Letteren, exact! / Humanities, exactly!

John Nerbonne

Rijksuniversiteit Groningen

Afscheidscollege

Jan. 27, 2017

Goals

- Background (early work, interests)
- Sketch of most important research line
- How some of it felt
- Some thanks
- *Valete*, Groningen!

Background: Computational Linguistics (CL)

- CL *now* well known – lots of smart phone apps
  - Search, spell-check, translate, speech, intelligent dictionaries, …
  - Popular!
- CL is theory & engineering behind apps
- "If I had asked people what they wanted, they'd have said faster horses." Henry Ford
- Own career shifted from application to theory

Research topics

- Varied, including language interfaces to software, computer-assisted language learning, evaluation
- Collaboration on transliteration for search, handwriting recognition, geo-referencing texts, text enrichment (for education)
- Several pure theory lines on syntax and semantics, hierarchical lexica, learning from simple data, detecting contact influences
- Over thirty languages

Dialectology

- It is one of the first duties of a professor […] to exaggerate a little both the importance of his subject and his own importance in it
  - G.H. Hardy, *A mathematician's apology*
- My best known and best developed research
- Started w. student project, replicating recent (1 yr. old) paper!
- Dialectology has/had a dusty image (Voskuil)
- But more abstract questions abound
  - How does geographic influence arise? What form does it take? Role in language change?

String comparison (edit distance)

- Levenshtein distance (LD, aka edit distance) aligns strings optimally, measures distance
- Dutch 'milk' in Grouw, Haarlem

m	ɔ	l		k	ə
m	ɛ	l	ə	k	
	1		1		1

∑ (distance) = 3

- Idea: Apply LD to phonetic transcriptions in dialect atlases

Traditional dialectology Problem 1

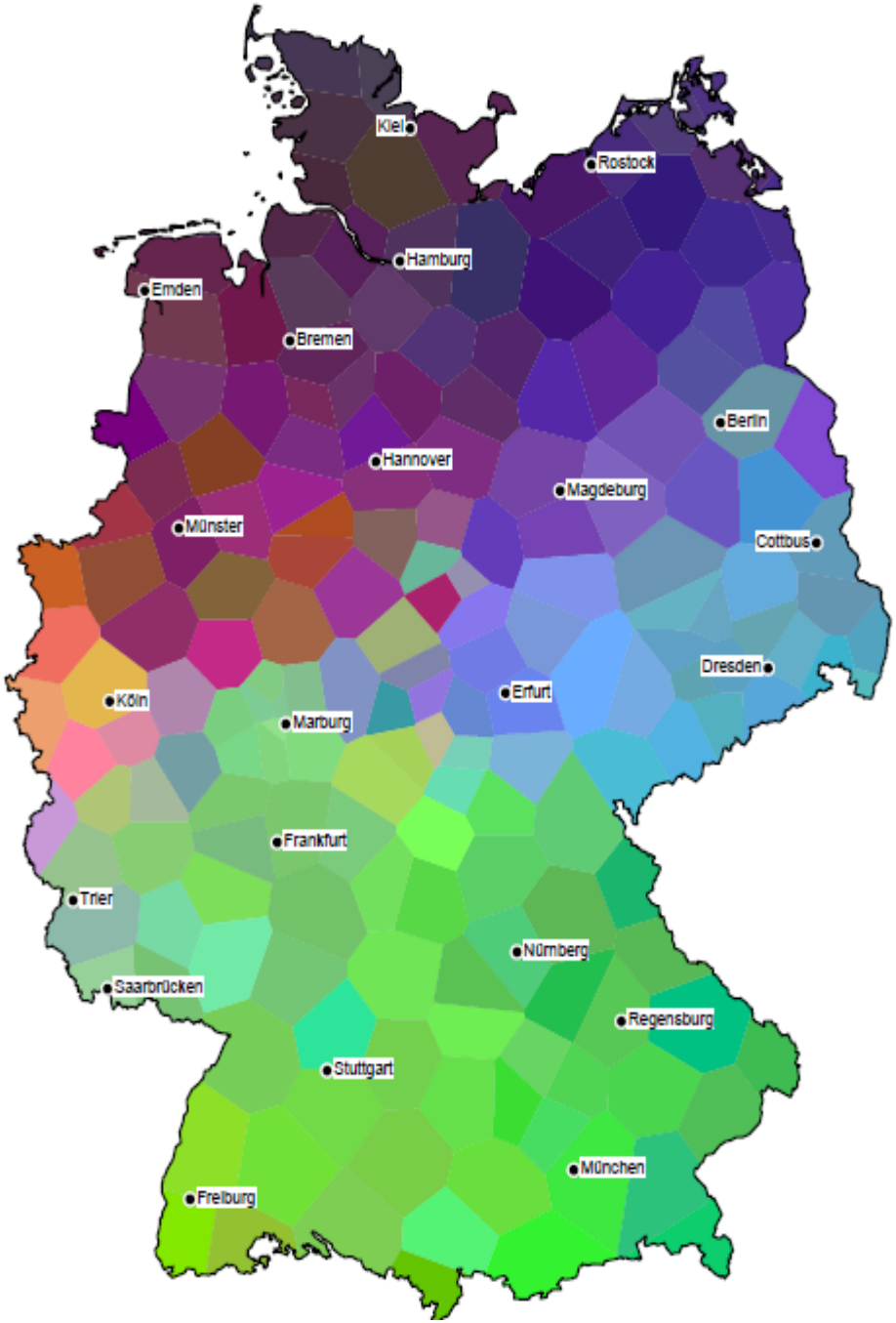
- ✓ Categorical level – same or different
- ✓ But some pairs are more similar than others!
- ✓ No access to more powerful numerical analyses

Pronunciations of *ich* 'I' in German atlas

Traditional Dialectology Problem 2

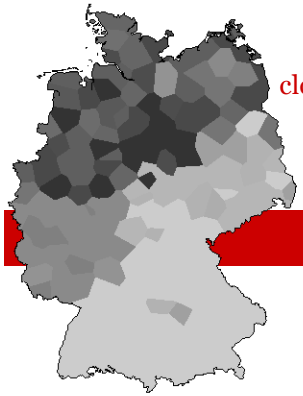
- ✓ No simple overlap in maps of individual features
- ✓ Noisy distribution
- ✓ Bloomfield (1933), summarizing Kloeke

FIGURE 6. Distribution of syllabic sounds in the words *mouse* and *house* in the Netherlands. — After Kloeke.

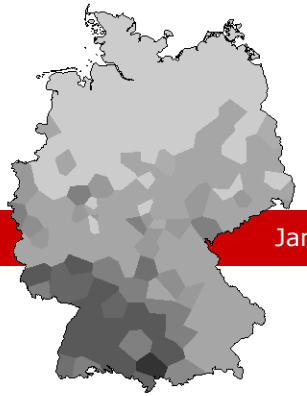


clcg

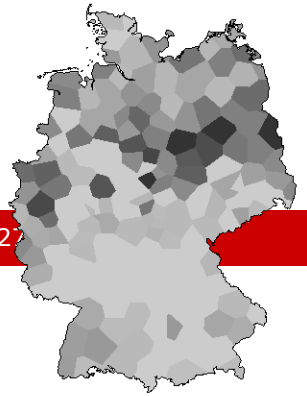
Jan. 2



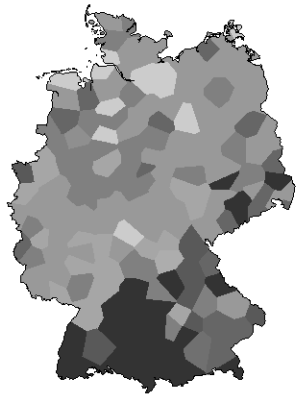
aggregate 2nd shift



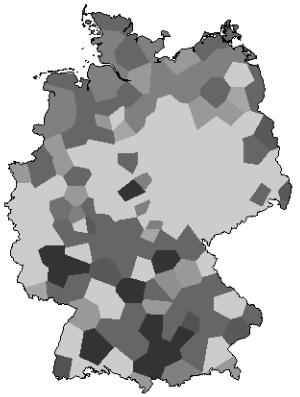
[ʃ] (dark) vs. s
(non-initially)



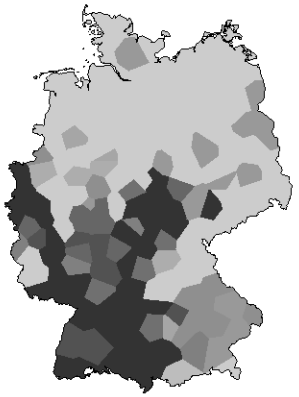
[z] (dark) vs. [s]
(initially)



post-nasal d/t (dark)
vs. deletion

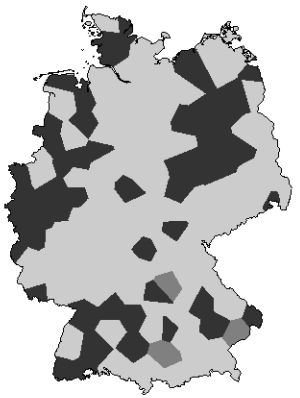


apical [r] (dark)
vs. uvular [ʁ]

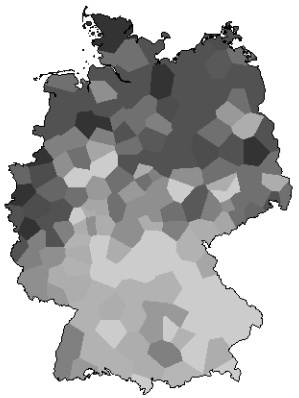


final [n] deletion (dark)
vs. retention

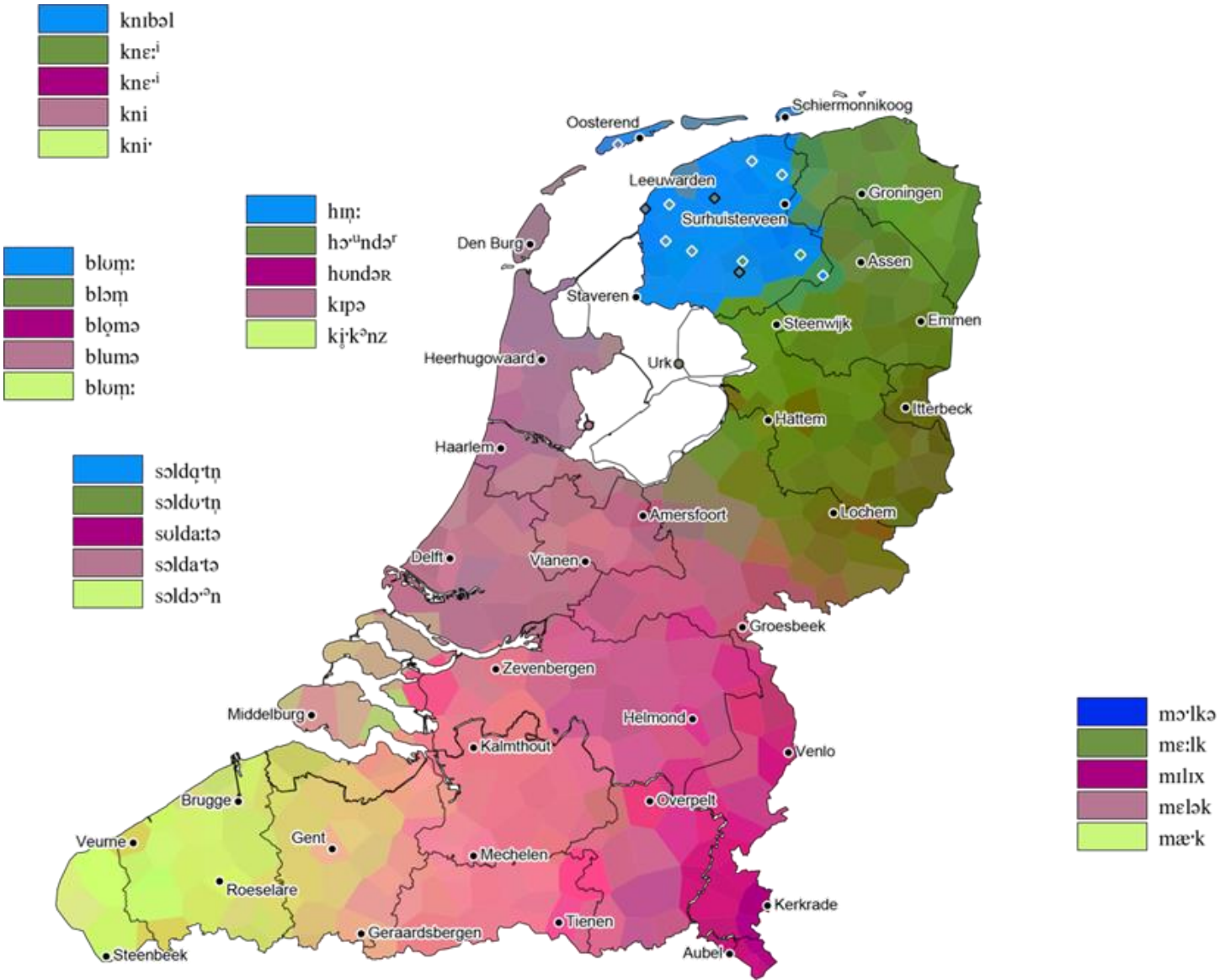
medial [t] vs. s



initial lenited /g/



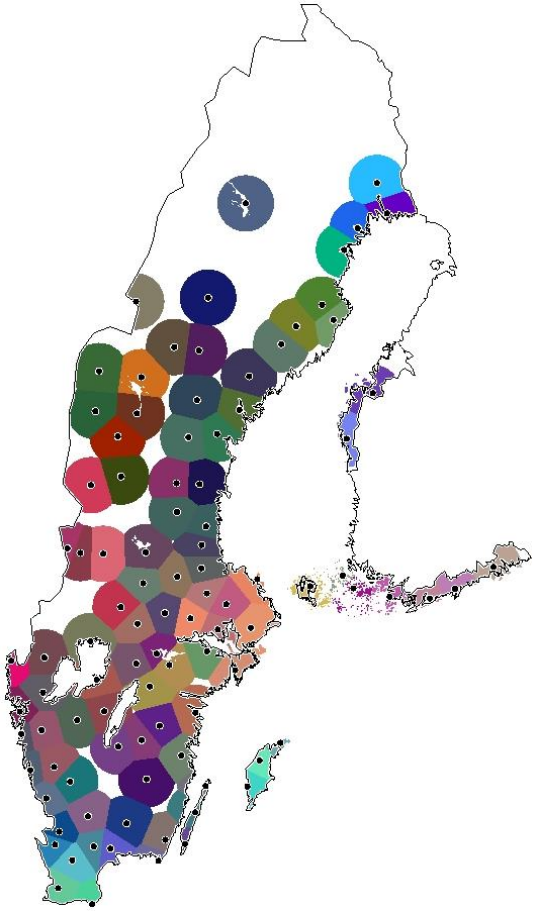
front or low V in *Hause*


knɪbəl
knɛːⁱ
knɛˑⁱ
kni
kniˑ
hɪn̩:
hɔˑᵘndəʳ
hʊndəʀ
kɪpə
kɨˑkᵊnz
bloːm̩:
bloəm̩
bloˌmə
blumə
bloːm̩:
səldɑˌtn̩
səldʊˑtn̩
sʊldaːtə
səldaˑtə
səldɔˑᵊn
mɔˑlkə
mɛːlk
mɪlɪx
mɛləx
mæˑk
Schiermonnikoog
Oosterend
Leeuwarden
Groningen
Surhuisterveen
Den Burg
Assen
Staveren
Steenwijk
Emmen
Heerhugowaard
Urk
Hattem
Itterbeck
Haarlem
Amersfoort
Lochem
Delft
Vianen
Groesbeek
Zevenbergen
Middelburg
Helmond
Kalmthout
Venlo
Brugge
Overpelt
Veurne
Gent
Mechelen
Roeselare
Kerkrade
Tienen
Geraardsbergen
Steenbeek
Aubel

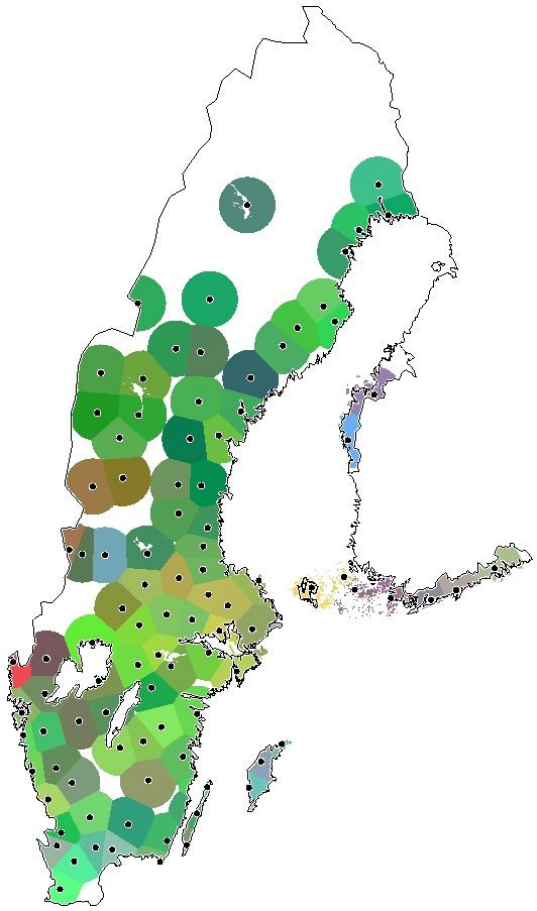
Lots of deeper, further work

- Heeringa (2004): Variations on edit distance, validation studies (also w. Gooskens)
  - Relation geo. and ling. differences
- Spruit (2008): Syntax, search for latent factors
- Shackleton (2010): Eng. sources of Am. dialects
- Prokić (2010): Bulgarian, phylogenetic inference
- Nabende (2011): Transliteration (Urdu, Russian)
- Wieling (2012): Non-linear regression, enabling comprehensive statistical model
- Hansen (2016) Spontaneous vs. elicited
- Manni (ongoing): Links to genetics, culture

Swedish Dialect Leveling



- Data (Eriksson, 2004)
  - > 1K speakers
  - 19 vowels, 5 recordings each
- 65-yr. olds (left), 27-yr. olds (right)
- Therese Leinonen, 2010 Royal Gustav Adolph Prize, Swedish Folk Culture
- N.B. “Leveling” good aggregate concept



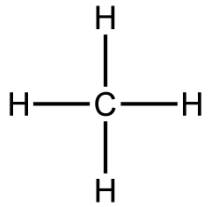
Lots of great collaborators

I not only use all the brains that I have, but also all that I can borrow (Woodrow Wilson).

- **Groningen**: Renée van Bezooijen, Leonie Bosveld, Çağrı Çöltekin, Bob de Jonge, Peter Houtzagers, Remco Knooihuizen, Sebastian Kürschner, Hermann Niebaum, and Ernst Wit
- **Elsewhere**: Harald Baayen, Erhard Hinrichs, Franz Manni, Philippe Mennecier, Bill Kretzschmar, Timo Lauttamus, Lisa Lena Opas-Hänninen, Simonetta Montemagni, Petya Osenova, Vladimir Zhobov, Lucija Simičić, and Esteve Valls.
- ***Prima inter pares***: Charlotte Gooskens – comprehensibility, w. Vincent van Heuven, Anja Schüppert, Femke Swarte, and Jelena Golubovic

Discrete micro-level, statistical macro-level

- Syllable structure
  - V, CV, CVn Japanese
  - V, VC Arandic (Aus.)
  - V, CV, VC, CCV, ... Dutch
- Dialects (10^4-10^5 wd)

  aggregate similarity
- Chemical Valence
  - Hydrogen H – H
  - Methane 
  - Water H – O – H
- Volumes of gas:

  statistical mechanics

Lots of open questions

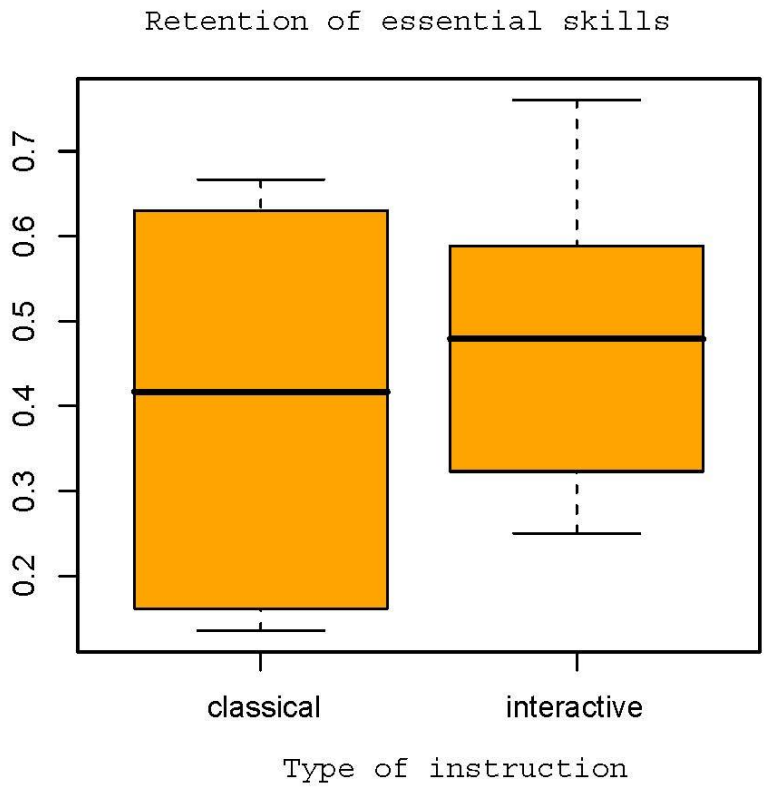
- How does linguistic structure influence aggregate differences, and how much?
- Can we develop better measures of syntactic differences?
- In morphology, should we measure allomorphy and morphotactics independently? How can we measure allomorphic variation independently of phonetic and phonological variation?
- Can we automate the detection of these differences well enough to enable corpus-based measurements?
- Can we bring this social perspective on language into closer contact with the dominant cognitive perspective of linguistics?

Teaching

- Logic → Language → Computation → Statistics
- Lots of statistics teaching in the last 15 years
- Rewarding, given how frequently simple statistical reasoning is invoked
  - Part of educating to autonomy, articulateness (Enlightment vision, Kant, von Humboldt)

Statistics & enlightenment goals

- (Discussion among parents):
  - A: Interactive methods are proven superior!
  - B: But I think kids can be very different!
- What's the best next step (in discussion)?



Pre-statistical heroism!

- Intellectual life emphasized discrete categories
  - Linguistics: Generative grammar (syntax), finite-state automata (phonology, morphology)
  - Logic: Modal logics, Intensional logics, Montague grammar
  - Computer Science: Worst-case complexity, comparison to exponential combinatorics
- Never tell me the odds! (Han Solo, *Return of the Jedi*) https://www.youtube.com/watch?v=gRvu0yHoHy8

Management

- "You may not be interested in war, but war may be interested in you." (Trotsky)
- Started for all the wrong reasons!
- Also rewarding, e.g., demanding review of graduate student projects after one year.
- Fantastic support from Wyke van der Meer!

Special thanks

- NUFFIC (Uganda project), also Gerard Renardel, Henk Sol & Erik Haarbrink
- RuG, CvB, FdL – Deans de Haan & Wakker
- CL community in NL/BE – engaged!
- Department
  - Gertjan & Gosse (Jake & Elroy), Johan, George, Malvina, Leonie, Greg, Barbara,...
  - Carel & CIW group

- Ellen on the home front

Thanks for your attention!

› *Valete*, Groningen!