Bridging the gap: Incorporating language variation into documentary and descriptive linguistics

Christopher Cox *(Carleton University)*
Olivia N. Sammons *(University of Alberta / Carleton University)*

LD&C 2019

Acknowledgments

Michif: Eliza Aubichon, Cecile Burroughs, †Daniel Daigneault, †Victoria Daigneault, Verna DeMontigny, †Louis Ducharme, Louise Dufour, Liz Durocher, Tony Durocher, James Favel, Marie Favel, †Rita Flamand, Edna Fleury, George Fleury, Harvey Fleury, Irene Fleury, Lawrance Fleury, Mary Fleury, †Mervin Fleury, Norman Fleury, Angus Gardiner, †Victoria Genaille, Louise Gregory, Thérèse Laliberté, Tony Laliberté, Shirley LaRocque, George Lavallée, †Clifford Ledoux, Grace Ledoux-Zoldy, Yvonne Longworth, Lawrence Morin, Max Morin, R.J. Morin, George Pelletier, Harvey Pelletier, Louis Roy, Thomas ("T.J.") Roy, Stanley Smith, Heather Souter, Edwin St. Pierre, Harriet St. Pierre, Marie Tanner, and Gail Welburn

Endangered Languages Documentation Programme

FONDATION TRUDEAU FOUNDATION

Killam Trusts

Social Sciences and Humanities Research Council of Canada

Conseil de recherches en sciences humaines du Canada

Canada

Acknowledgments

Plautdietsch: Alina Balzer, Hella Banman, Elmer Boehr, Nettie Boehr, Kathy Boldt, †Abe Braun, Bill Braun, Dick Braun, Kathy Braun, †Margaret Braun, Mary Ann Braun, Abe Buhler, Eva Buhler, Jake Buhler, Ruth Buhler, Wilf Buhler, Leonard Doell, †Tina Doell, Jack Driedger, Helen Dueck, †Mary Freistadt, †Menno Friesen, Tena Friesen, Anita Froese, Elsie Froese, Mary Froese, Julie Froess, †Frank Funk, Mary Funk, †Jacob G. Guenter, Cornie Guenther, Sarah Guenther, Henry Harms, †Joan Harms, †Ed Hildebrandt, Mary Hildebrandt, Margaret Janzen, †Beth Kobelsky, Bill Kruger, Helen Kruger, Abe Loewen, †Marie Loewen, Sarah Loewen, Erna Neufeld, Elmer Neufeld, Ann Peters, Bill Peters, Velma Regehr, †Emma Schidlowsky, †Art Zacharias, Edna Zacharias, †Evelyn Zerff

Endangered Languages Documentation Programme

FONDATION TRUDEAU FOUNDATION

Killam Trusts

Social Sciences and Humanities Research Council of Canada

Conseil de recherches en sciences humaines du Canada

Canada

Introduction

- Benefits of attention to language variation in language documentation and conservation (LDC) are manifold:
  - *For language revitalization:*
    - Variation a feature of healthy language use—and one that often requires explicit attention in language planning.[1]

1. Hinton (2001), Sallabank (2012)

“

[I]t is important for everyone in the community to understand that **there is not just one “right” way to speak** and that the variations that occur among the speakers are the vestiges of **healthy language variation.** *(...)*

[D]isagreements over which variety is the “correct” form of a language are worse than useless – they can destroy morale and short-circuit a revitalization program. **Tolerance of variation is essential.**

”

Hinton (2001: 15; emphasis added)

Introduction

- Benefits of attention to language variation in language documentation and conservation (LDC) are manifold:
  - *For language revitalization:*
    - Variation a feature of healthy language use—and one that often requires explicit attention in language planning[1]
  - *For linguistic theory:*
    - Relevance to understanding the development of linguistic diversity[2]

1. Hinton (2001), Sallabank (2012)

2. Dorian (2010), Seifart et al. (2018)

“

[T]here has been a **striking neglect of sociolinguistics within the documentarist program**, despite the gathering of much primary usage data. We actually know little about the sociolinguistics of the small, rural, usually unwritten languages typically targeted by this research (...)—yet these were the societies in which the current linguistic diversity largely evolved. Without this, **we remain ignorant about the key engines of linguistic diversity throughout most of human history.**

”

Seifart et al. (2018: e335; emphasis added)

…but it's not that easy

- **Linguistic variation**, whether in terms of dialects or regional varieties, is often seen as a **barrier** to the goals of LDC:[1]
  - Tendency to focus on single "ancestral code" varieties, with idealized notions of speech communities and their linguistic practices[2]
- This can have implications for:
  - Language instruction
  - Curriculum development
  - Materials development
  - Orthography
  - Community cohesion
  - Allocation of resources

1. Sallabank (2012) 2. Woodbury (2005), Woodbury (2011), Childs, Good & Mitchell (2014)

Recommendations vs. realities

Sociolinguistics[1]

- **Narrow focus:** Single linguistic variables, limited sets of social factors
- Demographically balanced population sampling

Language documentation[2]

- **Broad focus:** Variation as reflected in larger patterns of language use in context
- Opportunistic representation of speakers

1. Meyerhoff (2017) 2. Woodbury (2003)

Q:

Rather than being seen as a barrier, what are **productive ways of addressing linguistic variation** in LDC?

1. Documenting variation: Issues and approaches

Issues in documenting variation

1. *Time constraints*: Multi-variety documentation may not be feasible, given availability of resources, community needs, and levels of endangerment

Issues in documenting variation

1. *Time constraints*: Multi-variety documentation may not be feasible, given availability of resources, community needs, and levels of endangerment
2. *Prior knowledge:* Limited information may be available about variability before documentation and revitalization activities begin[1]
3. *Endangerment:* Community-wide variation, idiolectal variation, and attrition

1. Mansfield & Stanford (2017: 124–127)

Some practical strategies

1. **Be inclusive:** Wherever possible, take advantage of opportunities to reflect as wide a range of uses/varieties of language as possible
2. **Talk to speakers:** Most speakers will certainly have impressions, opinions, and/or anecdotes about how things are said by different people
3. **Consult previous literature** (including on related languages) for reports of variation and **explore** as part of documentary process

Some practical strategies

4. **Take advantage of standardized tasks:** A Pear Film, Frog Story, common translation task or survey that many contributors take part in can help bring variation to the fore *(and make analysis easier later on)*[1]
5. **Make notes** of what comes up unexpectedly and incorporate into your planning

1. Lüpke (2009)

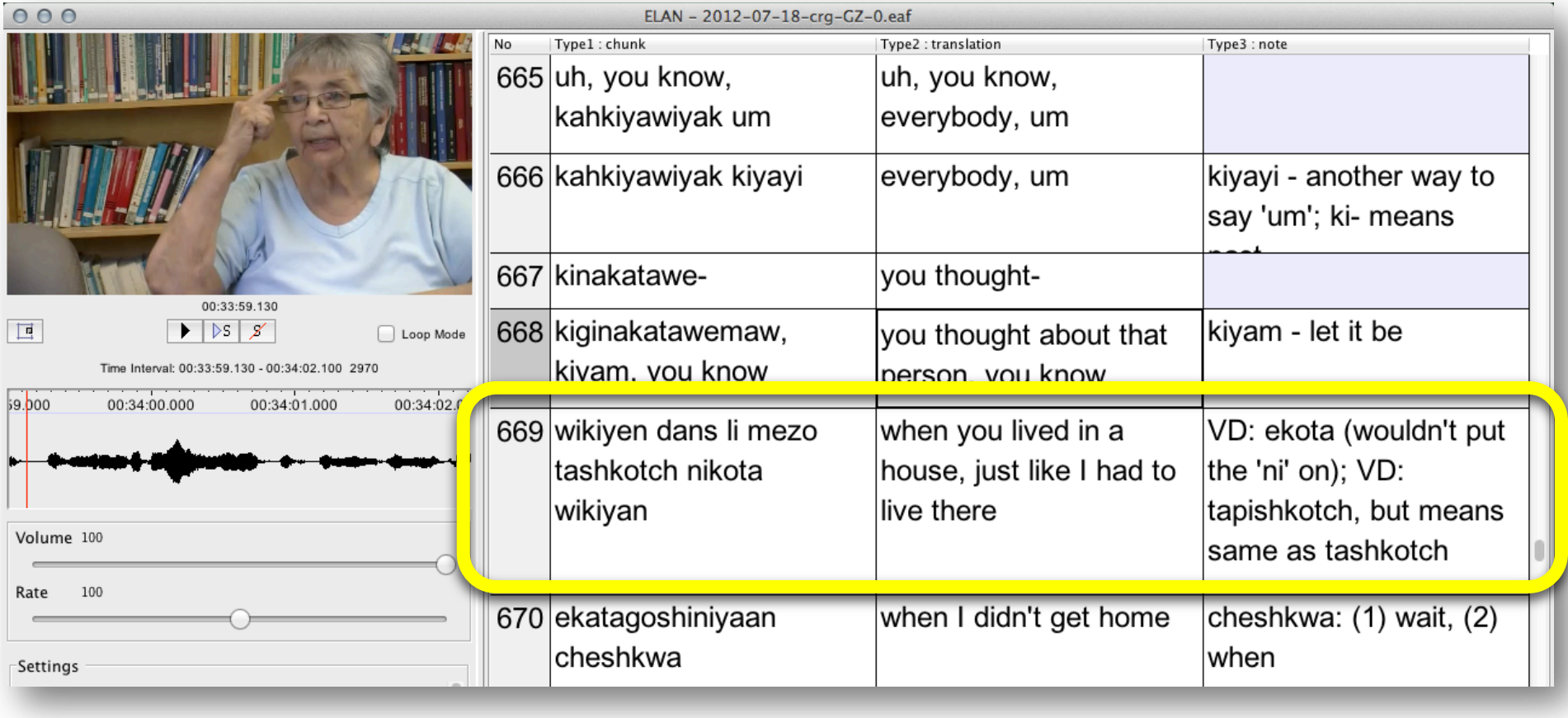
Some practical strategies

6. The process of **annotating** and **editing materials** can help reveal both variation and attitudes towards it[1, 2]
   - **Don't scrub prematurely!**

1. Jung & Himmelmann (2011) 2. Marten & Petzell (2016)

Some practical strategies

7. **Review language materials** featuring speakers of one background with speakers from another
    - Even careful repetitions can involve (unconscious or intentional) levelling of differences that can actually help highlight variation



Indicating dialect differences in ELAN

(Elder Grace Zoldy pictured; commentary by Verna DeMontigny)

not liiv-inaan
‘our books (exclusive)’

- **VD:** Acceptable
- **NF:** Acceptable *(but no inclusive-exclusive distinction)*
- **GZ:** Not acceptable (**moñ liiv-inaan**)

“He’d say **not liivr-inaani** (‘our books’).” 

“If he didn’t use that ‘*not*’ there, it would be good.” 

2. Analyzing variation in language documentation and conservation

Approaching variation

- It would often be helpful to have a clearer understanding of the **relationships** that exist in the variation that we encounter in LDC:
  - *Between individuals:* Who speaks what way? Who speaks most or least like one another? What linguistic or social features characterize a particular group of speakers?
  - *Between instances of variation:* Who tends to use *(or not use)* a particular form? What other variables seem to pattern the same way?

Approaching variation

- **Problem:** Quantitative methods for analyzing sociolinguistic variation may falter on relatively small, sparse, or socio-demographically 'uneven' samples of language use[1]

1. Blainey (2017: 588–590), Meyerhoff (2017: 545), *inter alia*;

Approaching variation

- **Problem:** Quantitative methods for analyzing sociolinguistic variation may falter on relatively small, sparse, or socio-demographically 'uneven' samples of language use[1]
  - Yet, documentary corpora tend to be exactly this: **smaller**, more **opportunistically assembled**, and not always planned with these kinds of questions in mind[2]

1. Blainey (2017: 588–590), Meyerhoff (2017: 545), *inter alia*

2. Woodbury (2003)

“

Contexts of **language endangerment** (…) may not provide the **quantity** or **type** of sociolinguistic data necessary for logistic regression analysis.

”

Blainey (2007: 588; emphasis added)

“

While **mixed-effects models** may be the ideal analytical tool for variationist sociolinguistic analysis, research involving endangered [languages or varieties] **may not be in a position to use them.**

”

Blainey (2007: 588; emphasis added)

Example: Plautdietsch in Saskatchewan

- **Plautdietsch** (Indo-European; ISO 639-3: `pdt`):
  - Traditional language of diasporic **Dutch-Russian Mennonite** communities
  - Complex history of settlement and migration in western Canada in late 19th–early 20th century
  - Rapid decline in language use following forced closure of Mennonite school system (1916–1919), mass emigration to Latin America (*ca.* 1922–1929)

Thiessen (2003)

Example: Plautdietsch in Saskatchewan

- Growing interest in **language education and revitalization** in many communities—but
  - … whose language should be taught?
  - … how many **different varieties** are actually present in the community? How are they **distributed**, socially and geographically? What features **characterize** them?
  - … how can these differences be respected when the language is taught?

"*Eene Plautdietsche Fibel*"

- **Idea:** Survey variation while developing basic documentation and educational resources *(here, traditional Mennonite primer)*
  - **108 linguistic features** known to show variability reflected in 42 contributed responses

Multiple Correspondence Analysis *(MCA)*

- Statistical method that aims to identify regular patterns of variation in *(mostly)* categorical data[1]
  - Unlike regression-based methods, does well with **large numbers of variables** over relatively **few respondents**
  - Identifies both similarities **between observations** *(speakers) and* **variables** *(linguistic or social features)*
  - Allows for **visualization** of patterns of variation that might otherwise be difficult to interpret

1. Abdi & Valantin (2007), Lê, Josse & Husson (2008)

Applying MCA in documentation

Q: How do we use MCA here?

1. Gather together instances of variation from our documentation into a **spreadsheet** *(speakers as rows, variables as columns)*

	Speaker	vAuEeGave	vCanPLVowel	vEnInf	vGirls	vRealizationU	Gender	ParentsPOB
1	F00	AU	Ä	-e	Mäakjes	BACK	F	UkraineUkraine
2	F01	AU	Ä	-e	Mäakjes	FRONT	F	UkraineUkraine
3	F02	EE	E	-e / -en	Me(r)jallen	FRONT	F	CanadaCanada
4	F03	EE	E	-e	Me(r)jalles / Mäakjes	BACK / FRONT	F	USAUSA
5	F05	AU	E	-e / -en	Me(r)jalles	FRONT	F	CanadaCanada
6	F06	AU	E	-e	Mäakjes	BACK	F	UkraineUkraine
7	F07	AU	Ä	-e	Mäakjes	BACK	F	UkraineUkraine
8	F08	EE	Ä	-e / -en	Me(r)jalles	FRONT	F	UkraineUkraine
9	F09	AU	E	-en	Mäakjes	FRONT	F	CanadaUkraine
10	F11	EE	E	-en	Me(r)jalles	FRONT	F	CanadaCanada

Applying MCA in documentation

Q: How do we use MCA here?

2. Import this spreadsheet into the statistical software package **R**[1], then use the `mca` function provided by the **FactoMineR**[2] library to perform Multiple Correspondence Analysis

http://factominer.free.fr

1. R Core Team (2018)

2. Le, Josse & Husson (2008)

Applying MCA in documentation

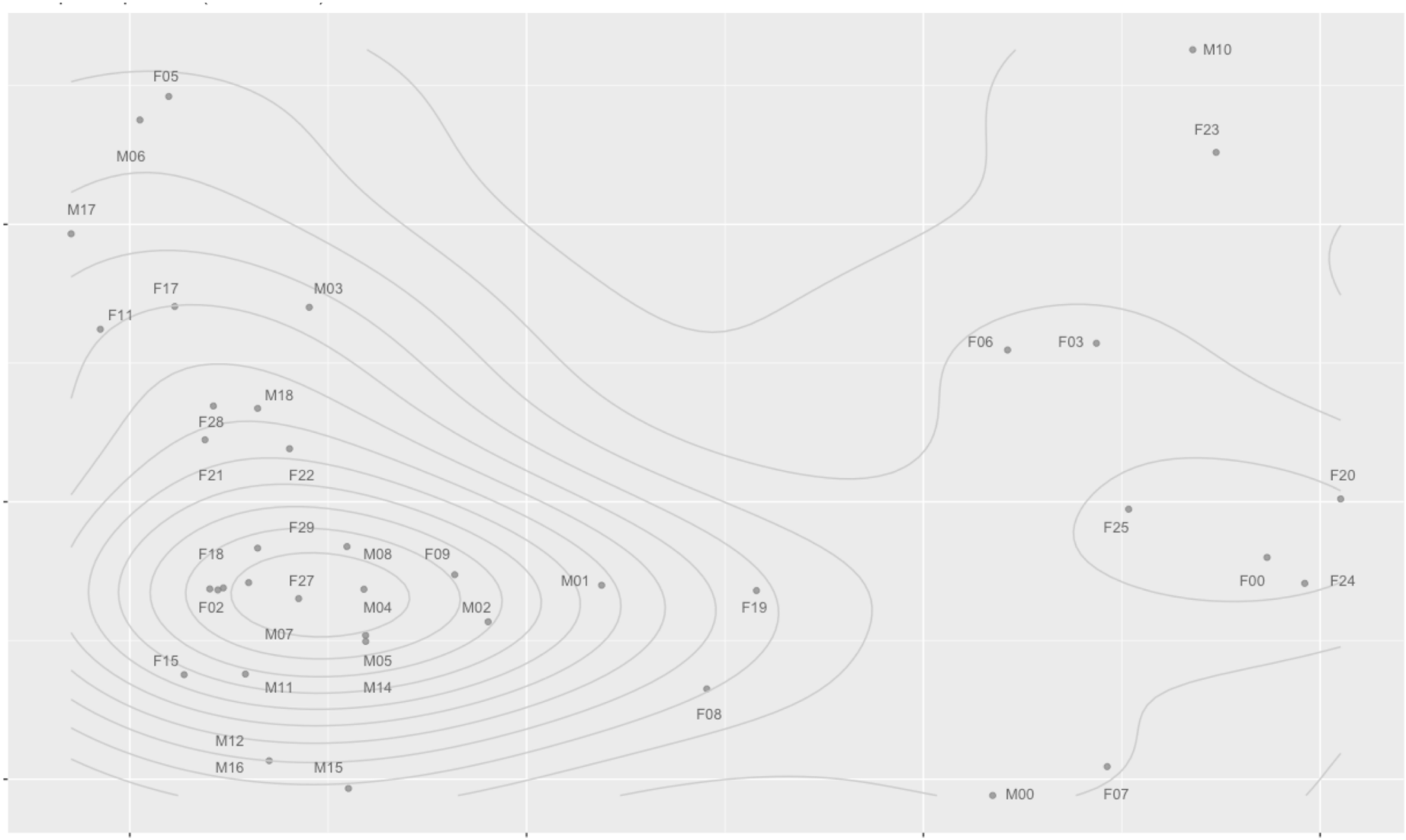
Q: How do we use MCA here?

3. Visualize the results, using FactoMineR's built-in methods or the `ggplot2`[1] library

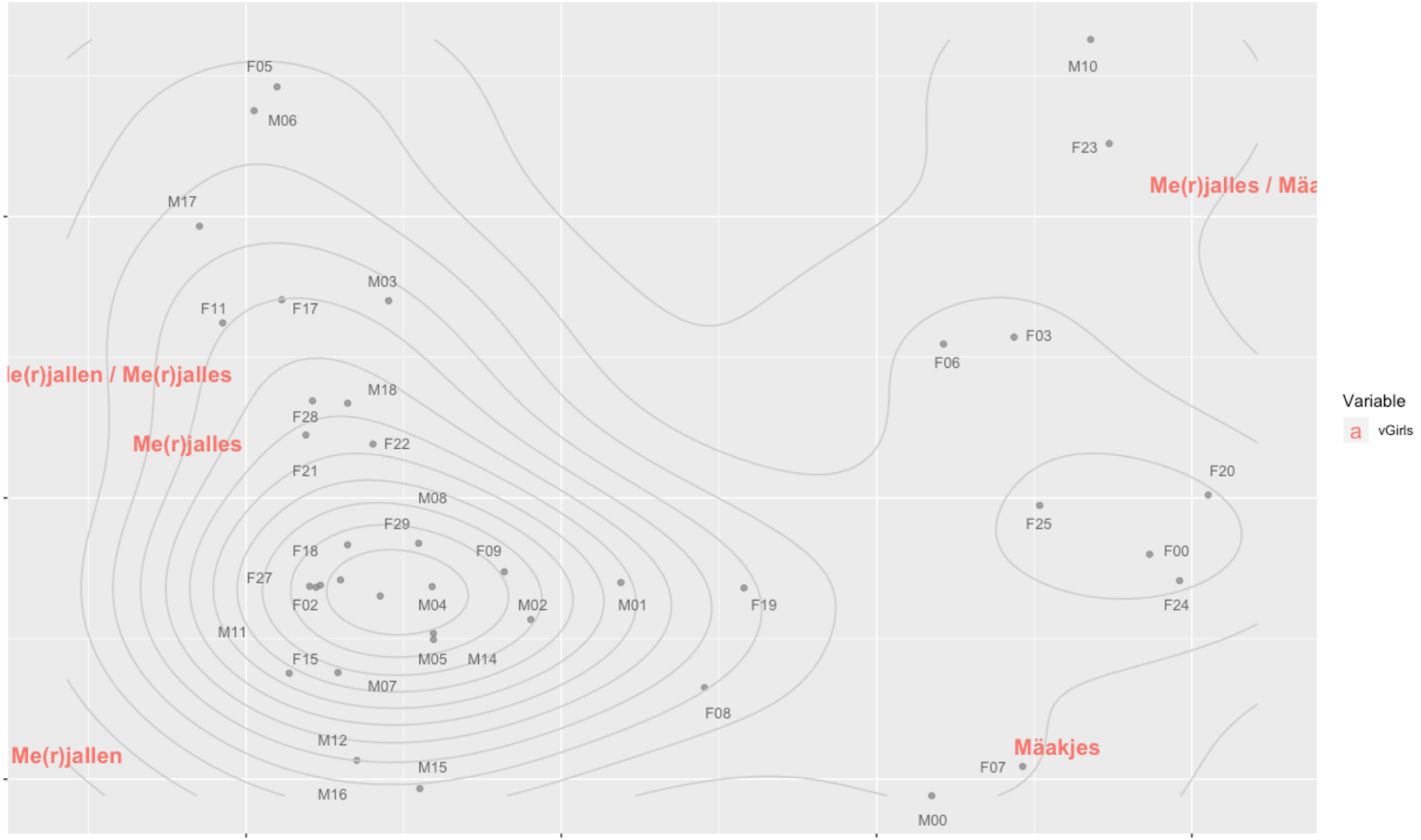
https://ggplot2.tidyverse.org/

1. Wickham (2016)

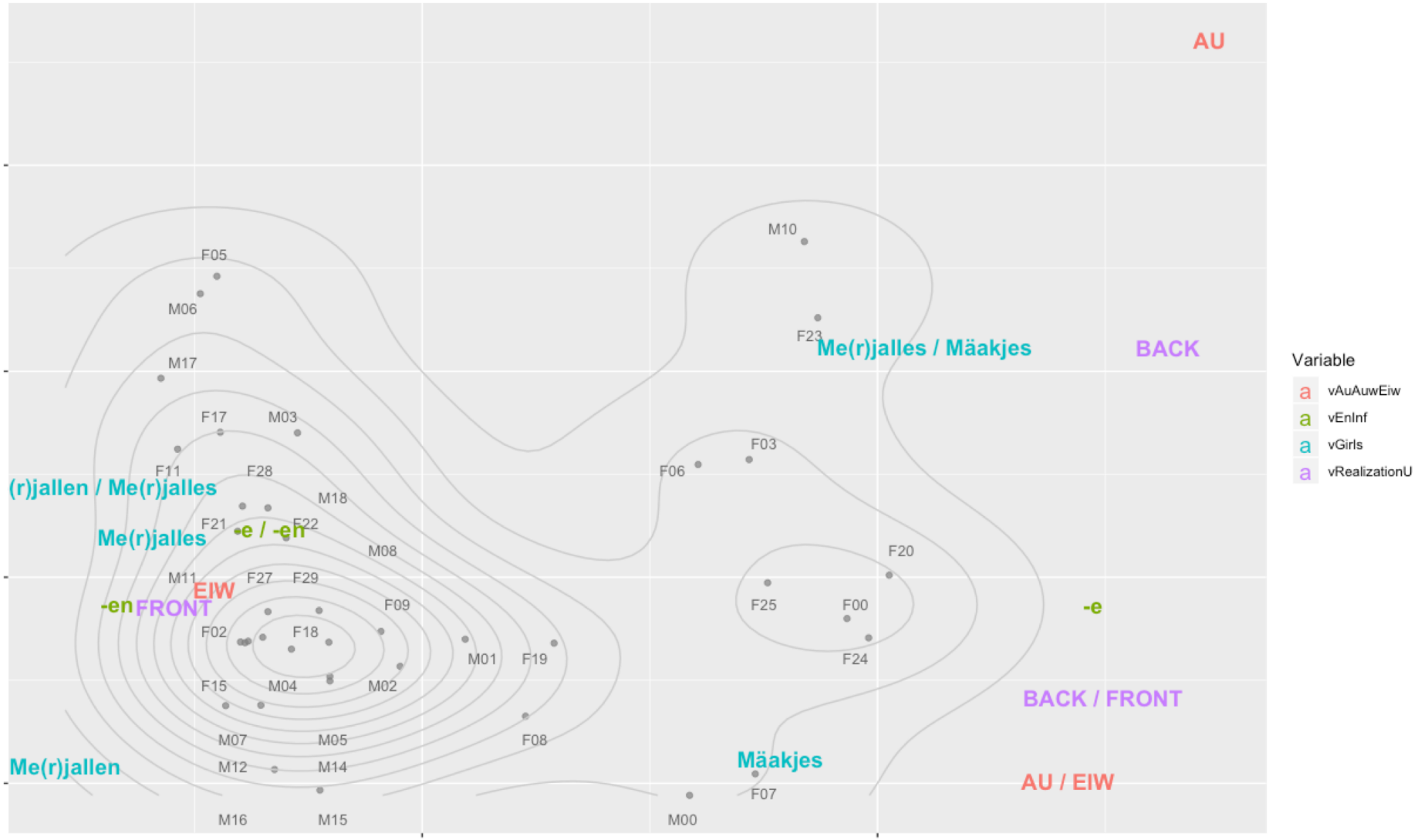
MCA: Grouping speakers together



MCA: Speakers and linguistic features



MCA: Speakers and linguistic features



Beyond MCA in documentation

- Can apply **Hierarchical Clustering on Principle Components** (HCPC) to learn more about the linguistic and social features of each group of speakers:
  - What linguistic and social features are **most distinctive** for that group? Which **specific variants** are associated with this group?
  - Who is the **most 'prototypical'** speaker in this group? Who in this group speaks the ***least*** **like other groups**?

3. Conclusions

Conclusions

1. As challenging as variation may sometimes seem to be to target in documentation, there are **important benefits** to its inclusion in the record, both for revitalization and for linguistic theory.
2. There are ways of incorporating variation into documentation that **don't have to be overly onerous**—but this requires awareness of variation and potentially some planning.

Conclusions

3. **Methodological challenges** in analyzing variation in documentary corpora **can be addressed**, albeit possibly in different ways from other sociolinguistic analyses.

Thanks!