Reduction of survey sites in dialectology:

A new methodology based on clustering

Péter Jeszenszky¹, Carina Steiner¹, Adrian Leemann¹

¹Center for the Study of Language and Society, Faculty of Humanities, University of Bern
What is dialectology?

Sicoli & Holton, 2014

Christen, Glaser & Friedli, 2019

Montgomery, 2018
• 2019-2024 – Investigating dialect change in Swiss German in comparison to ~1950

• LARGE-SCALE SURVEY in 125 Survey sites

• 1000 Speakers
  • 8 per location; 4 young, 20-35 (2F/2M); 4 old, 65+ (2F/2M)
  • Lived most of their life at the surveyed location
  • At least one parent from the region
  • Max. commute per day: 2h

• 315 Questions/Items, spontaneous speech, text translation, draw-a-map task etc.

• 300+ items of metadata recorded
Survey site reduction for dialectology

• Dialect change $\rightarrow$ dialects become similar to each other $\rightarrow$ less survey sites are enough to present the variation

• RQ: How to efficiently find a subset of survey sites that represent contemporary dialectal variation?

• Traditional dialectology: slower, purely qualitative, more bias

• Digital data available $\rightarrow$ higher objectivity possible

• We suggest a general methodology
Procedure of finding $n$ optimal survey sites

Requirements of SDATS:

• Reduce 565 $\rightarrow$ 125 survey sites
  Due to dialect change and project budget

• Represent *contemporary* variation
  Address dialect change of 70 years

• Sites regionally representative?
  • Theory of linguistic gravity (Trudgill, 1974)
    • Clustering in linguistic space – assumed to cluster in geographic space as well

STEPs of the General Methodology

- Digitisation of original data (from maps)
- Determine similarity between survey sites $\rightarrow$ linguistic distance
- Cluster analysis of survey sites; Find candidate sites
- Revise candidates qualitatively:
  - dialect expertise
  - sociogeographic changes
Linguistic distance calculation

For each survey site pair:

\[ D_{ij}^{\text{ling}} = \frac{\# \text{differing items}}{\# \text{items answered at sites } i \text{ and } j} \]

Survey sites of the original data (~1950)

Site × Site linguistic distance matrix

| Linguistic distance | AG_01 | ... | LU_05 | ... | ZH_07 |
|---------------------|-------|-----|-------|-----|-------|
| AG_01               | 0     |     | 0.38  | 0.23|       |
| LU_05               | 0.38  | 0   | 0     | 0.51|       |
| ZH_07               | 0.23  | 0.51| 0     | 0   |       |

Based on Goebl, 1983 and Jeszenszky et al., 2019
Clustering in dialectology

- Used for determining dialect areas
- Not used for site reduction

Zheng et al., 2017

Scherrer & Stoeckle, 2016
Partitioning Around Medoids (PAM)

Unweighted Pair Group Method with Arithmetic mean (UPGMA)

Ward’s method

Partitional clustering

Hierarchical clustering

R packages used:
- fpc (Hennig, 2020)
- cluster (Maechler et al., 2019)

Kaufmann & Rousseeuw, 1987
Park & Jun, 2009
In linguistics:
- Cheshire et al., 2011
- Syrjänen et al., 2016

Ward, 1963
Wilks, 1995
In linguistics:
- Heeringa, 2004
- Prokić & Nerbonne, 2008
- Grieve et al., 2011
Cluster validation / Cluster stability

• How stable are clusters?

à How representative?

• Noisy clustering, bootstrapping (Nerbonne et al., 2008)

PROBABILITIES

• Take balanced (lex, morph, phon) subsets

• How often do the same survey sites fall into the same cluster?
Final revision

- **Socio-demographic filtering:**
  Check for important changes at candidate sites
  → change in population →
  → change/mixture of dialects

- **Linguistic filtering:**
  Known, remarkable dialects;
  Interesting/representative local dialects;
  Documented change;
  Equidistant survey sites possible?

  → **Overwrite quantitative decisions**
Key findings

- **Main benefit:** Offer *n* candidate survey sites in a quantitative framework
- Arbitrary number of representative sites can be appointed
- Overlap of the original and intended studies with regards to their objectives and variables
- Locally representative site depends on the purpose of the intended study
- Dialect change to be considered as we want to represent the contemporary dialectal variation → qualitative revision needed

**STEPS OF THE GENERAL METHODOLOGY**

1. **(Digitisation of maps and data)**
2. Determine similarity between survey sites
3. Cluster survey sites; Find candidate sites
4. Revise candidate set: - dialect expertise - sociogeographic changes
Thank you very much!

Our homepage: www.sdats.ch

Funded by: FNS/SNF

Swiss German Dialects Across Time and Space
Der neue schweizerdeutsche Sprachatlas

@sdatsbern

Check out our preprint:
https://bit.ly/3sBiM20

Is it possible to estimate age, weight, height, origin etc. based on voice only? https://bit.ly/3gazzqb

@ Swiss German speakers: Participate in a study of ours!
References (selection based on preprint and presentation)

• Cheshire, J. A., Mateos, P., and Longley, P. A. (2011). Delineating Europe’s Cultural Regions: Population Structure and Surname Clustering. Human Biology 83, 573–598

• Christen, H., Glaser, E., Friedli, M., & Renn, M. (2010). Kleiner Sprachatlas der deutschen Schweiz. Frauenfeld: Verlag Huber.

• Goebel, H. (1982). Dialektometrie: Prinzipien und Methoden des Einsatzes der Numerischen Taxonomie im Bereich der Dialektgeographie (Wien: Verlag der Österreichischen Akademie der Wissenschaften)

• Grieve, J., Speelman, D., and Geeraerts, D. (2011). A statistical method for the identification and aggregation of regional linguistic variation. Language Variation and Change 23, 1–29. https://doi.org/10.1017/S095439451100007X

• Heeringa, W. (2004). Measuring dialect pronunciation differences using Levenshtein distance. Ph.D. thesis, University of Groningen

• Hennig, C. (2020). Package “fpc”: Flexible Procedures for Clustering v. 2.2-8 (pp. 1–164). pp. 1–164. Retrieved from https://cran.r-project.org/web/packages/fpc/fpc.pdf

• Hotzenköcherle, R., Schläpfer, R., Trüb, R., and Zinsli, P. (eds.) (1962-2003). Sprachatlas der deutschen Schweiz (Bern (I–VI)/Basel (VII–VIII)): Francke, 8 vols edn.

• Jeszenszky, P., Hikosaka, Y., Imamura, S., & Yano, K. (2019). Japanese Lexical Variation Explained by Spatial Contact Patterns. ISPRS International Journal of Geo-Information, 8(400), 1–30. https://doi.org/10.3390/ijgi8090400
• Jeszenszky, P., Steiner, C., Leemann, A. (in print). Reduction of survey sites in dialectology: a new methodology based on clustering. *Frontiers of Artificial Intelligence* (tba), doi: 10.13140/RG.2.2.31230.20807

• Kaufman, L. and Rousseeuw, P. J. (1987). Clustering by means of medoids. In *Statistical Data Analysis Based on the L 1–Norm and Related Methods*, ed. Y. Dodge (Amsterdam: North-Holland, Elsevier). 405–416

• Leemann, A., Jeszenszky, P., Steiner, C., Studerus, M., Messerli, J. (2020). Linguistic fieldwork in a pandemic: Supervised data collection combining smartphone recordings and videoconferencing. *Linguistics Vanguard*, 6.3. doi: 10.1515/lingvan-2020-0061

• [Dataset] Leemann, A., Jeszenszky, P., Steiner, C., Studerus, M., Messerli, J. (2020, May 7). SDATS Corpus – Swiss German Dialects Across Time and Space. Retrieved from osf.io/s9z4q

• Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., Hornik, K., Studer, M., ... Murphy, K. (2019). Package “cluster” for R - Finding Groups in Data: Cluster Analysis Extended. *R Package Version 2.1.0*. Retrieved from https://svn.r-project.org/R-packages/trunk/cluster

• Montgomery, C. (2018). The Perceptual Dialectology of England. In N. Braber & S. Jansen (Eds.), *Sociolinguistics in England* (pp. 127–164). https://doi.org/10.1057/978-1-137-56288-3_6

• Nerbonne, J., Kleiweg, P., Heeringa, W., and Manni, F. (2008). Projecting dialect distances to geography: Bootstrap clustering vs. noisy clustering. In *Data Analysis, Machine Learning and Applications*, eds. C. Preisach, H. Burkhart, L. Schmidt-Thieme, and R. Decker (Berlin, Heidelberg: Springer). 647–654. https://doi.org/10.1007/978-3-540-78246-976

• Park, H. S. and Jun, C. H. (2009). A simple and fast algorithm for K-medoids clustering. *Expert Systems with Applications* 36, 3336–3341. https://doi.org/10.1016/j.eswa.2008.01.039

• Prokić, J. and Nerbonne, J. (2008). Recognising groups among dialects. *International Journal of Humanities and Arts Computing* 1, 153–172. https://doi.org/10.3366/e1753854809000366
• [Dataset] Scherrer, Y. (2019). dialektkarten.ch – Interactive dialect maps for German-speaking Switzerland and other European dialect areas

• Scherrer, Y. and Stoeckle, P. (2016). A quantitative approach to Swiss German – Dialectometric analyses and comparisons of linguistic levels. Dialectologia et Geolinguistica 24, 92–125. https://doi.org/10.1515/dialect-2016-0006

• Sicoli, M. A., & Holton, G. (2014). Linguistic phylogenies support back-migration from Beringia to Asia. PLoS ONE, 9(3). https://doi.org/10.1371/journal.pone.0091722

• Syrjänen, K. J. J., Honkola, T., Lehtinen, J., Leino, A., and Vesakoski, O. (2016). Applying Population Genetic Approaches within Languages. Language Dynamics and Change 6, 235–283. https://doi.org/10.1163/22105832-00602002

• Ward, J. H. (1963). Hierarchical Grouping to Optimize an Objective Function. Journal of the American Statistical Association 58, 236–244

• Wilks, D. (1995). Statistical Methods in the Atmospheric Sciences, International Geophysics - Volume 59 (Academic Press), 1st edn.

• Willis, D. (2020). Using social-media data to investigate morphosyntactic variation and dialect syntax in a lesser-used language: Two case studies from Welsh. Glossa: a journal of general linguistics 5, 103. https://doi.org/10.5334/gjgl.1073

• Zheng, Y., Xu, S., Liu, J., Zhao, Y., & Liu, J. (2017). Genetic diversity and population structure of Chinese natural bermudagrass [Cynodon dactylon (L.) Pers.] germplasm based on SRAP markers. PLoS ONE, 12(5), 1–15. https://doi.org/10.1371/journal.pone.0177508