Clustering Dialect Data: Experiments with the Dutch SAND Database

Erik Tjong Kim Sang
Meertens Institute, Amsterdam
Netherlands eScience Center, Amsterdam
erik.tjong.kim.sang@meertens.knaw.nl

24 August 2017

SAND: Syntactic Atlas of Dutch Dialects

The SAND data set contains syntactic data of different dialects spoken in The Netherlands and the Dutch-speaking part of Belgium

The data set contains 219 syntactic features based on 671 sentences presented in 267 locations

Collected and analyzed data are available on maps in printed volumes, on a website and in a relational database

Website: www.meertens.knaw.nl/sand

28 **Het verbale cluster: doorbreking en morfosyntaxis** Verbal cluster: interruption and morphosyntax

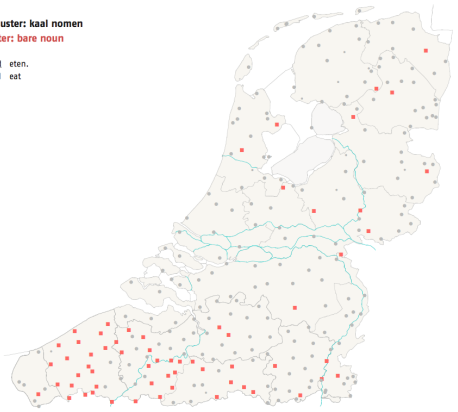
a

2.3.1.1 **Doorbreking van het verbale cluster: kaal nomen**
Interruption of the verbal cluster: bare noun

Ik weet dat Eddy morgen wil brood eten.
I know that Eddy tomorrow wants bread eat

'I know that Eddy wants to eat bread tomorrow.'

■ interruption by bare noun 62
● *no interruption by bare noun* 195



Research goal

Our goal is to use computational techniques for automatically deriving sensible dialect regions from the data

Our research question is: *Are the syntactic features of dialects sufficient for identifying dialect boundaries?*

We are also interested in the kind of dialect boundaries that can be derived: sharp boundaries or transition zones

Challenge

Most relevant automatic data clustering techniques require **complete** and **contrastive** data for creating models

The SAND data are neither complete nor contrastive

For example on contrastivity: a syntactic variable may have value A in one location, value B in another location and both values in yet another location

14 **Werkwoordclusters** Verb clusters

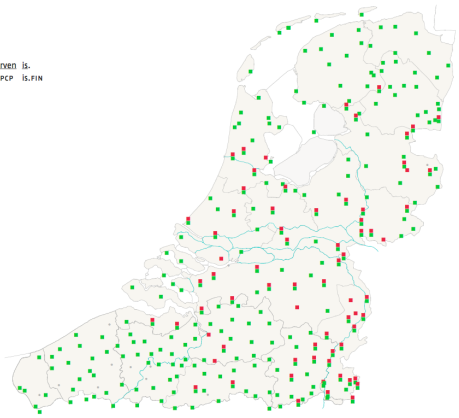
a

1.3.1.1 **Hww *zijn*$_{\text{FINIET}}$ · V$_{\text{PARTICIPIUM}}$**
Auxiliary be$_{\text{FINITE}}$ · V$_{\text{PARTICIPLE}}$

Ze weet niet dat Marie gisteren gestorven is.
she knows not that Marie yesterday died.PCP is.FIN

'She does not know that Marie died yesterday.'

■ V1–V2 (is gestorven) 71
■ V2–V1 (gestorven is) 247



Previous work

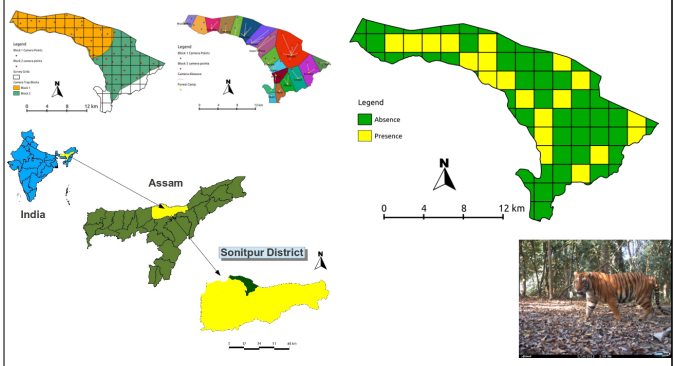
Spruit (2008) converted all feature values to binary classes, for example to AisPresent and AisNotPresent. Next he replaced all missing values by negative values (XisNotPresent)

Van Craenenbroeck (2014) used the same approach as Spruit, except for the location where the features were not tested: there he estimated missing values based on the distribution of the known values

Tjong Kim Sang (2016) used the unchanged SAND data and developed a distance function for comparing with sets of arbitrary strings and dealing with missing values

Data analysis for SAND2

Count	Type	List
2	With negative data	28a 28b
8	Reannotated	32b 41b 42a 42b 49a 50b 55a 57a
27	Sufficient coverage	14a 14b 15a 15b 17a 17b 18a 20a 31a 31b 33a 33b 35a 36b 37a 40a 41a 56a 57b 58b 60a 61a 61b 62a 63b 64a 64b
14	Information in comments	18b 29a 29b 29c 30a 34a 34b 37b 38b 40b 45b 49b 52a 63a
9	Application rule	38a 44b 44c 45a 46a 46b 46c 48a 58a
5	No information	19b 20b 35b 43a 59b
19	Ignore these maps	19a 21a 22a 23a 24a 25a 30b 32a 36a 43b 44a 50a 52b 53a 54a 54b 55b 56b 59a



Alternative data interpretation

We do not interpret the data as the local dialect at this location **has** a certain feature or the local dialect at this location **does not have** a certain feature

Instead we state: this certain feature **was observed** in the local dialect at this location or this certain feature **was not observed** in the local dialect at this location

This subtle interpretation difference solves our problem of modeling with few negative data

Modeling experiments

Repeating the work of the third chapter of Spruit (2008):

1. Extracted all 973 feature values from the SAND database[1]
2. Removed the 35 negative feature values
3. Converted 938 values to complete binary feature table
4. Computed Hamming distances between locations
5. Summarized dialect differences with multidimensional scaling (R)
6. Visualized the three most important axes with colors on a map

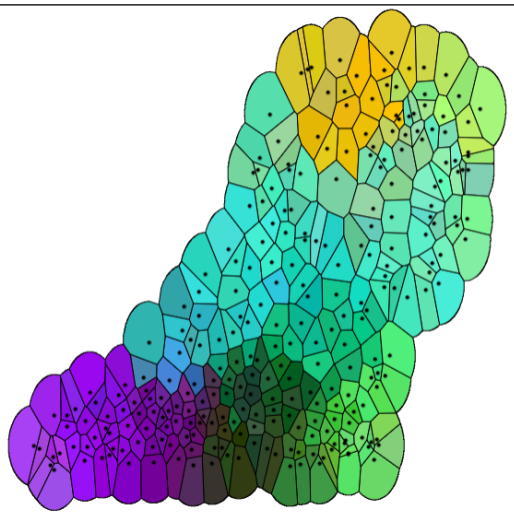
[1] Spruit (2008) only used SAND1: 507 values

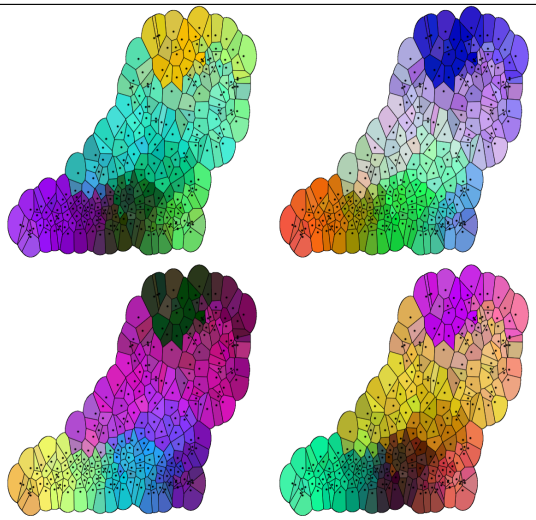
Modeling example

1. Each location is represented by a vector of features: (1,1,0,0,1)
2. The Hamming distance between two location vectors is equal to the number of different features in two vectors:
   Hamming distance between (1,1,0,0,1) and (0,1,0,1,0) is 3
3. The distances between a location and all other locations can be put in a distance vector, for example: (0, 3, 2, 5, 1)
4. Multidimensional scaling summarizes the distance vectors to vectors of three numbers, for example (0.9, 0.1, 0.0)

VISUALIZATION TOOL

ifarm.nl/maps/arvid





Issues with this approach

1. Color selection influences the observations
2. Exact boundary location hard to determine
3. Not all data are used but a summary

Can we get hard boundaries based on all data?

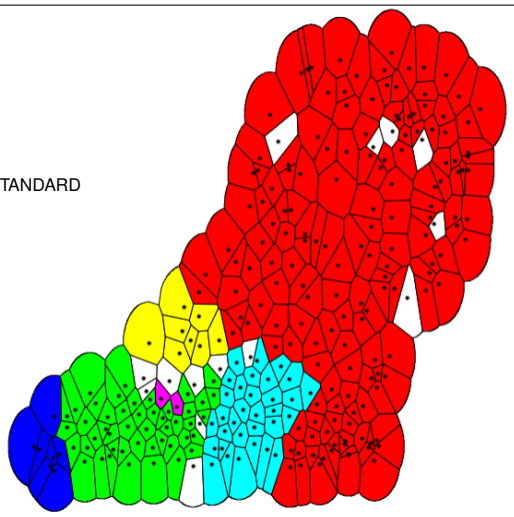
Hard clustering

We use single-linkage clustering:

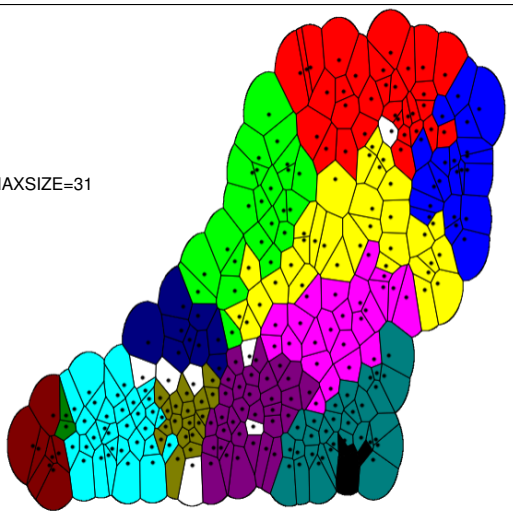
1. Find the two closest unprocessed pair of locations, based on their syntactic features
2. Combine their clusters
3. Go back to step 1

Cluster parameters: number of desired clusters, cluster size restrictions, method for computing node distances, type of input data, ...

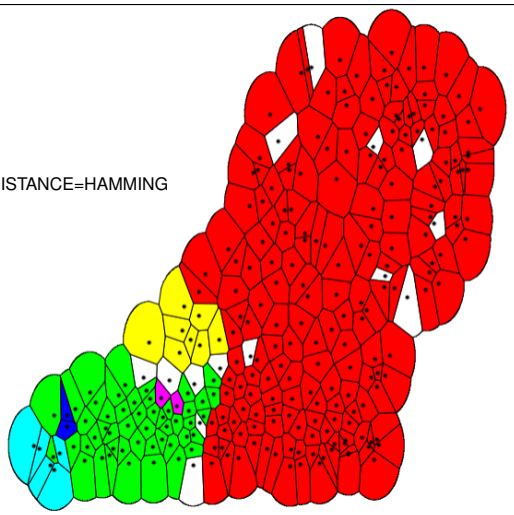
STANDARD



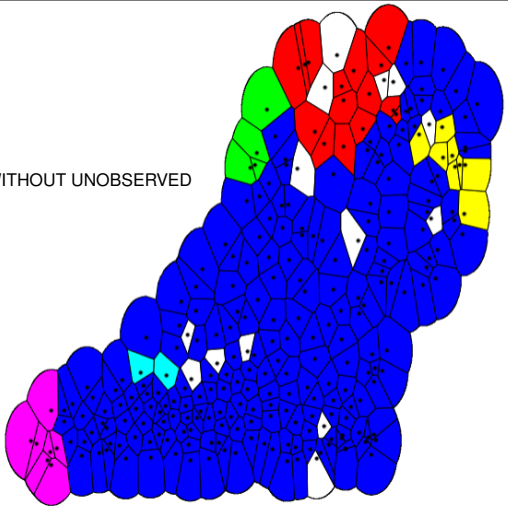
MAXSIZE=31



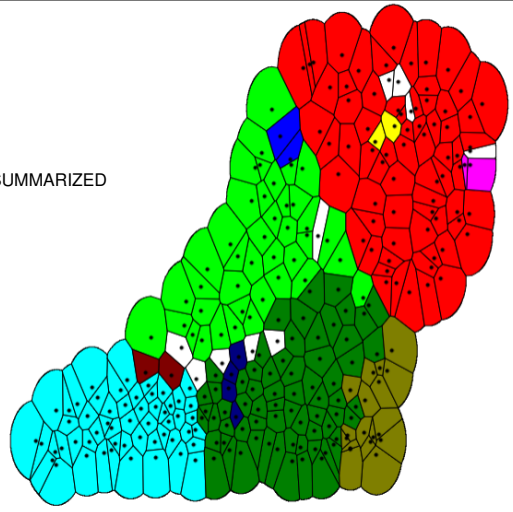
DISTANCE=HAMMING



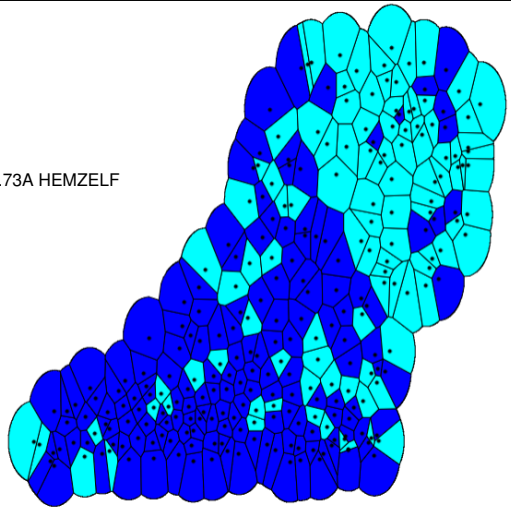
WITHOUT UNOBSERVED



SUMMARIZED



1.73A HEMZELF



Concluding remarks

We conclude that: **yes**, on their own, syntactic features of dialects are sufficient for identifying sensible dialect boundaries

Dialect area identification from inspecting maps with gradual differences (from soft clustering), can lead to different conclusions if different colors are chosen

Hard clustering can produce familiar dialect regions but the results depend on the chosen cluster parameters

THE END