Handling Syntactical Particularities of Western Swiss German Dialects: A Study of Embedding Vector Initialization in Machine Translation

Andreas Møller Belsager and **Linda Caroline Schombach** and **Mira Metzger**
IT-University of Copenhagen
{abel,lics,metz}@itu.dk

Abstract

Machine Translation systems face significant challenges when dealing with low-resource languages, primarily due to data scarcity. Particularly, Swiss German, being predominantly a spoken language with vast dialect differences, has been overlooked in text-to-text models. This study addresses this gap by investigating the potential of using word embeddings from a distinct task to initialize a machine translation model for Bernese Swiss German into High German, focusing on syntax particularities. Results highlight the potential of pre-trained embeddings, especially in handling difficult syntax nuances like verb positioning. Embeddings from Graubünden, with a moderate amount of data and on the other side of the isogloss of Bern, emerge as the most effective choice, surpassing those from the same side of the isogloss (Valais).[1]

1 Introduction

Recent advancements in natural language processing (NLP) have facilitated the integration of machine translation tools into diverse applications, enhancing the impact of technologies such as virtual assistants, language translation services, and automated customer support systems (Dogan-Schönberger et al., 2021). However, a large disparity in capabilities persist across languages. Low-resource languages present challenges for robust training of models due to data scarcity.

A particularly challenging low-resource language is Swiss German. Swiss German is a collection of Standard German (High German) dialects spoken by around five million people (approximately 60 percent of the population) in Switzerland. (Plüss et al., 2022; Schweiz, 2023) As it is predominantly spoken, access to text data is limited, a coherent writing system is missing, spelling ambiguities exist, and text is often restricted to informal contexts. (Plüss et al., 2022) Further, the different Swiss dialects differ substantially in regard to phonetics, grammar, and vocabulary which increases linguistic complexity while further reducing the available amount of data per dialect. (Paonessa et al., 2023; Siebenhaar and Wyler, 1997) These distinctions often comply with isoglosses, geographical boundaries that mark linguistic features. (Glaser, 2022; Seiler et al., 2021; Glaser and Ott, 1997; Linder et al., 2020) This paper addresses the described issues for Swiss German machine translation systems resulting from syntactical particularities.

Leveraging pretrained word embeddings from a different translation task to initialize the word embeddings of a machine translation model has the potential to enhance translation performance, particularly in low resource scenarios. By representing words as vectors in a continuous space based on their relationships and similarities, pretrained embeddings leverage the underlying semantic and syntactic relations among words, thereby enriching the model's linguistic understanding. (Qi et al., 2018; Soliman et al., 2017).

However, the optimal choice of embedding initialization for enhancing a Swiss German translation model's handling of syntax-specific nuances remains uncertain. Considering a dialect with specific syntactical nuances (Bern), three options are explored: initializing with embeddings trained on another Swiss-German dialect from a different side of the syntactical isogloss (Graubünden (GR)), on a Swiss-German dialect within the same side of the isogloss (Valais (VS)), or on the target language (High German). While all options share benefits like shared vocabulary and data efficiency, the extent of these advantages varies. Embeddings

[1] Code/Data available on: https://github.com/mira-me/German_to_Swiss_Translation_ANLP_2023/tree/main

from another Swiss-German dialect offer a higher shared vocabulary, syntax similarities, and dialect-specific knowledge transfer, while embeddings from the target language might excel in potential data efficiency. Initializing with the same side of the isogloss emphasizes syntax similarities, while a different side of the isogloss prioritizes data-driven learning.

This leads to the research question: **How can the initialization of embeddings improve the performance in handling syntax-specific nuances during Bernesen Swiss German to High German text-to-text machine translation?**, followed by the sub-questions:

- *SQ1: Do pretrained embeddings from a dialect close to the source (GR or VS) or the target language (DE) enhance performance?*
- *SQ2: Do pretrained embeddings from a dialect within the source language's syntactical isogloss side (VS) enhance performance more than those from a dialect outside the isogloss side (GR_S)?*
- *SQ3: Do pretrained embeddings from a dialect outside the source language's syntactical isogloss side, but with more data, (GR) enhance performance more than those from a dialect within the isogloss side (VS)?*
- *SQ4: Do pretrained embeddings from a dialect close to the source language (GR or VS) enhance performance more than those from the target language?*

Contributions 1) We provide a fine-tuned NLLB-200 model for text-to-text translation from one Swiss German dialect (Bern) into High German. 2) We evaluate and compare five different initialization for the embeddings of the model to improve handling of three syntax phenomena.

2 Related Work

Swiss German Machine Translation: As Swiss German is predominantly spoken, there has been primarily research of Swiss to Standard German speech-to-text translation, neglecting text-to-text tasks. For instance, Plüss et al. (2021) trained and Automatic Speech Recognition system to automatically transform Swiss German speech into a Standard German text corpus. Recent work also showed that dialect diversity and linguistic differences cause significant challenges in building Swiss German speech translation systems (Paonessa et al., 2023).

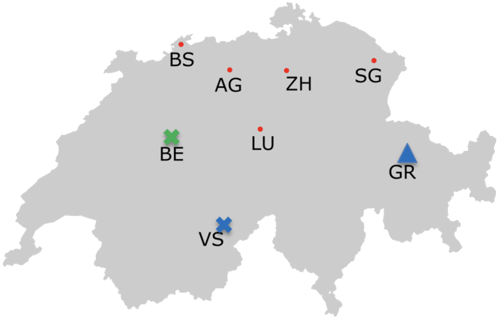

Figure 1: Swiss German dialects present in the used dataset: Aargau (AG), Bern (BE), Basel (BS), Graubünden (GR), Luzern (LU), St. Gallen (SG), Valais (VS) and Zürich (ZH).(*green = source language, blue = dialect for pretraining embeddings, cross = western synthetic isogloss, triangle = eastern synthetic isogloss*) Figure adapted from Dogan-Schönberger et al. (2021)

Swiss German Syntax and Isogloss: Strong evidence has been provided for syntactic differences between Swiss German dialects, particularly by the "Syntaktischer Atlas der Deutschen Schweiz" project (Glaser, 2022). This project systematically and extensively surveys syntax use across dialects since 2000 (Glaser, 2022). It investigates a wide array of syntax constructions, ranging from nominal phrases like possessive constructions, verb groupings with considerations for positions and doublings of verbs, to junctions such as relative clause connections (Glaser, 2022). Further, it presents results in an accessible online map (`https://dialektsyntax.linguistik.uzh.ch`), providing a visual representation of the syntactic landscape and revealing syntactical isoglosses (Seiler et al., 2021). While subtle variations exist between northern and southern dialects, the most significant division lies between West and East Swiss German (Glaser, 2022, 2014; Glaser and Bart, 2015; Seiler et al., 2021; Glaser and Ott, 1997). Although the exact delineation varies based on linguistic variables, we adopt a meticulous approach to ensure a valid separation, categorizing the most western dialects (VS and BE) as west and the most eastern (GR) as east while omitting intermediary dialect as illustrated in Figure 1.

Pretrained Embeddings Previous research has provided evidence that pretrained word embeddings are particularly effective for low-resource machine translation, underlining the prerequisite of sufficient data that allows to capture basic language characteristics (Qi et al., 2018). Further, pretrained embeddings are beneficial if the source and target language are semantically very similar, which is the case for Swiss and High German, as this leads to a more similar semantic neighborhood, improving the effectiveness of pretrained embeddings in capturing nuances and enhancing translation quality (Qi et al., 2018). However, no research presently exists on the effect pretrained embeddings have on handling syntactical differences for Swiss German.

3 Data

Parallel Multidialectal Corpus (Swiss Dial): An open-source dataset provided by ETH Zürich is used for training and testing purposes (Dogan-Schönberger et al., 2021). This dataset includes both audio and transcribed text across 8 key Swiss German dialects as shown in Figure 1, in addition to High German. Beyond being the first annotated parallel corpus across Swiss German dialects, this dataset is uniquely suited to our research due to its sentence-level annotations. The word-level correspondences of other datasets is insufficient for effective training of syntax related translation tasks.

In total, the dataset comprises 11,213 sentences. It is essential to note that while all sentences include a Standard German reference text (target language), only 2,529 provide translations across all 8 Swiss German dialects. Notably, for dialects featuring distinctive syntax constructions, such as Bernese (BE) and VS Swiss German, the dataset contains a limited number of sentences (2700 for BE, 2753 for VS). The GR dialect has the most sentences available with 10475.

Synthetic Test Set: To assess the translation system's proficiency in handling syntax issues, we manually created synthetic Bernese sentences paired with their High German translation. For focus, three specific syntax phenomena were chosen based on the results of the Zurich project "Syntaktischer Atlas der deutschen Schweiz" (Glaser, 2022; Seiler et al., 2021). The selection of the phenomena was based on their presence in the Swiss dialect dataset, divergence between Western (BR, VS) and Eastern dialects (GR), and suitability for synthetic example creation. Those criteria ensure the model is trained on known syntactical nuances, align with the research question, and facilitate practical evaluation by allowing controlled generation of synthetic examples.

Phe.	High German / Bernesen Swiss German
1	Er **lässt** den Senioren selber trinken. / Er **lat** dä Seniorä säuber **la** trinkä.
2	Früher habe ich mehr **bezahlen lassen**. / Früecher hani me **la zahlä**.
3	Die Frau fragte, ob er diesen Pullover **nehmen wolle**. / D Frou het gfragt, ob er dä Pullover **wöu nämä**.

Table 1: Examples of synthetic generated sentences by syntactical phenomenon (1 = verb doubling, 2 = verb positioning of "lassen", 3 = modal verb positioning).

The selected syntax phenomena specific to the western Swiss German region, are 1) the doubling of specific verbs such as 'lassen', 'anfangen', 'gehen' and 'kommen', 2) the intentional swapping of verb positions in sentences structured around 'lassen' plus infinitive, and 3) the reversal of verb groupings, featuring an infinitive and a modal verb ('dürfen', 'sollen', 'können', 'müssen', 'wollen' and 'mögen'). (Glaser, 2022; Seiler et al., 2021)

We build 20 sentences for each phenomenon by leveraging the existing vocabulary in the Swiss Dial data. Examples of the synthetic sentences are shown in Table 1. See Appendix B for examples of the phenomenon present in the SwissDial test set.

4 Methodology

4.1 Translation Task

To answer the research question, the Bernese western Swiss German dialect is selected as the source language. An High German translation task was choosen over an intra-dialect translation task driven by the consideration that significantly more data would be required for satisfactory results of the latter. A pretrained model on Swiss German is not deemed suitable as the training data is unknown, which hinders the evaluation of syntactic nuances.

4.2 NLLB-200 Model

For the translation from Bernese Swiss German into High German, we employ the NLLB-200 (No Language Left Behind) distilled 600M model variant. This transformer model has been trained on an extensive dataset of over 18 billion sentence pairs, spanning 202 languages. Its primary objective is to extend capabilities for lower-resource languages, beyond the usual top 100 languages typically utilized in machine translation tasks. During training, the model leverages a combination of bitext-data sourced from the internet (including Common Crawl and ParaCrawl) and pages where articles are manually translated by humans. The training data also incorporates monolingual data, utilized for various model tasks such as language identification and general language modeling. For evaluation, the model draws from a diverse set of datasets professionally translated by humans, including Flores-200 and NLLB-Seed.

4.3 Word Embedding Initialization Strategies

To enhance model's ability to handle syntactical particularities, we tested initializing the word embeddings of the Bern Swiss German to High German translation model with five different embeddings: 1) Pretrained on VS which aligns with the syntactical isogloss side of BE. (SQ 1 to 4). 2) Pretrained on $\mathrm{GR_L}$ which is outside the syntactical isogloss side of BE but has a larger dataset than VS and more linguistic similarities than DE to BE. (SQ 3 & 4). 3) Pretrained on an undersampled GR ($\mathrm{GR_S}$) dataset (same size as VS) to assess if the impact of embedding differences between VS and $\mathrm{GR_L}$ is attributable to syntax similarities of VS and BE. (SQ 2). 4) Pretrained on DE which has a large dataset and is the source language. (SQ 1 & 4) 5) Randomly initialized as baseline. (SQ 1)

To generate the word embeddings from a Swiss German dialect (1 to 3), it was necessary to first train NLLB-200 models on translating the respective dialect into High German. For this purpose, 90% of the SwissDial dataset was used. The stopping criteria for training varied among dialects due to differences in training data size, as larger datasets may require more time for effective learning. Taking a fixed number of steps as the stopping criterion for all models would hinder those trained on larger datasets from fully leveraging their data resources, which is essential for SQ 3. To reach minimum validation loss, $\mathrm{GR_L}$ required 3900 steps, VS 1600, and $\mathrm{GR_S}$ 2000. For the word embeddings trained on High German, existing ones from the NLLB-200 models were used.

4.4 Training Process

For each word embedding initialization strategy, an NLLB-200 model was trained to translate Bernese Swiss German into High German, using 90% of the SwissDial text corpus. All models shared consistent hyperparameters as listed in Appendix A to ensure a fair comparison, attributing differences to the embedding initialization.

Each model underwent training until reaching 2200 steps. This predefined stopping criterion ensures a fair comparison across various embeddings, avoiding bias introduced by the number of training epochs. The learning progress is visualized in Appendix D. The curves affirm that models initialized with random embeddings and embeddings from a more data-rich language (DE and $\mathrm{GR_L}$) begin with a higher loss. With more steps, curves converge. The final model was selected based on the lowest validation loss (randomly initialized & DE : 1800 steps, VS & $\mathrm{GR_L}$: 1500, $\mathrm{GR_S}$: 1400).

4.5 Evaluation Metrics

Firstly, the overall translation quality of the models resulting from different embedding intializations was assessed using the SwissDial test set and three metrics (BLEU, chrF2, and TER). Secondly, for an assessment of syntax handling, independent of other performance factors such as vocabulary correctness, the synthetic test set was employed. The translated sentences were manually categorized as either syntactical wrong or right. For each syntactical phenomenon and model, the percentage of correct cases was calculated (Syntax Accuracy Ratio). Statistical significance tests for all metrics involve paired bootstrap resampling according to Koehn (2004), comparing the performance of two machine translation systems. This process generates new test sets (10,000 times) by drawing sentences with replacement from the translated collections. For each set, metric scores are computed. Superiority is established if one system outperforms the other 95% of the time.

5 Results

5.1 Overall Translation Quality Evaluation

Considering overall translation performance, $\mathrm{GR_L}$ outperforms random initialization at a significance level of 0.05 in all three metrics (BLEU, chrF2 and

Embedd.	Phen. 1	Phen. 2	Phen. 3	All
Random	75%	35%	65%	58%
VS	80%	60%	75%	72%
GR_L	80%	90%	90%	87%
GR_S	64%	55%	65%	60%
DE	80%	75%	90%	82%

Table 2: Percentage of correctly handled syntactical sentences by phenomenon (1 = verb doubling, 2 = verb positioning of "lassen", 3 = modal verb positioning)

TER) and GR_S in two metrics (BLEU and chrF2). Complete results are presented in Appendix E.

5.2 Syntax Handling Assessment

Syntax Accuracy Ratio The syntax accuracy of each embedding intialization is summarized in Table 2. The highest overall performance is achieved with embeddings pretrained on GR_L, correctly handling syntax in 87% of the sentences. Random initialization performs the lowest with 58%. When examining different syntactical phenomena, the most significant impact of embeddings appears in Phenomenon 2 (baseline is 35%, GR_L is 90%). The significance test results of the relevant embedding comparisons are shown in Table 3. All pretrained embeddings, except GR_S, showed superior performance over random embeddings (SQ 1) in handling syntax nuances at a significance level of 0.05. VS embeddings yield significantly better results than GR_S (SQ 2) but worse than GR_L embeddings (SQ 3). No significant differences were observed between GR_L or VS and DE embeddings. However, GR_S was significantly outperformed by DE (SQ 4).

Error Categories Upon manual examination of translation results, six error categories emerged, see Appendix C for examples: 1) The syntactical nuances specific to Bernesen text persisted in the High German translation. 2) In instances of verb positioning, one verb (either 'lassen' or the modal verb) was omitted instead of being switched (not applicable for phenomenon 1). 3) The translation of the verb was inaccurate, resulting in a change in the sentence's meaning (particularly relevant for phenomenon 1). 4) The verb construction was not identified, leading to grammatically incorrect sentences or wrong meaning. 5) While the verb construction was acknowledged, the sentence conveyed a completely incorrect meaning. 6) Other errors that occurred rarely.

Embedding Pair	Score (%-p)	BS (%)
VS vs Rand.	14	**4.3**
GR_L vs Rand.	29	**.0**
DE vs Rand.	24	**.0**
GR_S vs Rand.	2	46.7
VS vs GR_S	12	**3.9**
VS vs GR_L	-15	**.4**
DE vs VS	10	5.6
DE vs GR_L	-5	24.8
DE vs GR_S	22	**.0**

Table 3: Scores for the Syntax Accuracy Differences between embedding initialization strategies, with P-values for the Paired bootstrap resampling significance test with 1000 resampling trials (BS) and difference of accuracy scores in %-points

E.	Rand.	VS	GR_L	GR_S	DE	All
1	24%	9%	13%	23%	11%	18%
2	30%	50%	13%	29%	21%	30%
3	9%	9%	40%	10%	11%	13%
4	24%	18%	33%	19%	32%	24%
5	9%	9%	0%	13%	16%	10%
6	3%	5%	0%	6%	11%	5%

Table 4: Distribution of errors by word embedding initialization (1 = Syntax remains, 2 = Verb dropped, 3 = Verb translation, 4 = Verb construction, 5 = Very wrong meaning, 6 = others)

The overall distribution of errors, both collectively and based on word embedding initialization, is shown in Table 4. It indicates that the most prevalent error is wrongly dropping a verb (Error 2) accounting for 30% of all errors, followed by a wrong verb construction (Error 4) with 24%. Further, the various embedding initializations exhibit variations in the types of errors they predominantly make. For example, VS embeddings maintain correct syntax in only 9% of cases, contrasting with random initialization at 24%. However, VS makes more errors in terms of incorrectly dropping a verb, with a rate of 50%, compared to 30% for random initialization. Refer to Appendix F for a breakdown of errors by phenomenon.

6 Discussion and Error Analysis

The results affirm that initializing with pretrained embeddings can enhance the handling of syntax-specific nuances in Bernese Swiss German to High German text-to-text machine translation. However, as anticipated, the effectiveness depends on the choice of pretrained embeddings.

SQ 1 Embeddings prove beneficial when trained on data with syntactical similarities to the source language (VS), large data quantity (DE), or medium data quantity combined with linguistic similarities beyond syntax to the source language (GR_L). General similarities to the source language alone do not improve performance (no significant difference of GR_S and baseline).

SQ 2 Furthermore, findings show that embeddings trained on data with syntactical similarities effectively facilitate the accurate handling of those particularities. This is evident in the superior performance of VS compared to GR_S. Additionally, an examination of error types reveals that VS exhibited the lowest error rate in preserving the syntactical structure of the source language at 9%, suggesting a more adept learning of syntactical nuances by the system.

SQ 3 In case of a trade-off between embeddings prioritizing data-driven learning or syntax similarities from a dialect close to the source translation, the former is more effective in handling syntax particularities, evident from GR_L outperforming VS embeddings by 15 percentage points.

SQ 4 The choice of embeddings from a dialect close to the translation system's source or target language does not significantly impact performance (no statistical difference between embeddings trained on VS to DE or GR_L to DE).

Syntax Phenomena The effectiveness of pretrained embeddings appears to differ across syntactical phenomena. They are most effective for addressing challenging syntax issues, exemplified by supporting phenomenon 2 the most which has the highest error rate with randomly initialized embeddings at 37% (1: 25% and 3: 23%). The system shows a preference for omitting a word over altering its order, evident in common errors for both verb positioning syntax phenomena (62% and 24%, respectively). Despite Phenomenon 1's lower occurrence in SwissDial training data, it exhibits a comparable error rate, reinforcing the difficulty of altering verb order. Additionally, the third phenomenon is handled more effectively, possibly due to increased occurrence in the training data. Another contributing factor may be the inclusion of 'lassen' in Phenomenon 1 and 2, intensifying the system preference to drop 'lassen'.

Overall Performance While GR_L does not exhibit a significant increase in syntactical performance compared to DE, it achieved the highest accuracy score of 87%. It also outperformed other initializations in overall translation quality, securing the highest BLEU and chrF2 scores. The disparity in syntactical handling can be attributed to two factors: 1) the limited number of test cases for syntax evaluation and 2) GR_L being the closest in linguistic features, beyond syntax, to the source language BE. VS has been identified as most distant to other Swiss German dialects in regard to vocabulary and other linguisitic differences (Paonessa et al., 2023).

7 Concluding Remarks

In this study, we fine-tuned the NLLB-200 model for translating Bernese Swiss German into High German and found that initializing the model with pretrained embeddings significantly improves syntax handling during translation. While embeddings from data similar to the source language (VS), larger data sets (DE), or a mix of medium data quantity and linguistic similarities (GR_L) are beneficial, embeddings trained on more data outperform the ones trained on syntax similar data. Additionally, embeddings are more helpful for challenging phenomena such as verb positioning.

However, it is essential to acknowledge that syntax nuances are not the sole determinants of machine translation performance. Future evaluations should consider further linguistic aspects and the training capacity of various embedding strategies, such as utilizing existing embeddings like DE to avoid unnecessary model training. Further, resource constraints, evident in a limited number of syntactical issues explored and a small synthetic test set. The latter poses challenges for meaningful significance tests, warranting cautious interpretation of results. The manual evaluation restricted a potential exploration of different models and leveraging their results. Future research could enhance robustness through multiple comparison hypothesis tests. Additionally, the presence of High German in the NLLB model is a noteworthy limitation potentially influencing the translation task impact.

References

Pelin Dogan-Schönberger, Julian Mäder, and Thomas Hofmann. 2021. Swissdial: Parallel multidialectal corpus of spoken swiss german. *arXiv preprint arXiv:2103.11401*.

Elvira Glaser. 2014. Wandel und variation in der morphosyntax der schweizerdeutschen dialekte. *Taal en Tongval: Lanquage Variation in the Low Countries*, 66.

Elvira Glaser. 2022. *Syntaktischer Atlas der deutschen Schweiz (SADS) - Band 1: Einleitung und Kommentar*. A. Franke Verlag, Tübingen, Germany.

Elvira Glaser and Gabriela Bart. 2015. *Dialektsyntax des Schweizerdeutschen. Regionale Variation des Deutschen - Projekte und Perspektiven*. Berlin, München, Boston: De Gruyter, Zürich, Switzerland.

Elvira Glaser and Peter Ott. 1997. Dialektsyntax: eine forschungsaufgabe. In *Bericht über das Jahr 1996. Schweizerisches Wörterbuch*, pages 11–32. Zürcher Druck + Verlag AG.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.

Lucy Linder, Michael Jungo, Jean Hennebert, Claudiu Cristian Musat, and Andreas Fischer. 2020. Automatic creation of text corpora for low-resource languages from the Internet: The case of Swiss German. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2706–2711, Marseille, France. European Language Resources Association.

Claudio Paonessa, Yanick Schraner, Jan Deriu, Manuela Hürlimann, Manfred Vogel, and Mark Cieliebak. 2023. Dialect transfer for swiss german speech translation. *arXiv preprint arXiv:2310.09088*.

Michel Plüss, Manuela Hürlimann, Marc Cuny, Alla Stöckli, Nikolaos Kapotis, Julia Hartmann, Malgorzata Anna Ulasik, Christian Scheller, Yanick Schraner, Amit Jain, et al. 2022. Sds-200: A swiss german speech to standard german text corpus. *arXiv preprint arXiv:2205.09501*.

Michel Plüss, Lukas Neukom, Christian Scheller, and Manfred Vogel. 2021. Swiss parliaments corpus, an automatically aligned swiss german speech to standard german text corpus. *arXiv preprint arXiv:2010.02810v2*.

Ye Qi, Devendra Sachan, Matthieu Felix, Sarguna Padmanabhan, and Graham Neubig. 2018. When and why are pre-trained word embeddings useful for neural machine translation? In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 529–535, New Orleans, Louisiana. Association for Computational Linguistics.

EDA Präsenz Schweiz. 2023. Die sprachen – fakten und zahlen. Eidgenössisches Departement für auswärtige Angelegenheiten EDA, Schweiz.

Guido Seiler, Sandro Bachmann, Johannes Graën, Nikolina Rajović, Adrian van der Lek, Ghazi Hachfi, Igor Mustač, Elvira Glaser, Peter Ranacher, and Robert Weibel. 2021. Syntaktischer atlas der deutschen schweiz online (sads online), iv.23. Deutsches Seminar / Linguistic Research Infrastructure / UFSP Sprache und Raum, Universität Zürich. `https://dialektsyntax.linguistik.uzh.ch/`.

Beat Siebenhaar and Alfred Wyler. 1997. *Dialekt und Hochsprache in der deutschsprachigen Schweiz*. Pro Helvetia, Zürich, Switzerland.

Abu Bakr Soliman, Kareem Eissa, and Samhaa R. El-Beltagy. 2017. Aravec: A set of arabic word embedding models for use in arabic nlp. *Procedia Computer Science*, 117:256–265.

A Hyperparameter for Fine-tuning NLLB-200 Model

Dialect	**Sentence**
Batch Size	16
Maximum token length	128
Warm-up steps	1000
Optimizer	Adafactor
Learning Rate	1e-4
Clip Threshold	1.0
Weight Decay	1e-3
Scale Parameter	False
Relative Step	False

Table 5: Hyperparameters of the NLLB-200 Model for finetuning

B Example Sentences with Syntactical Phenomena in SwissDial Training Dataset

Dialect	Sentence
High German (DE)	Ich wühle mich währenddessen durch die alten Akten und gucke, ob sich da nicht etwas finden lässt.
Bern (BE)	I wüelemi währenddessä dür die autä Aktä und luege, obsech dä nid öpis **lat la** fingä.
Wallis (VS)	Ich wiehlu mi währenddessu durch d'altu Akte und lüegu, ob schich da nit eppis **laht lah** findu.
Graubünden (GR)	I wüahla mi währenddessa dur dia alta Akta und luaga, ob sich do öpis finda loht.
Aargau (AG)	Ich wüehle mech währenddesse dor die alte Akte ond luege, obsech dete ned öppis **loht loh** fende.
Basel (BS)	Ich wüehl mi währenddesse durch die altä Akte und lueg, ob sich do nid öbbis finde losst.
Luzern (LU)	Ech schnoigge währendem i de alte Akte ond luege, öb ech öppis fende.
St. Gallen (SG)	I wüehl mi währenddesse dur di alte Akte und lueg, ob sich do nöd öpis finde loht.
Zürich (ZH)	Ich wüehle während däm dur di alte Akte und lueg ob sich da no öpis finde laht.

Table 6: Example of Verb Doubling (Phenomenon 1) in SwissDial training dataset (ID: 1736)

Dialect	Sentence
High German (DE)	Im Gegenzug werde er die Klage fallenlassen.
Bern (BE)	Im Gägäzug wirder d Chlaag **la gheiä**.
Wallis (VS)	Im Gäguzug wärde är d'Chlag **lah kiju**.
Graubünden (GR)	Im Gegazuug wür er d Klag keia loh.
Aargau (AG)	Em Gegezög wärdi er d'Klag falle lo.
Basel (BS)	Im Gegezug wird är d Klag falle loh.
Luzern (LU)	Em Gägezög werd är d Chlag gheie loh
St. Gallen (SG)	Im Gegezug wird er dChlag falle loh.
Zürich (ZH)	Im Gägezug werdi er d Klag falle lah.

Table 7: Example of Verb Positioning 1 (Phenomenon 2) in SwissDial training dataset (ID: 422)

Dialect	Sentence
High German (DE)	Das geht aber nur gut, wenn man die eigenen Daten ohne Probleme aus der Hand geben kann.
Bern (BE)	Das geit aber numä, wemä die eigetä Datä ohni Problem us dr Hand **cha gä**.
Wallis (VS)	Das geit aber nit nur, wemu d'eigenu Date ohni Problemi üser Hand **chänne gäh**.
Graubünden (GR)	Das goht aber nur guat, wemma dia eigna Data ohni Probelm us dr Hand geh kann.
Aargau (AG)	Das goht aber nome guet, wennmer die eigete Date ohni Problem os de Hand geh chan.
Basel (BS)	Das goht aber nur gued, wenn me die eigene Date ohni Problem us dr Hand gäh ka.
Luzern (LU)	Das gohd aber nor guet, wenn mer di eigete Date ohni Problem us de Händ **cha gäh**.
St. Gallen (SG)	Da goht aber nur guet, wenn me di eigete Date ohni Problem us de Hand **cha geh**.
Zürich (ZH)	Das gaht aber nur guet, wemmer di eigene Date ohni Problem usde Hand geh chan.

Table 8: Example of Verb Positioning 2 (Phenomenon 3) in SwissDial training dataset (ID: 24)

C Example Errors of Translated Sentences

Error	Sentence
1 Syntax remains	Sie kann die Hunde durch das ganze Meer schwimmen lassen. (DE) Sie cha d Hüng dürs ganzä Meer la schwümmä. (BE) Sie kann die Hunde durch das ganze Meer la schwimmen. (Random Embedding)
2 Verb dropped	Die Frage ist, was noch helfen dürfte. (DE) D Fraag isch, was no dörft häufä. (BE) Die Frage ist, was noch hilft. (VS Embedding)
3 Verb translation	Er fängt bald an arbeitsunfähig zu sein.(DE) Er fat gli aafa arbeitsunfähig z si. (BE) Er ist fett bald arbeitsunfähig zu sein.(Random Embedding)
4 Verb construction	Dienstags und samstags fängt der Friseur an zu öffnen. (DE) Am Zischtig und Samschtig fat dr Coiffeu aafa z uftuä. (BE) Am Dienstag und Samstag beginnt der Friseur mit der Aufarbeitung. ($\mathrm{GR_L}$ Embedding)
5 Very wrong meaning	Nie würde ich meine Katze ertrinken lassen. (DE) Ni würdi mini Chatz la etrinkä. (BE) Noch eine Katze würde mich ausdrinken lassen. (Random Embedding)

Table 9: Example of Error Categories of the Translation Results

D Validation and Training Loss Curve

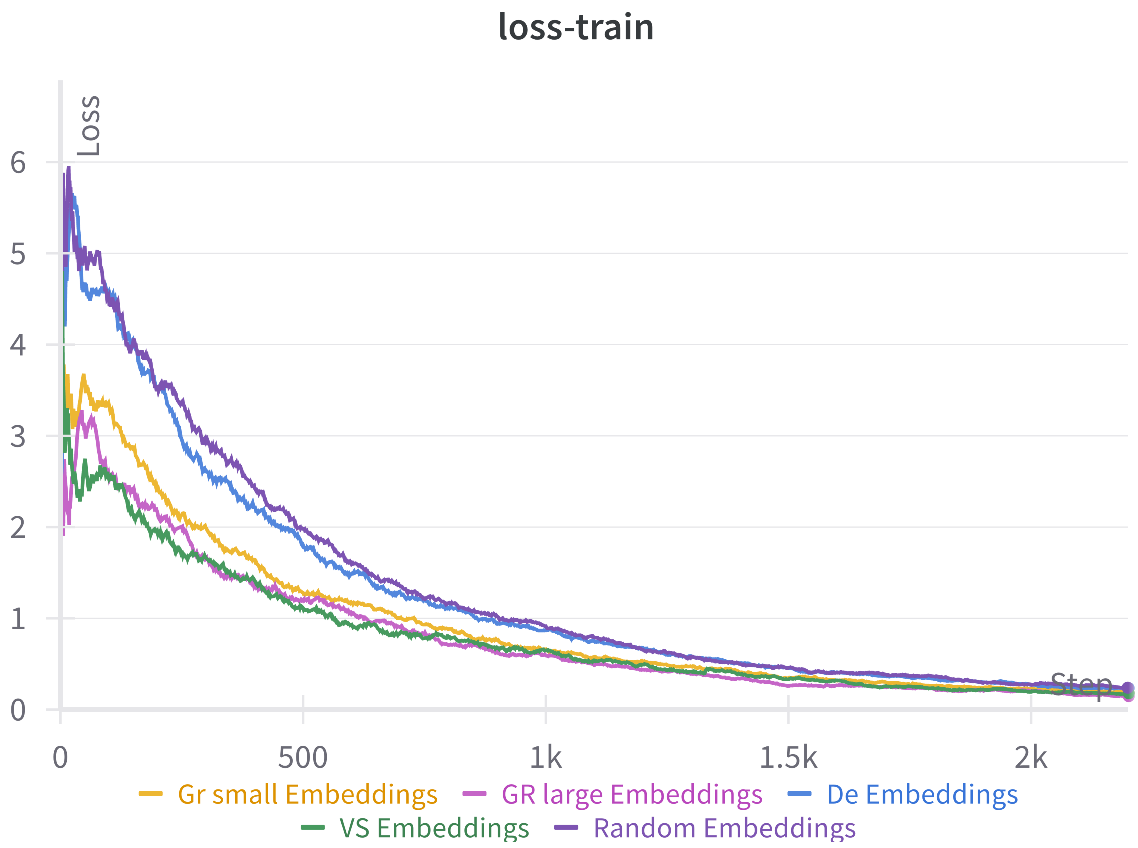

Figure 2: Training Loss of the NLLB-200 models using different word embedding initialization (*Smoothing with exponential moving average, parameter of 0.96*)

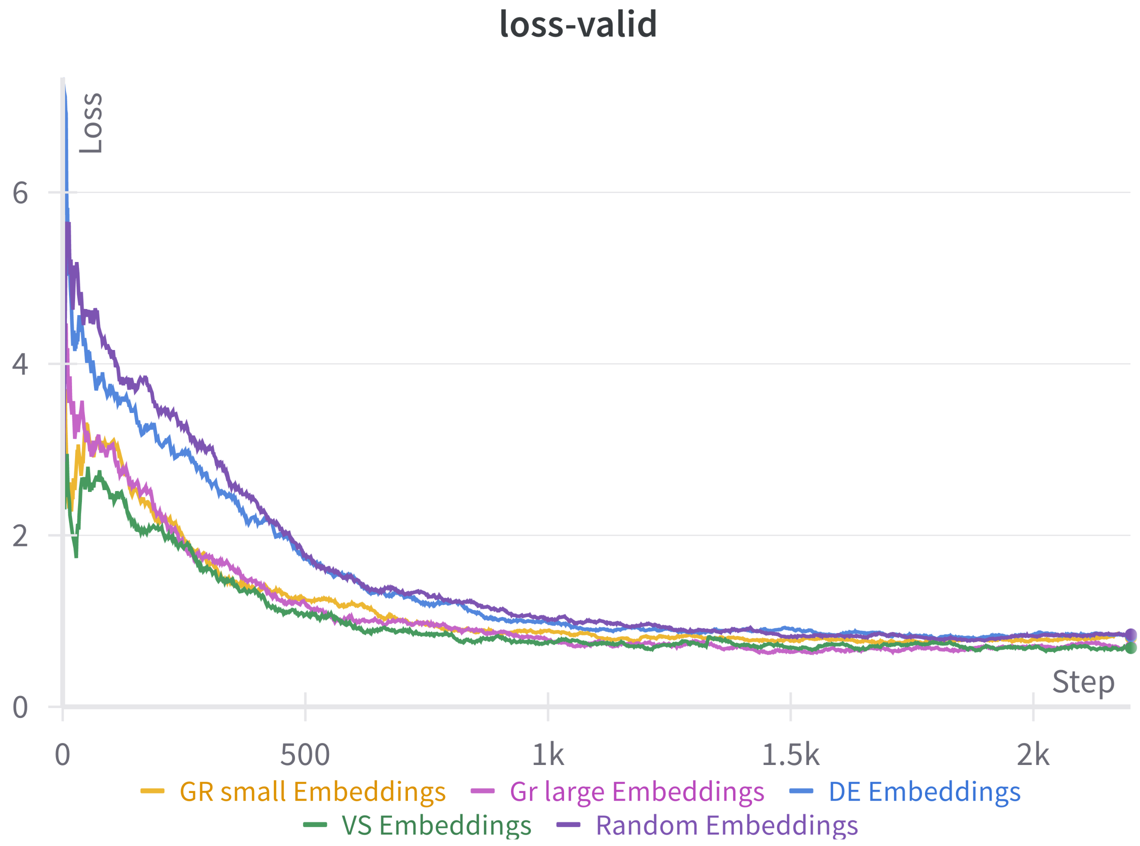

Figure 3: Validation Loss of the NLLB-200 models using different word embedding initialization. (*Smoothing with exponential moving average, parameter of 0.96*)

E Overall Translation Quality Score

Embeddings	Metric	Score	BS (%)
Random Baseline	BLEU	51.7	N/A
	chrF2	77.1	N/A
	TER	31.2	N/A
VS	BLEU	51.1	20.48
	chrF2	76.7	20.28
	TER	30.4	17.48
GR_L	BLEU	55.8	**.50**
	chrF2	79.4	**.60**
	TER	26.9	**.10**
DE	BLEU	51.8	40.76
	chrF2	76.1	10.09
	TER	31.4	35.16
GR_S	BLEU	55.6	**1.10**
	chrF2	78.3	6.79
	TER	29.1	**4.70**

Table 10: The scores for the different evaluation metrics, for the different translation models, with P-values for the Paired bootstrap resampling significance test with 1000 resampling trials (BS)

Embedd.	Metric	Score	AR (%)	BS (%)
VS Baseline	BLEU	51.1	N/A	N/A
	chrF2	76.7	N/A	N/A
	TER	30.4	N/A	N/A
GR_S	BLEU	55.6	**.54**	**.20**
	chrF2	78.3	**4.53**	**2.80**
	TER	29.1	25.92	11.49

Table 11: The scores for the different evaluation metrics, for the different translation models, with P-values for two different significance tests: Paired approximate randomization test with 10000 trials (AR) and Paired bootstrap resampling test with 1000 resampling trials (BS)

Embedd.	Metric	Score	AR (%)	BS (%)
VS Baseline	BLEU	51.1	N/A	N/A
	chrF2	76.7	N/A	N/A
	TER	30.4	N/A	N/A
GR_L	BLEU	55.8	**.18**	**.20**
	chrF2	79.4	**.16**	**.20**
	TER	26.9	**.27**	**.40**

Table 12: The scores for the different evaluation metrics, for the different translation models, with P-values for two different significance tests: Paired approximate randomization test with 10000 trials (AR) and Paired bootstrap resampling test with 1000 resampling trials (BS)

Embedd.	Metric	Score	AR (%)	BS (%)
DE Baseline	BLEU	51.8	N/A	N/A
	chrF2	76.1	N/A	N/A
	TER	31.4	N/A	N/A
VS	BLEU	51.1	59.23	20.88
	chrF2	76.7	44.13	14.99
	TER	30.4	41.51	15.88
GR_L	BLEU	55.8	**2.08**	**1.10**
	chrF2	79.4	**.10**	**.20**
	TER	26.9	**.11**	**.40**
GR_S	BLEU	55.6	**2.93**	**1.70**
	chrF2	78.3	**1.47**	**1.00**
	TER	29.1	7.73	**4.50**

Table 13: The scores for the different evaluation metrics, for the different translation models, with P-values for two different significance tests: Paired approximate randomization test with 10000 trials (AR) and Paired bootstrap resampling test with 1000 resampling trials (BS)

F Translation Errors by Phenomenon

Error	Phen. 1	Phen. 2	Phen. 3	All
1	22%	2%	32%	18%
2	-	62%	24%	30%
3	17%	9%	15%	13%
4	39%	16%	18%	24%
5	12%	7%	12%	10%
6	10%	4%	0%	5%

Table 14: Distribution of errors by syntactical phenomenon ((1 = Syntax remains, 2 = Verb dropped, 3 = Verb translation, 4 = Verb construction, 5 = Very wrong meaning, 6 = others; Phen. 1 = verb doubling, Phen. 2 = verb positioning of "lassen", Phen. 3 = modal verb positioning)

G Group Contributions

Project Part	**Group Member**
Download of Dataset	Andreas
Pre-processing	Andreas and Mira
Model Research	Andreas and Mira
Fine-tuning of NLLB-200 Model	Mira
Generating of Pretrained Embeddings	Mira
Development of Hypothesis	Mira and Linda
Creation of Synthetic Test Set	Linda
Manual Syntax Error Analysis	Linda
Statistic Tests	Mira
Report: Introduction, Related Work, Data	Linda
Report: Methodology	Andreas and Linda and Mira
Report: Results	Linda
Report: Discussion and Error Analysis	Linda
Report: Formating and Citations	Andreas and Linda

Table 15: Group Contributions by Project Part