Péter Jeszenszky, Department of Geography, Ritsumeikan University, Kyoto, *pjeszenszky@gmail.com;* **Panote Siriaraya**, Kyoto Institute of Technology, *spanote@gmail.com*; **Philipp Stoeckle**, Austrian Centre for Digital Humanities, Austrian Academia of Science, *philipp.stoeckle@oeaw.ac.at;* **Adam Jatowt**, Department of Social Inf., Kyoto University, *adam@dl.kuis.kyoto-u.ac.jp*

SPATIO-TEMPORAL PREDICTION OF DIALECTAL VARIANT USAGE

Using logistic predictors to **predict dialect variants** in Swiss German **syntax** based on **age** at **global and local** scales

1st International Workshop on Computational Approaches to Historical Language Change @ **ACL 2019** Annual Meeting of the Association for Computational Linguistics, Aug 2nd, Florence

Summary

The distribution of most dialectal variants have not only spatial but also temporal patterns. Based on the *apparent time* hypothesis, idiolects remain mostly unchanged after the acquisition of the mother tongue. Besides, *contact* between speakers and speaker communities is held responsible for language change.

We model the *contact potential* based on two of the most important factors assumed in sociolinguistics to affect language change: *age and distance*.
We test the following two hypotheses:

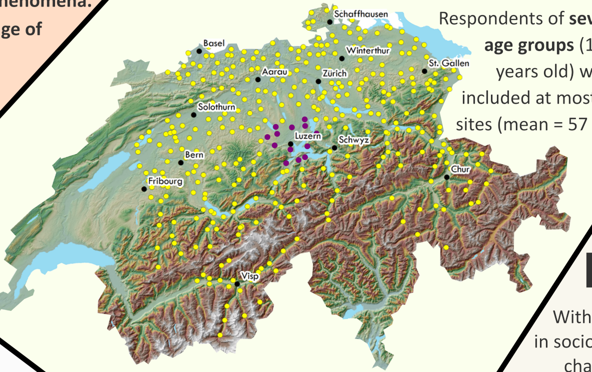
- **At the *global scale*, age explains the usage of dialect variants in some linguistic phenomena.**
- **Age is a better predictor for the usage of dialect variants at the *local scale*.**

Our results show the relative success of a logistic prediction approach implemented and show the potential of the method at the local scales.

Data

The **Syntactic Atlas of German-speaking Switzerland** *(SADS)* probed *(morpho)syntactic* phenomena using 118 written survey questions in 383 relatively homogeneously distributed locations all over Switzerland. A total of 3174 respondents mean **multiple answers for each question** at each survey site (3-26 answers/survey site, *median* = 7).

Respondents of **several age groups** (12-94 years old) were included at most survey sites (mean = 57 years).

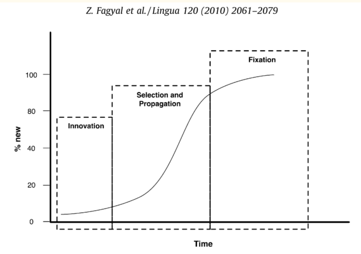


SADS survey sites in Switzerland with an overlay of 13 nearest neighbours of Luzern (Lucerne), as an example of the local scale approach.

Apparent time

The ***apparent time* hypothesis** assumes that mother tongue is mostly acquired until the late teenage, after which the **idiolect** is more resistant to change. If not uprooted, idiolect is assumed to reflect the contact patterns of early life. Thus, **synchronic diversity can be interpreted diachronically**.

Based on the apparent time hypothesis, our database contains information about the whole 20th century and different environments. The rate of change in the linguistic level of syntax is assumed to be low, which is a pro and a con at the same time for the predictive powers of apparent time approach.



With time, innovations fixate in the language. Local patterns in this fixation are shaped by contact.

Methods

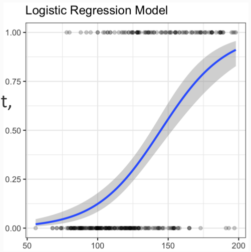
Global scale

We test the association between linguistic variation as a categorical (nominal) variable and age as a continuous predictor variable. We train a **logistic predictor** and apply it with a **10-fold cross-validation** approach, using all 383 survey sites in the training and testing set.

The **precision** of the model is given by the **correct predictions** of the observed data in the testing set. This is reported by the ***AUC***, the area under the ROC curve, which is generated by plotting the **true positive rate** (TPR) against the **false positive rate** (FPR) at various threshold settings.

Local scale

Our model has to decide for each variant whether respondents at a **central survey site *s*** used it or not, based on age as the predictor variable, in a set of ***k* nearest neighbour** (kNN) survey sites.



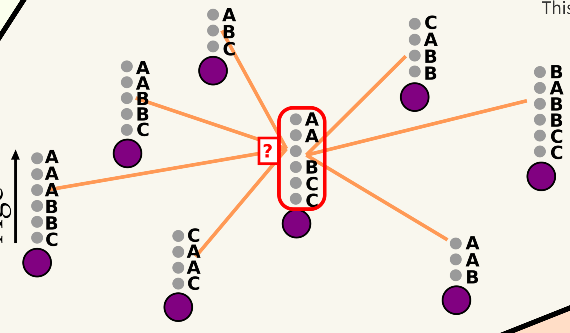
The logistic regression model used here predicts the variant usage for each respondent at each survey site. **kNN** are chosen based on **Euclidean distance**. Such a distance cut-off allows all survey sites to have the same effect on the masked central survey site.

Research aims

With **predicting the usage of variants based on contact**, core issues in sociolinguistic and diachronic linguistics – tracing back and forecasting change in language – can be addressed with a **better granularity**.

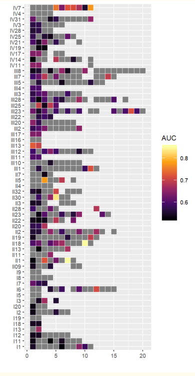
Keeping all other variables constant, the language of two people with similar age, having spent their youth near each other, is expected to be similar. Based on the **social network modelled by distance and age**, and based on sparse dialect data, we aim to predict variant usage.

This is a first step towards assembling a model for **predicting** diachronic usage of dialectal variants, and thereby, **language change** by means of taking many extralinguistic variables, **estimating contact potential**, into account.



We approach predicting dialect usage of survey respondents based on surrounding survey locations and similar age cohorts.

Global and local results

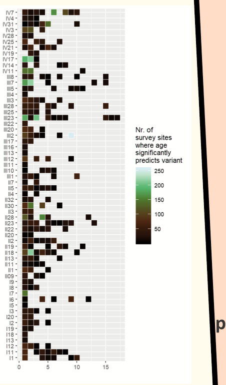


From SADS, 60 survey questions are used. The figure on the left reports whether the **usage prediction** of a certain variant is **significant** or not (dark grey). The ***AUC***, shown by the colour, represents **separability**.

AUC values are relatively low overall, leading us to investigate the prediction power of age at the regional scale, the patterns of which are possibly **concealed by the global patterns**.

In the figure on the right, **kNN** with *k* = 13 was used to present the local predicting power of age. It is visible that age is a significant predictor in many survey sites **only for a few variants**.

Variant coding includes the survey question number and a variant ID. For example, II5_3 is Variant #3 in the 5th question in the second survey sheet.



Interpretation of the results

A logistic prediction based on age has not been done previously on many phenomena at the same time. Age alone does not prove to be an exceptionally good predictor of syntactic variation. This is partly due to the nature of our data. Significance of age as a predictor variable is spatially autocorrelated. When present, the usage of certain variants is characteristic of certain age groups in some spatial clusters, while in other clusters it is used regardless of age.

At the local scale testing different *k* values showed that taking more survey sites into account does not necessarily increase predictive power, especially without distance decay.

Logistic regression is robust as the continuous variables do not have to be normally distributed. Yet, as it is sensitive to class imbalances, it might not always be the best choice as a predictor when there are a lot of 0s and only a few 1s in the data, as it might result in false accuracy by predicting 0s only and not the 1s.

As for outlooks, we will implement a distance decay approach and weights based on different parameters (including age), and add different distance matrices as predictors in a mixed model, estimating potential contact, as we assume closer dialect varieties to be the outcomes of tighter (historical) contact. Besides, we will look beyond logistic models as well.

FNSNF Participation funded by:
FONDS NATIONAL SUISSE
SCHWEIZERISCHER NATIONALFONDS
FONDO NAZIONALE SVIZZERO
SWISS NATIONAL SCIENCE FOUNDATION