce the utterances was conducted anonymously, therefore there are no details about the speaker demographic. The chosen distribution channels (Discord, word-of-mouth recruiting) allow us to assume that students between the ages 19 - 26 (highest student counts by age in Germany in 2022/23 according to [statista](#)⁴) or native speakers between the ages 30 and 55 participated.

⁴Statista: Student Counts by Age at German Universities

- *nat:lt*: The data was generated by 19 Lithuanian native speaking respondents living in Lithuania, mostly in capital Vilnius. 3 participants were aged 14 or younger, 3 participants 15 - 24, 12 participants 25 - 54 and one participant was 55 or older.

D. ANNOTATOR DEMOGRAPHIC

- *de-ba, MAS:de-ba, nat:de-ba*: The annotator is a native speaker of the Bavarian dialect within the age range of 18-24 years and lives in the southern region of Bavaria in Germany. The annotator has a Bachelor’s degree in Computational Linguistics. The translation was carried out by the annotator
- *lt, nat:lt*: The annotator is native Lithuanian speaking person aged within the range of 35-40 years. The annotator has acquired secondary education in Lithuanian language, a Bachelor’s degree in Statistics in Lithuanian language and another Bachelor’s degree in Data Science in English. The translation was carried out by the annotator.

E. SPEECH SITUATION

- *de-ba, lt*: The xSID data that our translations are based on stems from Snips ([Coucke et al., 2018](#)), generated in June 2017, and Facebook ([Schuster et al., 2019](#)), probably generated in 2019. The chosen method was a textual collection of sentences that crowd workers would say to a digital assistant given a specific topic (intent).
- *MAS:de-ba*: MASSIVE are translations based on the SLURP dataset ([Bastianelli et al., 2020](#)). SLURP was probably produced in 2020 by presenting crowd-workers with 200 pre-defined prompts asking them to write how they would ask a digital assistant to perform tasks. The MASSIVE translations were probably performed in 2022 by asking professional translators to translate the SLURP data into their local language, resulting in 50 translation.
- *nat:de-ba*: The participants were presented a survey that asked them to write down how they would ask a digital assistant for certain tasks exactly the way that they would speak it in order to capture the individual dialect.
- *nat:lt*: The data has been gathered in form of digital questionnaire asking respondents to write the sentences as how they would ask digital assistant in spoken form to complete a task (intent). In case of the children, the questions

have been asked in person and written down by an adult.

F. TEXT CHARACTERISTICS

- *de-ba, lt*: The data spans several different topics as defined by the following intents:

AddToPlaylist

BookRestaurant

PlayMusic

RateBook

SearchCreativeWork

SearchScreeningEvent

alarm/cancel_alarm

alarm/modify_alarm

alarm/set_alarm

alarm/show_alarms

alarm/snooze_alarm

alarm/time_left_on_alarm

reminder/cancel_reminder

reminder/set_reminder

reminder/show_reminders

weather/find

- *MAS:de-ba*: The MASSIVE data was re-annotated to fit the xSID intents and slots. However, the data only suited 14 out of 16 intents. Therefore, the genre of *MAS:de-ba* is determined by the same set of intents as *de-ba* except for “RateBook” and “alarm/time_left_on_alarm”
- *nat:de-ba, nat:lt*: The collection of *nat:de-ba* and *nat:lt* was performed in such a way that utterances for the same set of intents as in xSID ([van der Goot et al., 2021a](#)) was produced. Therefore, the genre of the data is determined by the same intents.