Abstract

Dialectology is concerned with the study of language variation across space. While dialect atlases and dictionaries have been produced over the last 150 years for almost all linguistic areas of Europe, recent dialectological research increasingly focuses on corpus-based approaches. However, carrying out quantitative studies with dialect corpora has proven challenging because corpus data are not directly comparable. If informant A does not use word x, this does not necessarily mean that the word does not exist in A's dialect. It may just be that A chose to talk about topics that did not require the use of word x. This project proposes a new take on corpus-based dialectology that relies on automatic normalization to provide comparability across dialects.

Normalization is defined as the annotation of every dialectal word with a canonical word form, for example the standardized spelling of the word. It disambiguates dialectal word forms and provides a basis of comparison of different dialects. Automatic normalization can be viewed as a particular case of machine translation. The first goal of the project will be to improve current normalization methods with techniques from state-of-the-art neural machine translation.

Normalization introduces comparability in dialect corpora. In particular, the parameters of the normalization models provide a condensed and abstract representation of the normalization process, which allows us, for example, to investigate the status of particular characters in different dialects and to test the validity of traditional dialectal classifications. The second goal of the project will thus be to extract, visualize and interpret dialectal patterns emerging from the normalization models.

The third goal of the project is to investigate to what extent user-generated content (UGC), i.e. texts published by diverse users on social media platforms, contains dialectal signals. We will collect UGC data and contrast them with existing dialect corpora, again using normalization methods to provide comparability.

The experiments will initially focus on Swiss German and Finnish dialects, for which relevant resources are available. We will extend our investigations to other dialect areas yet to be defined. The results of this research, obtained through the unique combination of machine learning methods and spontaneously occurring data, will yield new visualizations of dialect landscapes, showcasing the richness of linguistic variation.

1 Aim and objectives

1.1 Significance of the research project in relation to current knowledge, premise underpinning the research:

Dialectology is concerned with the study of language variation across space. Dialectological inquiries over the last 150 years have led to a wealth of dialectological atlases and dictionaries. Most of this work followed a traditional approach where linguistic items were considered out of their context of usage. More recently, dialectological research has increasingly focused on corpus-based approaches. Dialect corpora are typically compiled by transcribing semi-directed interviews between a researcher and an informant, with the goal of obtaining more realistic, everyday speech (Szmrecsanyi & Anderwald 2018).

Recent quantitative studies in dialectology, generally subsumed under the name *dialectometry* (Goebl 2010; Wieling & Nerbonne 2015), focus almost exclusively on atlas data, due to their more fine-grained geographical coverage and more systematic presentation of results. Dialect corpora, on the other hand, are mostly used for qualitative research. They do not lend themselves well to quantitative studies because the different interviews are not directly comparable (Goebl 2005). If informant A does not use word *x*, this does not necessarily mean that *x* does not exist in A's dialect. It may just be that A chose to talk about topics that did not require the use of *x*. **In this project, we fix this methodological problem. Our underlying premise is that dialect corpora can be made comparable, enabling their quantitative analysis.**

In an entirely different area, namely historical computational linguistics and digital humanities, researchers working on historical texts have been confronted with massive spelling variation over time, due to changing orthographic conventions. Their answer lies in **text normalization, where each text is annotated with an additional layer that contains a standardized spelling for each word.** This facilitates keyword search in historical corpora and allows historians and linguists to obtain reliable frequency information for the keywords they are interested in (Hämäläinen et al. 2018). Most recent approaches assume that normalization can be framed as a particular type of machine translation from old to modern spelling (Pettersson et al. 2014; Scherrer & Erjavec 2016; Bollmann 2019).

The first aim of this project is to enable corpus-based dialectology by introducing comparability through text normalization. In contrast to Szmercsanyi (2013), who employs syntactic annotation to ensure comparability in British English dialect corpora, we will apply text normalization techniques to annotate dialectal texts with standard spellings (Scherrer & Ljubešić 2016; Scherrer et al. 2019). This new take on corpus-based dialectology will allow us to focus on other linguistic levels such as phonology, morphology, and the lexicon. In particular, the **parameters of the normalization models provide a condensed and abstract representation of the normalization process**. The **second aim** of the project is to analyze and interpret these representations and compare them with results from atlas-based dialectology and from representations obtained without normalization.

Dialect corpora are costly to produce: informants need to be found and interviewed, and the recorded interviews need to be transcribed as consistently as possible. To circumvent this data bottleneck, researchers have increasingly turned to **user-generated content** (UGC) in recent years. UGC typically refers to texts published by diverse users on social media platforms such as forums, internet chat rooms, messaging applications and microblogging services. UGC is typically written, but frees itself from most orthographic and stylistic norms of the written genre to resemble spontaneous speech (Crystal 2011). These properties make UGC particularly well-suited for the study of linguistic variation. Indeed, there is a growing body of research on the analysis of dialectal variation in UGC (e.g. Siebenhaar 2006; Eisenstein et al. 2014; Ljubešić et al. 2015; Hovy & Purschke 2018; Grieve et al. 2019).

The **third aim** of the project builds on these recent developments. Although UGC data are inherently noisier and contain less dialect-specific signals that purpose-built corpora, the recent case studies show that it is possible to extract meaningful patterns of language variation from such datasets. However, the popularity of particular social media services varies widely by country, and the linguistic structures and devices used in such informal communication also differ across language areas, such that results obtained for one language area cannot be reproduced easily for another one. We will investigate this aspect by **collecting UGC data and contrasting them with existing dialect corpora**, using text normalization methods to provide comparability.

The objectives of the project are thus:

1) to improve the automatic normalization of dialect texts by using state-of-the-art machine translation methods,
2) to extract, visualize, compare and interpret the dialectal patterns emerging from the normalization models, and
3) to use these techniques to contrast the dialectal patterns found in purpose-built corpora with those of user-generated content.

The combination of machine learning methods and spontaneously occurring data provides unique opportunities to renew the research methodologies in dialectology. The results of this research will yield new visualizations of dialect landscapes, displaying the richness of linguistic variation. Thereby, they can strengthen the self-esteem of dialect speakers and reduce potential negative stereotypes related to dialectally marked speech practices.

1.2 Research questions and/or hypotheses:

1. Extended context and new neural network architectures improve text normalization.

Normalization is the annotation of every dialectal word with a canonical word form, for example the standardized spelling of the word. Normalization disambiguates dialectal word forms and provides a basis of comparison for different dialects, as illustrated in the following example from the *Samples of Spoken Finnish* corpus (Institute for the Languages of Finland, 2014):

Dialectal transcription:	ja	se	kuali	siälä	,	AmeriikkàsA	.
Normalization:	ja	se	kuoli	siellä	,	Ameriikassa	.
English gloss:	and	he	died	there	,	in America	.

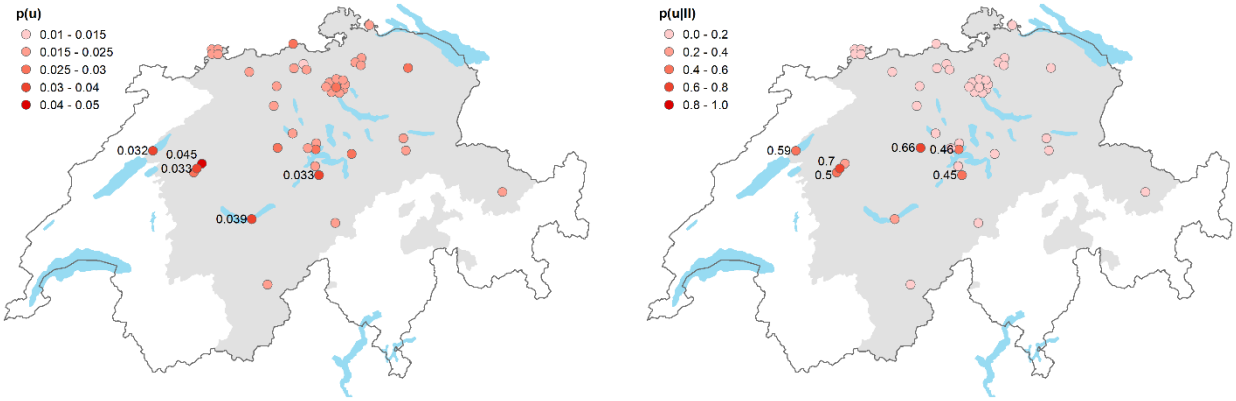
As soon as an initial set of manually normalized utterances is available, automatic normalization tools can be trained using machine learning. The normalization task is typically framed as a case of machine translation where every dialectal utterance is "translated" into its respective normalized utterance. From a wide variety of explored approaches, one of the most successful has been character-level statistical machine translation (**CSMT**, cf. Tiedemann 2009; Scherrer & Erjavec 2016). In this paradigm, the dialectal utterance is translated character by character into its normalized form. In the last five years however, researchers in machine translation have almost completely abandoned the statistical paradigm in favor of models based on deep neural networks. While some success has been reported on the normalization task with neural machine translation methods (**NMT**, cf. Lusetti et al. 2018; Partanen et al. 2019), it has been much more modest than in other applications of machine translation.

We see two main reasons for this lack of progress. First, many experimental setups normalize each word in isolation, without taking context into account. This is an unrealistic setup that hampers normalization accuracy for ambiguous words. Second, currently used NMT approaches tend to be too powerful for the text normalization task and do not perform well with limited amounts of training data. This problem can be

tackled from two angles, either by choosing simpler model architectures or by artificially generating synthetic training data. Both ideas have been successfully exploited in recent research in language technology.

2. Patterns of dialectal variation that emerge from normalization models correlate with traditional dialectological research results.

Normalization models learn correspondences between characters or sequences of characters. Since the frequency distributions of these correspondences vary across dialects, the normalization can serve as a basis for comparisons between dialects (Scherrer et al. 2019). An example may illustrate this: in some Swiss German dialects, /l/ becomes /u/ in certain phonological contexts. To define this vocalization area geographically, it is not sufficient to compute the frequency of /u/ in each text (left figure below), because /u/ occurs in other phonological contexts in all dialects. Normalization allows us to define phonological contexts easily and hence to restrict our search to those occurrences of /u/ that appear in /l/-vocalization contexts. **As a result, we obtain a clearer and more accurate picture of the geographical extent of /l/-vocalization** (right figure below).



We will investigate how such patterns of change can be automatically extracted from normalization models. Models based on the CSMT paradigm use phrase tables that provide frequency information about correspondences, whereas neural models use attention matrices that indicate which dialectal characters are most important when deciding about the normalized characters to generate.

A major advantage of current NMT architectures is their ability to combine multiple language pairs and translation directions within the same model (Johnson et al. 2017). Such multilingual models benefit most from combinations of related languages, and *a fortiori* of dialects. Another advantage of NMT is that the linguistic material (characters, graphemes or words, but also dialect and language identifiers) is "embedded", i.e. converted to high-dimensional numeric representations. Multi-dialectal normalization models therefore infer **implicit representations of the different dialects**, which can then be visualized and compared with traditional atlas-based dialect classifications, analogously to earlier work in computational language typology (Östling & Tiedemann 2017, Abe et al. 2018).

Patterns of dialectal variation are emerging properties of the normalization models. We expect them not only to yield dialect classifications that are coherent with traditional approaches, but also to detect when and where diachronic change has occurred.

3. User-generated content contains dialectal signals discernible through the use of normalization.

We assume that corpora extracted from social media are noisier than purpose-built dialect corpora, but that they still contain dialectally differentiated signals that can be analyzed. We will test this hypothesis by

contrasting UGC corpora with existing dialect corpora. In this setting as well, we expect text normalization to play a crucial role in providing comparability across different users and messages.

UGC contains various types of non-standard features: specific symbols and lexical items (emojis, hashtags, etc.), absence of orthographic norms, creative spellings (abbreviations, letter repetitions, etc.), syntactic and pragmatic structures related to spoken language. While the lack of orthographic norm precisely enables the use of dialectal markers, other non-standard features are irrelevant for the investigation of regional variation. A major part of our work will thus be concerned with **teasing apart those non-standard features that are related to dialectal variation from those that are not**.

1.3 Expected research results and their anticipated scientific impact, potential for scientific breakthroughs and for promoting scientific renewal:

Although most recent research in dialectology is based on quantitative methods, the use of neural networks is largely unknown, just as the use of text normalization as a tool to introduce comparability in corpora. This project will **provide dialectologists with new tools for their research**.

Dialects are – almost by definition – *low-resource languages*. We **test the capabilities of data-driven methods in such challenging settings** that contrast starkly with the 'big data' approaches typically used in language technology. In a research area where a large number of publications relies on data sources that are only available for English, our project can stand out in multiple ways and showcase the importance of research for linguistic varieties with few resources – including most of the languages of Europe – and in some cases low social status.

Text normalization is mostly considered an auxiliary task that enables research in digital humanities or linguistics. As a result, normalization does not get a lot of publicity in language technology venues, and cutting-edge research in machine translation does not immediately find its way to the normalization community. As a research group strongly embedded in a machine translation background, we intend to change that. We aim to **establish normalization** – for historical, dialectal or social media texts – **as a task on its own** in yearly research competitions.

With the emergence of deep neural networks, researchers in language technology have become increasingly concerned with their interpretability. The visualization and interpretation of internal model parameters fits well into this thriving research field commonly called *representation learning*. Whereas most work investigates representations of words and sentences, this project innovates through its **focus on smaller-scale units such as characters and morphemes, which are more relevant for dialectology**. On the other extreme, we also investigate the emerging representations of dialect areas and compare them with findings from existing dialectological research.

1.4 Special objective of call (concerns Academy Programmes and other thematic calls):

None

2 Implementation

2.1 Work plan and schedule:

The three main research hypotheses formulated above also guide the work plan. A fourth work package is included to foster international collaboration and dissemination.

1. Set up improved text normalization models

We will start by collecting a number of existing benchmark datasets. Depending on the interests of the research group members, additional datasets requiring curation or manual annotation may be chosen (**1a**). We will characterize the corpora in terms of variation types and frequencies to guide the modelling choices for text normalization. We will then implement and evaluate different model architectures and training setups on these corpora (**1b**). We will also investigate the feasibility of bootstrapping approaches (Scherrer & Erjavec 2016) for datasets that do not contain a manually annotated normalization layer. Although the project itself deals with dialectal variation, the benchmark datasets may include historical language variation data to assess the generalization capabilities of the models and to increase the visibility of the results. Work packages **1a** and **1b** are scheduled for the first two years of the project. As deliverables of this work package, we plan to publish two papers on our normalization experiments in language technology venues. These could include major conferences like ACL or COLING or international workshops like WMT or VarDial, depending on their timeline.

2. Extract, visualize, compare and interpret dialectal variation patterns emerging from normalization models

When the normalization models are stabilized, we will proceed to extract dialectal variation patterns from the normalization models trained on the various datasets, in years 2 and 3 (**2a**). In parallel with the normalization models created as part of work package **1a**, we intend to create contrastive models that are trained only on the dialectal transcriptions, but not on the normalization, relying on the widely used BERT training procedure (Devlin et al. 2019). This will allow us to assess the impact of normalization on the quality of the emerging variation patterns (**2b**). Visualization of our findings will be provided in years 3 and 4 (**2c**). Deliverables are one or two publications on this topic in dialectology-oriented publication channels, and the visualization web site.

3. Extend investigations to user-generated content

With the foundations of our approach in place, we will tackle the most challenging part of the project: UGC data are noisier and it is harder to extract dialectal signals from them. Collection, compilation and (if required) annotation of UGC corpora will already start in year 1 to make sure that the datasets are available when they are required for the experiments (**3a**). The application and adaptation of the normalization and feature extraction pipeline to UGC data (**3b**) is scheduled for years 3 and 4. The visualization tools will be updated with UGC results in year 4 (**3c**). This work package will lead to one publication about the collected resource(s) and one about the normalization and dialectology experiments.

4. International collaboration and dissemination

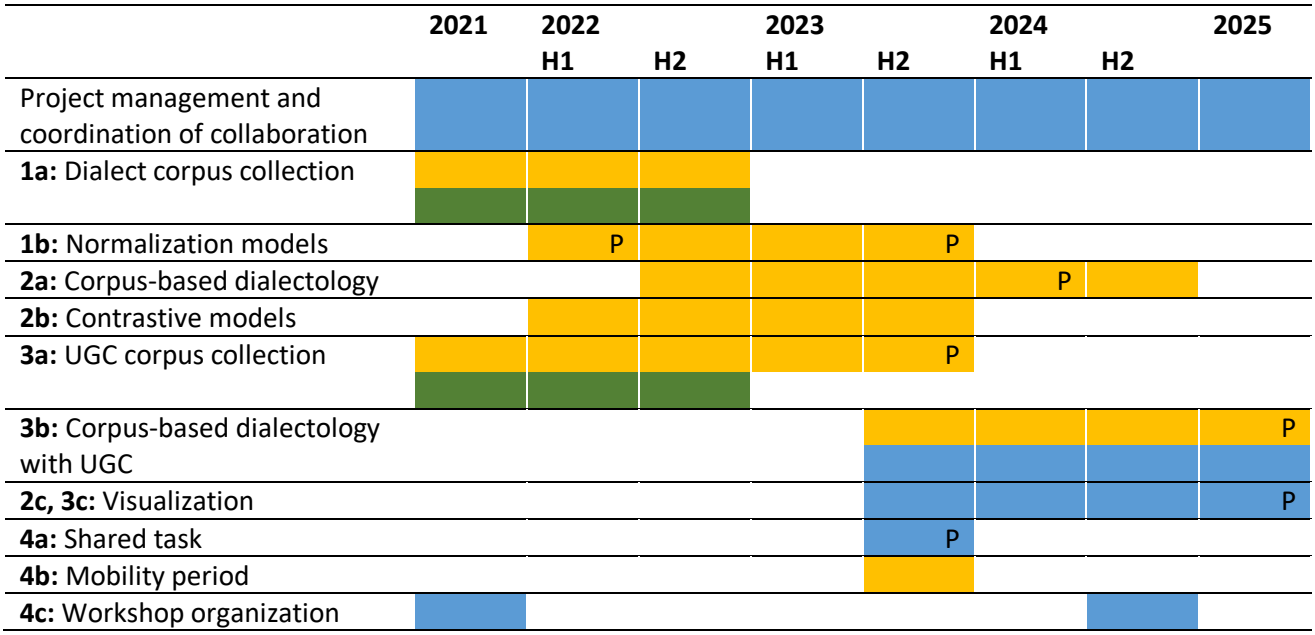
In order to foster interest in dialect text processing and normalization, we plan to **organize a shared task on text normalization or dialect identification**, centered around the datasets used in the project (**4a**). Two venues are potential candidates for such a shared task: the *Conference on Machine Translation (WMT)* hosts yearly competitions related to machine translation, and the *Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)* hosts yearly competitions related to the identification and annotation of

similar language varieties. We plan to submit a shared task proposal in the third year of the project. If accepted, the shared task will be described and summarized in an overview paper.

In order to promote the career prospects of the post-doctoral researcher, financial provisions are made to fund a six-month mobility period (**4b**). The host institution and the exact date will be defined to comply with the language areas under investigation.

We will organize a workshop at the beginning of the project with all team members and collaborators to guide the first stages of the project. A second workshop or colloquium is planned towards the end of the project (**4c**).

The chart below summarizes the different work packages and the responsibilities of the team members.



Blue: principal investigator; yellow: post-doc; green: research assistant. *P*: planned publications.

2.2 Research data and material, methods, and research environment

Research data

The primary material to be used in this project are dialect corpora, i.e. collections of texts annotated with their geographical provenience. In order to train and evaluate normalization models, parts of the corpora have to be annotated manually with word-level normalizations. Initially, the research within this project will rely on the following two dialect corpora:

- The *ArchiMob* corpus of Swiss German (Scherrer et al. 2019) contains transcriptions of 43 interviews with informants from different regions of German-speaking Switzerland. It is partially annotated with normalizations and is freely available for research purposes.
- The *Samples of Spoken Finnish* corpus (Institute for the Languages of Finland, 2014) contains transcriptions of interviews with informants from 23 areas of the Finnish language area. The entire corpus is normalized. It is freely available for research purposes from the Language Bank of Finland.

Two project collaborators, Janine Siewert and Noëmi Aepli, will work with their own datasets of Low Saxon (Siewert et al. submitted) and Swiss German, according to their respective research plans. We may reuse these datasets within the project. Furthermore, we plan to include one or two additional corpora from other linguistic areas. Through existing collaborations of the PI, we will be able to obtain datasets for German, Russian, Arabic and French dialects and regional varieties, but the final decisions will be made in

accordance with the interests of the other project members. Depending on the datasets, manual curation and/or annotation may be required.

In order to evaluate the viability of corpus-based dialectology, the results obtained from these corpora need to be compared to some ground truth. For Swiss German, the *Linguistic Atlas of German-speaking Switzerland* (SDS) and its derived quantitative analyses (Hotzenköcherle et al. 1962-1997; Scherrer & Stoeckle 2016) can serve as ground truth. For Finnish, the *Dialect Atlas of Finland* is available in various electronic formats (Kettunen 1940; Embleton & Wheeler 1997). Similar resources are available for other language areas.

The third work package requires text collections of user-generated content. Such collections can be compiled from Twitter, Jodel or similar platforms, taking advantage of the geographical coding added to the messages by the users' devices (Ljubešić et al. 2016; Hovy & Purschke 2018; Grieve et al. 2019). Pre-compiled UGC corpora exist for various language areas and could be reused in our project if the sharing conditions permit. The exact set of languages and data sources will be defined at the beginning of the project, and manual annotation may be required to provide training and test data for the normalization experiments.

Research methods - Text normalization

We will focus both on statistical (CSMT) and neural (NMT) approaches to text normalization. We propose three directions to improve the current state of the art.

First, we make sure that the normalization models have access to sentential context. Word-by-word normalization models that are inaccurate for ambiguous words by design, are still common nowadays. We have already proposed sentence-level normalization in the context of CSMT (Scherrer & Ljubešić 2016), and similar extensions can also be envisaged in NMT models (Partanen et al. 2019). Furthermore, a recently proposed approach to learn contextualized string embeddings (Akbik et al. 2018) has the potential to provide an adequate basis for modelling the contextual dependencies in text normalization.

Second, the text normalization task can be viewed as a particular type of the noisy channel paradigm with a simple channel model but a powerful language model. Neural machine translation architectures do not follow the noisy channel paradigm and therefore cannot capture this imbalance between models satisfactorily. We propose on the one hand to choose a model architecture that mimics the simple channel model, such as the recently proposed Levenshtein transformer (Gu et al. 2019). On the other hand, we aim to strengthen the language modelling part by generating synthetic data (Sennrich et al. 2016). This is an established method in machine translation, but has not yet been explored for normalization.

Third, we intend to address the data bottleneck caused by supervised normalization models, which require large, manually annotated datasets for training. Scherrer & Erjavec (2016) showed that it is possible to create normalization models without manual annotation in an unsupervised way, and we will update this technique to current NMT-based methods and compare it with supervised approaches.

Research methods - Corpus-based dialectology

The main assumption of the project is that normalization models trained on dialect corpora are able to pick up dialectological regularities and distill them in their parameters. We plan to devise methods to extract local as well as global features from normalization models. Local features are obtained from separate normalization models for each dialect, whereas global features are based on a single model that encompasses data from all dialects.

Local features give answers to concrete questions about character correspondences and sequences: in which dialects is the /u/-/l/ correspondence most frequent? In which dialects is this correspondence restricted to intervocalic contexts? What is the distribution of diminutive suffixes across dialects? Local features are mainly based on alignment and frequency information that is computed during the training process of CSMT-based normalization systems, but similar information can be extracted from attention matrices in neural models. We will use statistical techniques to find the most characteristic features for each dialect and aggregate them to provide numerical scores of "dialectality". The obtained results will then be compared with traditional dialectological knowledge.

Global features are more abstract representations that emerge from NMT-based multi-dialectal normalization models. They can answer more general questions about similarities and differences between dialects, but without being able to trace them back to particular words or structures (Abe et al. 2018): does /u/ behave more like a vowel or like a consonant? Which dialects are most similar from a normalization point of view? Which are the dialects whose vowel spaces differ most? These results will be correlated with existing dialect classifications if available, or with classifications inferred automatically using traditional methods such as hierarchical clustering.

The impact of normalization on the precision of the emerging dialect features can be assessed through the comparison with similar training setups that do not make use of normalization. BERT (Devlin et al. 2018) is an example of such a setup. It is trained to predict the items (words, parts of words, or characters) that have been intentionally masked in the input. For example, a source sentence could contain three masked items *"ja [redacted] kuali siä[redacted], Ame[redacted]àsA"*, and BERT would be tasked to restore the complete sentence *"ja se kuali siälä, AmeriikkàsA"*. The internal representations of BERT models have been shown to be competitive for a wide range of tasks. We will contrast the normalization-based models with BERT models trained on the same datasets.

Research infrastructure

The development of text normalization algorithms is computing-intensive: neural models require the use of GPUs during training, and preprocessed data and trained models require large amounts of storage space. We will rely on the infrastructure provided by the Finnish Center for Scientific Computing (CSC) for computing resources and storage needs. The CSC infrastructure has proved satisfactory in earlier research of the PI; it is free of charge for Finland-based research groups. The visualization web pages will be hosted on servers provided either by CSC or by the University of Helsinki.

The access to existing dialect corpora will be facilitated by international data provision services such as CLARIN, or more locally, the Language Bank of Finland (Kielipankki).

2.3 Risk assessment and alternative implementation strategies:

- **Recruitment difficulties (likely):**
  Most of the work will be carried out by a post-doctoral researcher, who should ideally have previous experience both in machine translation and variational linguistics. At a time where numerous companies and research institutes vie for graduates with knowledge in artificial intelligence, finding a motivated and competent person for this project may prove challenging. We will keep the recruitment requirements as open as possible and also consider PhD candidates. As a last resort, the PI could reduce his lecturer duties to take a more active role in project-related research. These difficulties may negatively affect the timely completion of work packages, especially in the first half of the project.
- **Pandemic-related restrictions and delays (possible):**
  Recruiting personnel from abroad may be delayed by pandemic-related travel restrictions. The

choice of languages and corpora is partially determined by the main researcher to be recruited and may get delayed as well.
A first workshop with external collaborators is planned for autumn 2021, but can be postponed to 2022 or organized online should the situation require it.

- **Data access restrictions (possible):**
  The legal status of many linguistic corpus collections is unclear, such that some corpora available today may become unavailable at short notice. However, the goals of the project are largely language-independent and can still be achieved by backing off to datasets from different languages.
- **Absence of expected results (unlikely/possible):**
  The success of the proposed method will be judged by its correlation with traditional methods and by its ability to predict well-known patterns of dialectal variation. It is possible that such correlations do not obtain, or that the predicted variation patterns cannot be linked sensibly to dialectological knowledge. Preliminary studies (Abe et al. 2018, Scherrer et al. 2019) suggest that a complete absence of results is unlikely, but partially negative results can occur. Such negative results may hint at unforeseen factors of dialectal variation and are in fact particularly relevant from a dialectological point of view.
  We may also find evidence that invalidates our initial hypothesis that normalization is crucial for providing comparability between dialects. This would also be an interesting dialectological finding, although it will not challenge the normalization task *per se*, which remains relevant for other purposes.

3 Research team and collaborators

3.1 Project personnel and their relevant merits:

Principal investigator: Yves Scherrer, PhD, University lecturer

The proposed project follows directly from different strands of the PI's current research. Yves Scherrer has been involved in various projects in the areas of language technology, dialectology, and corpus linguistics. His current research focuses on the automatic analysis and annotation of variational data (such as dialectal and diachronic data), machine translation, crowdsourced data collection, and the dialectometrical analysis of corpora and inquiries. The project combines these areas of research in original ways.

The PI will **oversee the research activities** carried out within the project, namely: management of recruitment processes, coordination of research (guidance on language choice, corpus collection processes, normalization experiments and dialectology experiments), coordination of international collaboration, organization of research meetings, assistance with publications, data management, and career development of main researcher and research assistant(s).

During the second half of the project, the PI will **participate in the following research packages**: Normalization and dialectological experiments with user-generated content, preparation and organization of a shared task on normalization or identification, dissemination and online visualization of results, assistance with publications and coordination of research.

Main researcher: N.N.

Most of the work will be carried out by the main researcher, who will have previous experience in machine translation and linguistic variation and be able to work independently. The main researcher should ideally be a **post-doctoral researcher** who can be part of the project throughout its duration. If it is not possible to fill the vacancy in this way, alternative arrangements will be considered, such as filling the position with a **PhD candidate**. The main researcher will play a crucial role in defining additional language areas to be investigated.

The main researcher will be responsible for the following duties: coordination of corpus collection and annotation processes, normalization experiments and related publications, dialectological experiments with dialect corpora and related publications, and assistance with experiments on user-generated content.

Research assistant(s): N.N.

Auxiliary tasks especially related to corpus collection, compilation, cleanup and annotation will be handled by a research assistant. Preferably, this could be one or several **Masters' student(s)** in language technology or linguistics. These auxiliary tasks are due in the first two years of the project.

3.2 Collaborators and their key merits in terms of the project:

The PI of the project currently co-supervises two PhD students working on topics related to the proposed project. While their respective research plans will remain independent from the project, they will be integrated in the project team. We will encourage the exchange of data, tools and methodological advances. Joint publications with project team members are also envisaged.

- ***Noëmi Aepli, University of Zurich, Switzerland***
  Noëmi Aepli is a PhD student fully funded by the Swiss National Science Foundation. She has started her PhD project on *Sustainable natural language processing for low-resource language variations* in April 2020. Her research focuses on the improvement of natural language processing

tools for languages with high internal variation, and normalization is a crucial aspect of her thesis proposal. Her funding plan provides for a six-month mobility period in Helsinki in 2022 or 2023.

- ***Janine Siewert, University of Helsinki, Finland***
  Janine Siewert is a PhD student fully funded by the University of Helsinki. She has started her PhD project on *Corpus-based cross-border dialectometry for Low Saxon* in January 2020. While her research focuses on the varieties of Low Saxon, their historical evolution and the border effect of this pluricentric language, the methodology partially overlaps with the project to the extent that normalization and quantitative dialectological analysis play crucial roles in her thesis proposal.

We will continue existing collaborations with the following researchers:

- ***Nikola Ljubešić, Jožef Stefan Institute, Ljubljana, Slovenia***
  Nikola Ljubešić has collected social media data, in particular from Twitter, to investigate linguistic variation. He also leads several corpus creation and annotation efforts for South Slavic languages. He has worked with Yves Scherrer on CSMT-based normalization methods and on BERT-based dialect identification. We will rely on his knowledge for social media data collection and collaborate with him on text normalization.
- ***Tanja Samardžić, University of Zurich, Switzerland***
  Tanja Samardžić is the director of the Language and Space Lab and researches various aspects of the interaction between language and space. She co-created the ArchiMob corpus of Swiss German with Yves Scherrer and leads a working group on Swiss German language resources. The project relies on her competences in corpus compilation and annotation.

In addition, we will initiate collaborations with the following researchers:

- ***Jack Grieve, University of Birmingham, UK***
  Jack Grieve is a professor of corpus linguistics. His research involves the analysis of large corpora, especially from social media, to understand language variation and change. Most of his recent work relies on social media data from various dialectal and sociolectal varieties of English. His experience will help us collect, analyze and interpret social media corpora.
- ***Benedikt Szmrecsanyi, KU Leuven, Belgium***
  Benedikt Szmrecsanyi is professor in linguists with a focus on linguistic variation. He has founded the research area of corpus-based dialectometry and currently leads a research project on syntactic variation in Dutch corpora. His work on the theoretical foundations of corpus-based dialectometry will be of crucial importance to our research project.

Financial provisions are made to invite all collaborators to Helsinki twice, at the beginning of the project and towards the end of the project.

4 Responsible science

4.1 Research ethics:

Most data (dialect corpora and ground truth atlas data) used in the project have been compiled by other research institutions. We trust these institutions that the datasets respect the ethical standards in vigor at the time of their compilation.

If data is collected from social media sites within the context of this project, we will make sure that only public posts are collected and that the terms of service of the platform are respected. This will entail that the data will have to be anonymized for publication, that geolocation information will be blurred, and/or that only incomplete datasets may be published. We will not use or store any speech or audio data in the context of our project.

Within the research project, we will follow the Finnish and European codes of conduct for research integrity. The PI of the project will make sure that all members of the research group adhere to these codes of conduct.

4.2 Equality and non-discrimination:

Any type of research on dialects takes the stance - implicitly or explicitly - that the range of linguistic expressions is vast and diverse and that all types of linguistic variation shall be treated with respect and without discrimination. The scientific investigation of dialects can therefore strengthen the self-esteem of dialect speakers and reduce potential negative stereotypes related to dialectally marked speech practices.

Within the project, we will pay particular attention to the transparency and non-discrimination of recruitment processes. Human resources specialists of the host university will assist us in this task. We will also promote equality in the workplace, as stipulated by the Finnish Equality Act.

4.3 Open science:

We plan to publish the research results both in language technology and dialectology venues. In language technology, conference proceedings are the main means of publication. These have always been freely and openly accessible. Some journals in the areas of language technology and dialectology are Hybrid Access Journals and/or require an APC fee. We make financial provisions for covering the APC of one journal publication, which should be sufficient for all publishing needs related to the project.

The project will use and produce three types of data:

- **Existing datasets:** These datasets will be stored internally on version control systems of the University of Helsinki. Distribution of these datasets is under the responsibility of the original data providers.
- **Collected datasets:** These datasets will be stored internally on version control systems of the University of Helsinki. These corpora may contain personal or otherwise sensitive information, and we will make sure that these are properly addressed before further processing and publication. Data repositories like CLARIN or Kielipankki will be considered for publication.
- **Derived data** (results of experiments): We expect that a wealth of derived data will be produced throughout the project. We will use electronic laboratory notebooks to summarize the experiments and data related to them, and store the derived data on HPC storage archives provided by CSC. Data marked for long-term storage will be supplemented with metadata and made available through IDA or Zenodo. This also applies to the data underlying the visualization web site.

4.4 Sustainable development objectives:

The main goal of the project is to open up existing resources for scientific research. The traditional resources used for dialectological research, atlases and purpose-built dialect corpora, generally take years or decades to produce and demand a lot of active, focused attention from researchers and informants alike. We plan to show that spontaneously occurring data in the form of user-generated content contain precious information about dialectal features that are relevant for research. In this sense, our project contributes to a sustainable handling of linguistic resources and increases the value of citizens' participation in online communities.

5 Societal effects and impact

5.1 Effects and impact beyond academia:

Recent research activities on dialects and dialectology have shown tremendous success beyond academia. A smartphone application able to localize speakers of Swiss German dialects on the basis of their pronunciation was used by more than 70,000 participants (Leemann et al. 2016). Although our project is not explicitly geared towards popular science, we expect a positive societal impact of our research:

- The visualizations of research results will be made freely available online with a user interface that is intuitive for laypeople. They will rely on the technology of our existing web site www.dialektkarten.ch. Interactive maps provide an attractive and playful means to enhance the level of knowledge about linguistic variation in the population.
- Dialect identification tools similar to those of the mentioned smartphone applications could also be developed on the basis of dialect corpora. This is not the main goal of the present research project, but could develop naturally from it as a side project.
- In many language areas, dialects are seen as archaic and stigmatized ways of linguistic expression, and speakers using dialectal features may suffer from various types of discrimination. Even in those language areas that have naturally embraced dialectal variation, some varieties may lack prestige and may be discriminated against. With our project, we show that dialectal variation is precious for science and that dialects can be studied with modern tools like social media data and neural networks.

6 Bibliography

6.1 List of all the sources used in the research plan:

Abe, K., Y. Matsubayashi, N. Okazaki & K. Inui (2018). *Multi-dialect Neural Machine Translation and Dialectometry*. In: *Proceedings of PACLIC*.

Akbik, A., D. Blythe & R. Vollgraf (2018). *Contextual string embeddings for sequence labeling*. In: *Proceedings of COLING*.

Bollmann, M. (2019). *A Large-Scale Comparison of Historical Text Normalization Systems*. In: *Proceedings of NAACL.*

Crystal, D. (2011). *Internet Linguistics: A Student Guide*. Routledge.

Eisenstein, J., B. O'Connor, N. A. Smith & E. P. Xing (2014). *Diffusion of lexical change in social media*. In: *PLoS ONE*.

Embleton, S. & E. S. Wheeler (1997). *Finnish dialect atlas for quantitative studies*. In: *Journal of Quantitative Linguistics* 4:1-3, 99-102.

Goebl, H. (2005). *Dialektometrie*. In: *Quantitative linguistics / Quantitative Linguistik. An international handbook / Ein internationales Handbuch*. Eds. R. Köhler, G. Altmann & R. G. Piotrowski. 498–531.

Goebl, H. (2010). *Dialectometry and quantitative mapping*. In: *Language and Space. An International Handbook of Linguistic Variation. Vol. 2: Language Mapping*. Eds. A. Lameli, R. Kehrein & S. Rabanus. 433–457.

Grieve, J., C. Montgomery, A. Nini, A. Murakami & D. Guo (2019). *Mapping lexical dialect variation in British English using Twitter*. Computational Sociolinguistics Research Topic, Frontiers in Artificial Intelligence (Language and Computation).

Gu, J., C. Wang & J. Zhao (2019). *Levenshtein Transformer*. In: *Advances in Neural Information Processing Systems*, 11181-11191.

Hämäläinen, M., T. Säily, J. Rueter, J. Tiedemann & E. Mäkelä (2018). *Normalizing early English letters to Present-day English spelling*. In: *Proceedings of the LaTeCH Workshop*.

Hovy, D. & C. Purschke (2018). *Capturing Regional Variation with Distributed Place Representations and Geographic Retrofitting*. In: *Proceedings of EMNLP*.

Hotzenköcherle, R., R. Schläpfer, R. Trüb & P. Zinsli (Eds.) (1962–1997). *Sprachatlas der deutschen Schweiz*. Francke. Partial digitization available at www.dialektkarten.ch

Institute for the Languages of Finland (2014). *The Helsinki Korp version of Samples of Spoken Finnish* [text corpus]. Kielipankki. http://urn.fi/urn:nbn:fi:lb-2016042702

Johnson, M., M. Schuster, Q.V. Le, M. Krikun, Y. Wu, Z. Chen, N. Thorat, F. Viégas, M. Wattenberg, G. Corrado & M. Hughes (2017). *Google's multilingual neural machine translation system: Enabling zero-shot translation*. In: *TACL* 5, 339-351.

Kettunen, L. (1940). *Suomen murrekartasto*. Suomalaisen kirjallisuuden seura. Digitization available at http://kettunen.fnhost.org/

Leemann, A., M.-J. Kolly, R. Purves, D. Britain & E. Glaser (2016). *Crowdsourcing language change with smartphone applications*. In: *PLoS ONE*.

Ljubešić, N., D. Fišer, T. Erjavec, J. Čibej, D. Marko, S. Pollak & I. Škrjanec (2015). *Predicting the Level of Text Standardness in User-generated Content*. In: *Proceedings of RANLP*.

Ljubešić, N., T. Samardžić & C. Derungs (2016). *TweetGeo – A Tool for Collecting, Processing and Analysing Geo-encoded Linguistic Data*. In: *Proceedings of COLING*.

Lusetti, M., T. Ruzsics, A. Göhring, T. Samardžić & E. Stark (2018). *Encoder-decoder methods for text mormalization*. In: *Proceedings of VarDial*.

Östling, R. & J. Tiedemann (2017). *Continuous multilinguality with language vectors*. In: *Proceedings of EACL*.

Partanen, N., Hämäläinen, M., & Alnajjar, K. (2019). *Dialect Text Normalization to Normative Standard Finnish*. In: *Proceedings of W-NUT*.

Pettersson, E., B. Megyesi & J. Nivre (2014). *A multilingual evaluation of three spelling normalisation methods for historical text*. In: *Proceedings of the LaTeCH Workshop*.

Scherrer, Y. & T. Erjavec (2016). *Modernising historical Slovene words*. In: *Natural Language Engineering*, 22(6), 2016.

Scherrer, Y. & N. Ljubešić (2016). *Automatic normalisation of the Swiss German ArchiMob corpus using character-level machine translation*. In: *Proceedings of KONVENS*.

Scherrer, Y., & P. Stoeckle (2016). *A quantitative approach to Swiss German – Dialectometric analyses and comparisons of linguistic levels.* In: *Dialectologia et Geolinguistica* 24(1), 92–125.

Scherrer, Y., T. Samardžić & E. Glaser (2019). *Digitising Swiss German - How to process and study a polycentric spoken language*. In: *Language Resources and Evaluation* 53(4).

Scherrer, Y. & N. Ljubešić (submitted). *HeLju@VarDial 2020: Social Media Variety Geolocation with BERT models*. In: *Proceedings of VarDial 2020*.

Sennrich, R., B. Haddow & A. Birch (2016). *Improving Neural Machine Translation Models with Monolingual Data*. In: *Proceedings of ACL*.

Siebenhaar, B. (2006). *Code choice and code-switching in Swiss-German Internet Relay Chat rooms*. In: *Journal of Sociolinguistics* 10(4), 481-506.

Siewert, J., Y. Scherrer, M. Wieling & J. Tiedemann (submitted). *LSDC – A comprehensive dataset for Low Saxon Dialect Classification*. In: *Proceedings of VarDial 2020*.

Szmrecsanyi, B. (2013). *Grammatical variation in British English dialects: A study in corpus-based dialectometry*. Cambridge University Press.

Szmrecsanyi, B. & L. Anderwald (2018). *Corpus-based approaches to dialect study*. In: *The Handbook of Dialectology.* Eds. C. Boberg, J. Nerbonne & D. Watt.

Tiedemann, J. (2009). *Character-based PSMT for closely related languages*. In: *Proceedings of EAMT*.

Wieling, M. & J. Nerbonne (2015). *Advances in dialectometry*. In: *Annual Review of Linguistics* 2015, 243-264.