The Future of Dialect Data

Representation & Analysis

Jelena Prokić, Michael Cysouw, Johann-Mattis List
Deutscher Sprachatlas
Philipps-Universität Marburg

Dealing with Survey Data

1. **Case Study:**
   Phonetischer Atlas von Deutschland

2. **Methods:**
   Multiple sequence alignment and
   Automatic isogloss drawing

3. **Looking forward:**
   The git-way of data management

Ort der Mundart/Kreis	Aufnahme-Nr.	Transkribent	Listentyp
Astfeld/ Gandersheim	I/62	Angelika Braun	A
Planrechteck X 29	Aufnahmedatum 20.11.1965	Transkribiert von 14.6.1985 bis 24.7.1985	

Besprochen von 24.07.1985 25.07.1985 L/SN

Lfd. Nr.	Stichwort		Transkription	Bemerkungen
1	Winter	178	ˈvɪntʰəʁ	ʁ = kontinuant a bereits velarisiert
* 2	fliegen	56	ˈflaɪ̯ɐ̯ɔtʰə	'fliegen die', Sequentierung unklar - kein gemeinsames [L]
3	Blätter	23	ˈblɛ̝ːdə̯ʁ̥	
4	Luft	103	lʊˈftʰ‿	ʁ = kontinuant
5	hört	89	hɪ̯ʁt‿	ʁ = kontinuant
6	gleich	78	‿klɛ̝ɔɪ̯k‿	folgt P
7	schneien	130	ˈʃnɐɔɪ̯ən	
8	Wetter	174	/	statt dessen ˈvɪtəʁʊŋə
9	tu	151	dɐɔʊ	

Phonetischer Atlas von Deutschland

- Wenker-sentences recorded in the 1960s (with additions in the 1970s)
- Selected words from the recordings were transcribed on paper in the 1980s
- A joint project between Marburg and Groningen digitised the data in the 2000s
- In total 29530 words distributed over 183 locations and 186 cognate sets

Ort der Mundart/Kreis	Aufnahme-Nr.	Transkribent	Listentyp
Astfeld/ Gandersheim	I/62	Angelika Braun	A
Planrechteck X 29	Aufnahmedatum 20.11.1965	Transkribiert von 14.6.1985 bis 24.7.1985	

Besprochen von 24.07.1985 25.07.1985 L/SN

Lfd. Nr.	Stichwort		Transkription	Bemerkungen
1	Winter	178	ˈvɪntʰəʁ	ʁ = kontinuant ə bereits velarisiert
* 2	fliegen	56	ˈflaɪ̯ɐ̯ɔtʰə	'fliegen die', Sequentierung unklar - kein gemeinsames [L]
3	Blätter	23	ˈblɛ̝ːdə̯ʁ̥	
4	Luft	103	lʊrˈftʰ‿	ʁ = kontinuant
5	hört	89	hɪ̯ʁt‿	ʁ = kontinuant
6	gleich	78	‿klɛ̝ɐ̯ɪk‿	folgt P
7	schneien	130	ˈʃnɛɐ̯ɪən	
8	Wetter	174	/	statt dessen ˈvɪtəʁʊŋə
9	tu	151	dɛɐ̯ʊ	

Phonetischer Atlas von Deutschland

- Digitised in X-SAMPA, converted back to match original transcriptions, minor corrections for consistency of encoding
- The data is transcribed in high phonetic detail (3786 different phonetic segments)
- We will make the complete data available
  - electronically, separated by phonetic segments
  - as close as possible to the original source
  - including all idiosyncrasies

Multiple Sequence Alignment

- Just a fancy name for sound correspondences
- Each sound correspondence is "aligned" in a column, possibly adding empty cells
- It is a useful and consistent way to represent comparative data (both between languages or dialects)

LOCATION	WORD
Aachen	a̱ːph
Adorf	a̱ːbʱə
Ahrbergen	o̝→o̞phə̹
Albersloh	ɑːpᵑ̈ə̈
Allna	ɑɸh
Altenberg	ʌfɛ̈
Altentrüdin	a̱f
Altlandsberg	ɑʾfəʾ
Altwarp	oːph
Astfeld	pˤːpʰə
Atzendorf	afɛ̱
Ballhausen	ʌʾfə
Bardenfleth	o̞ːp͡ɸ
Barssel	o̞ːpʲə
Bempflingen	afː
Bennin	o̞pʰ
Billingsbach	a̱f
Bockelwitz	ʌ̹və̈
Bonn	ɑːpʾ
Borstendorf	ɤ̞fː
Breddin	pːph
Brelingen	ɑ̹fβ͡ə
Bremscheid	pˤːpḫə̹
…	…

A	FF	E
a̱ː	ph	-
ɑ̹ː	bʱ	ə
o̝→o̞	ph	ə̹
ɑː	pᵑ̈	ə̈
ɑ	ɸh	-
ʌ	f	ɛ̈
a̱	f	-
ɑʾ	f	əʾ
oː	ph	-
pˤː	pʰ	ə
a	f	ɛ̱
ʌʾ	f	ə
o̞ː	p͡ɸ	-
o̞ː	pʲ	ə
a	fː	-
o̞	pʰ	-
a̱	f	-
ʌ̹	v	ə̈
ɑː	pʾ	-
ɤ̞	fː	-
pː	ph	-
ɑ̹	f͡β	ə
pˤː	pḫ	ə̹
…	…	…

- **Workflow:**
  - Tokenisation of segments
  - Automatic alignment using LingPy (github.com/lingpy)
  - Manual correction
  - Separation of cognates (e.g. *Samstag* vs. *Sonnabend*)
  - Annotation of columns (e.g. many-to-one alignments, metathesis)
  - Merging of complex columns and removing boundaries

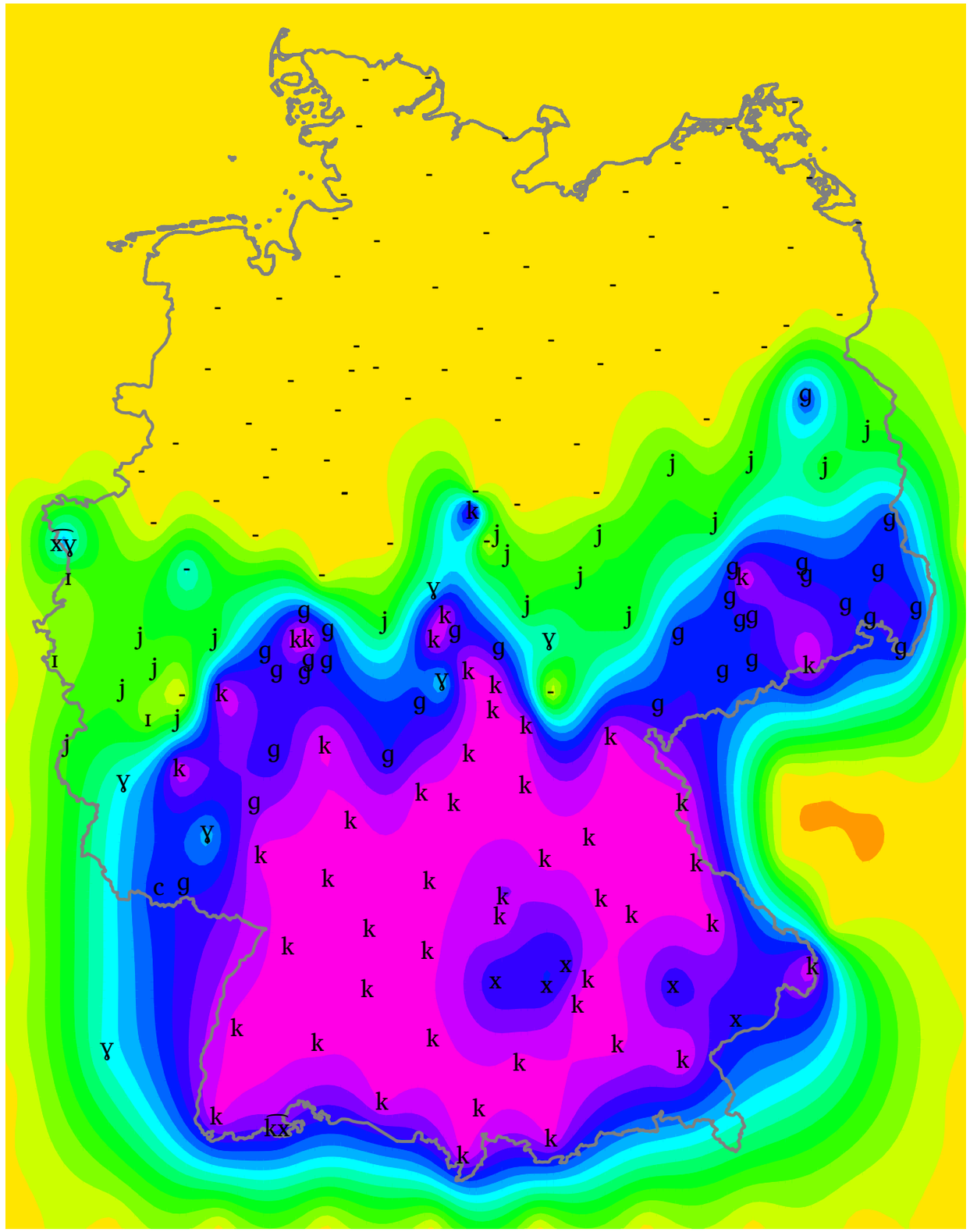
Automatic Isoglossing

- Each column represents an isogloss
- Plotting one column gives a dialect map (after some interpretation of the data)
- Areas can be added by using 3D-interpolation (or any other of the many methods being developed currently)

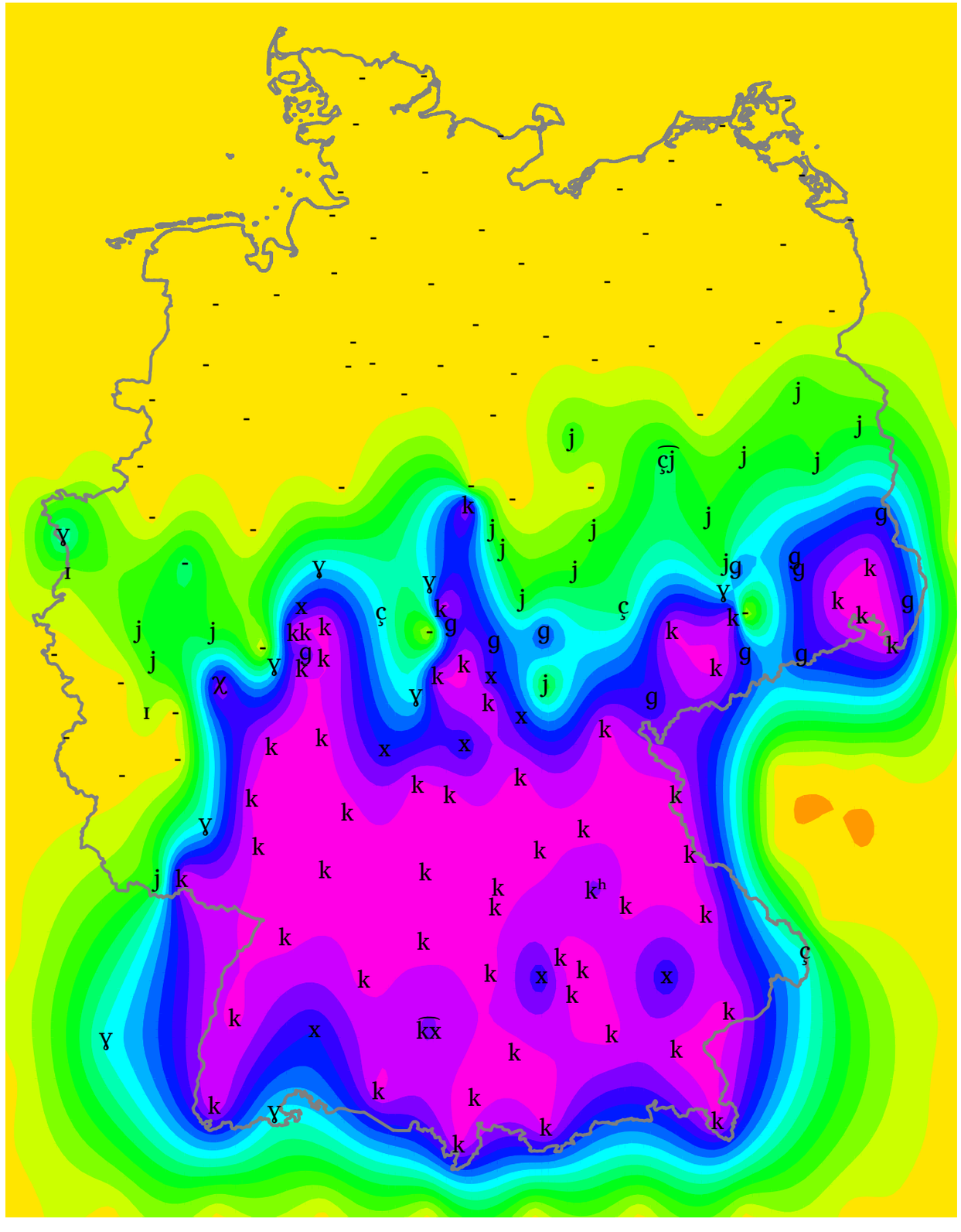
Example 1

- **Consonant** in German Perfect-prefix ge-
- Ordered to ‘strength’, visualised as ‘height’
- (NULL) ʏ̟ ɪ j̩ j j̲ ʔ c͡ɕ z̩ ʑ̩ ʒ ɕ̩ ɣ̩ ɕ c͡ɕ ɣ ɟ̩ ɕ̲ ɣ̲ t͡ɕ g͡ɣ x͡ɣ g̩ x̩ c g k͡ɕ k͡j x ç̲ x̲ kʰ k͡x k̩ ʀ̥ χ k k̲ q k’

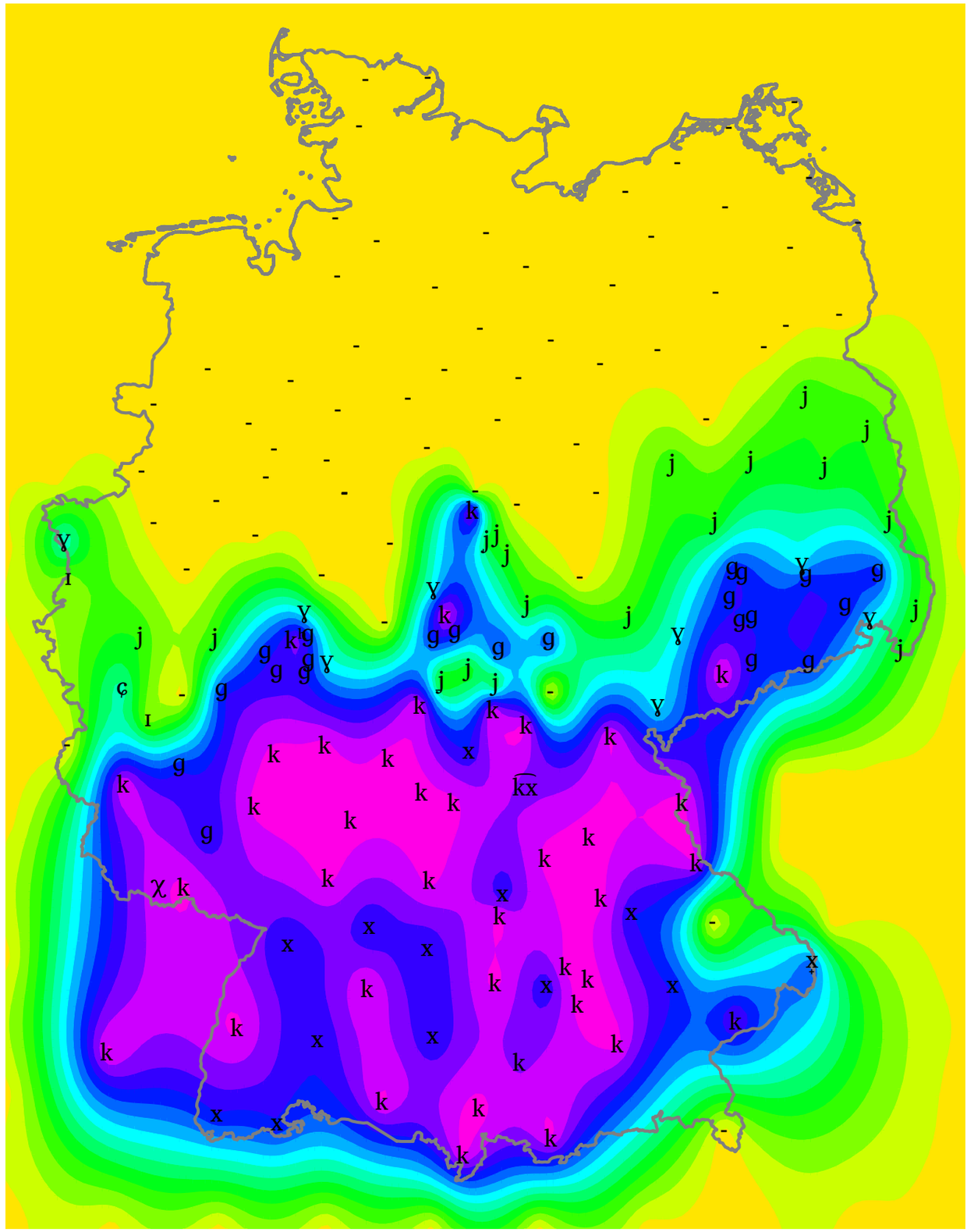
gefahren



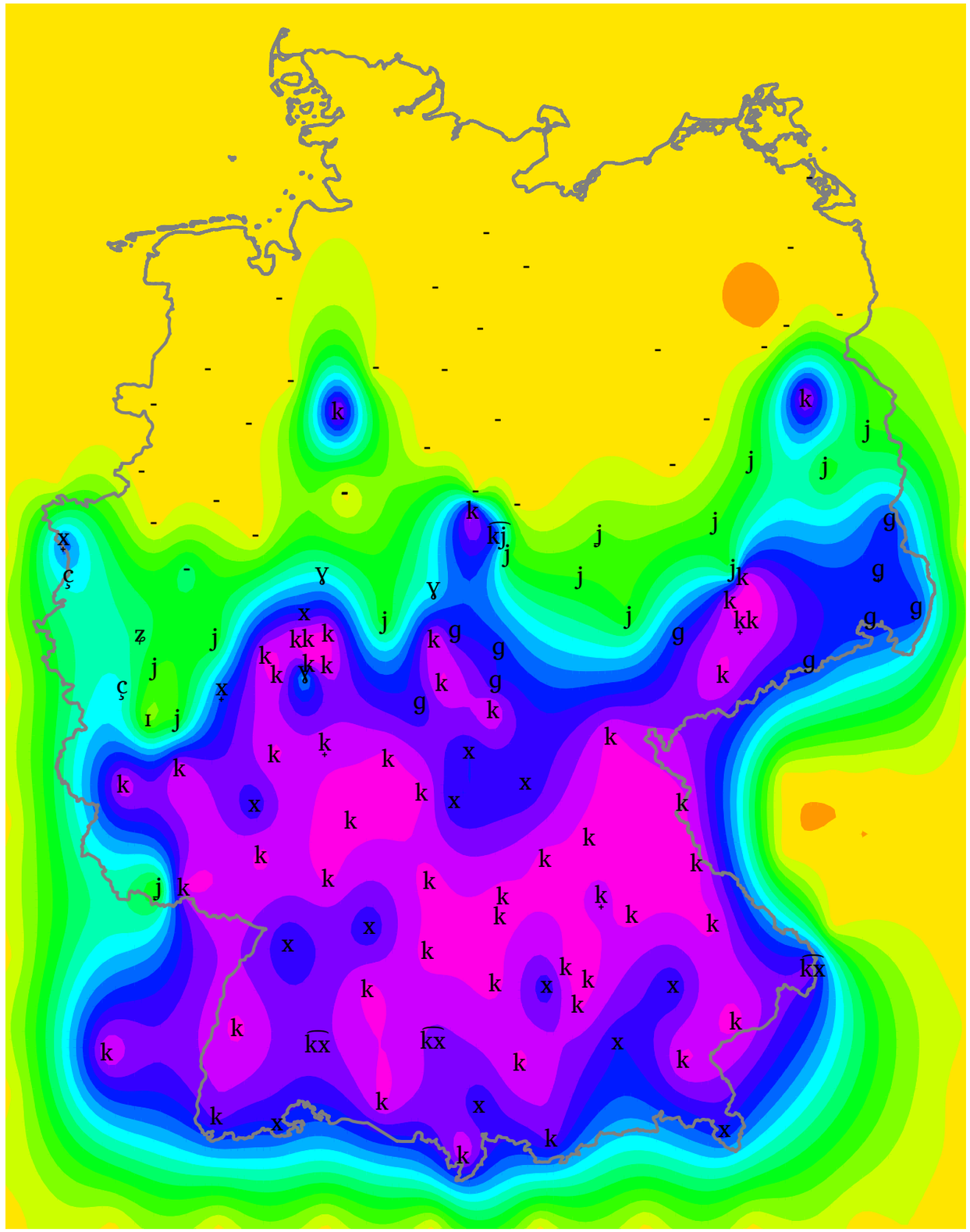
gefunden



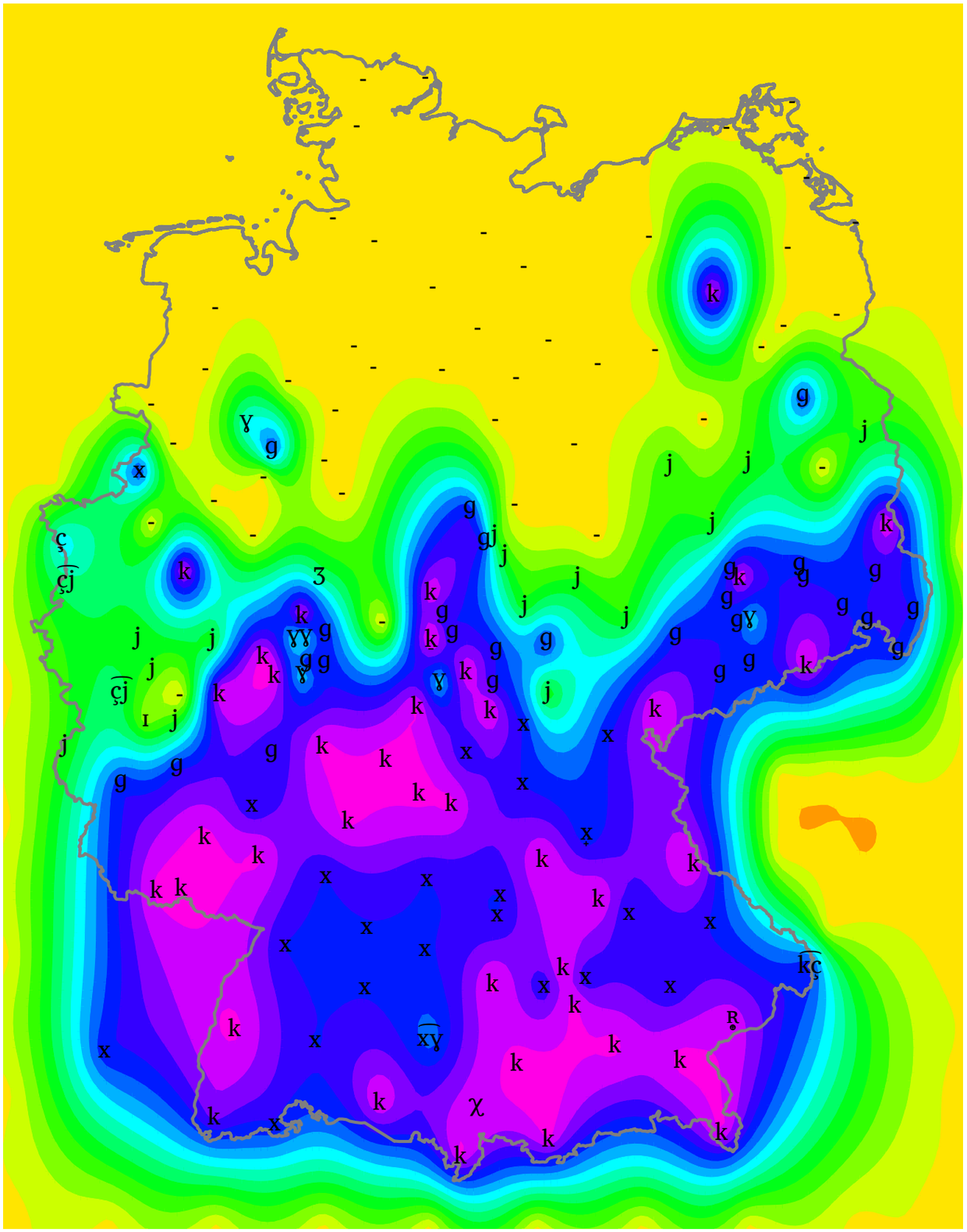
eingeschlafen



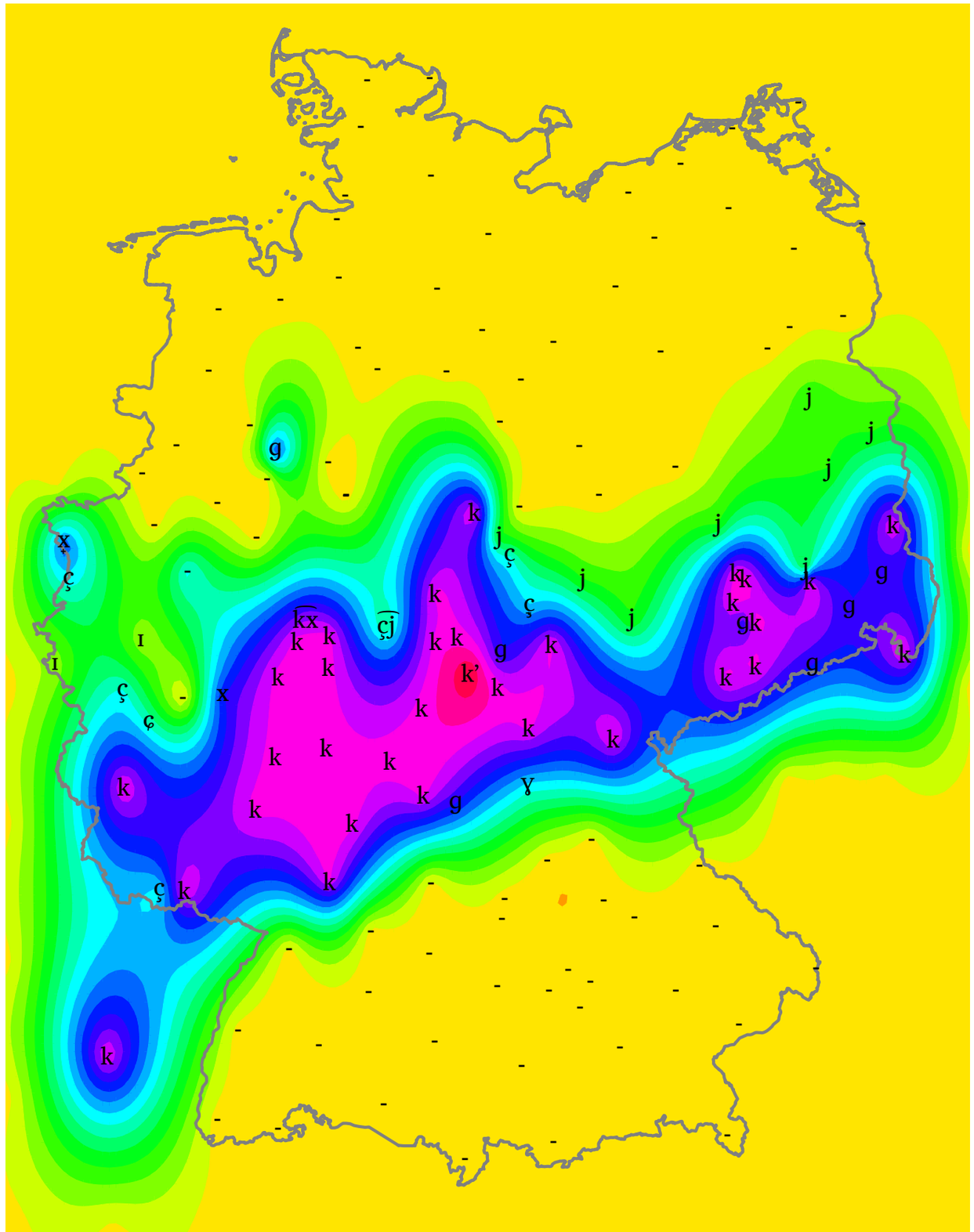
gestohlen



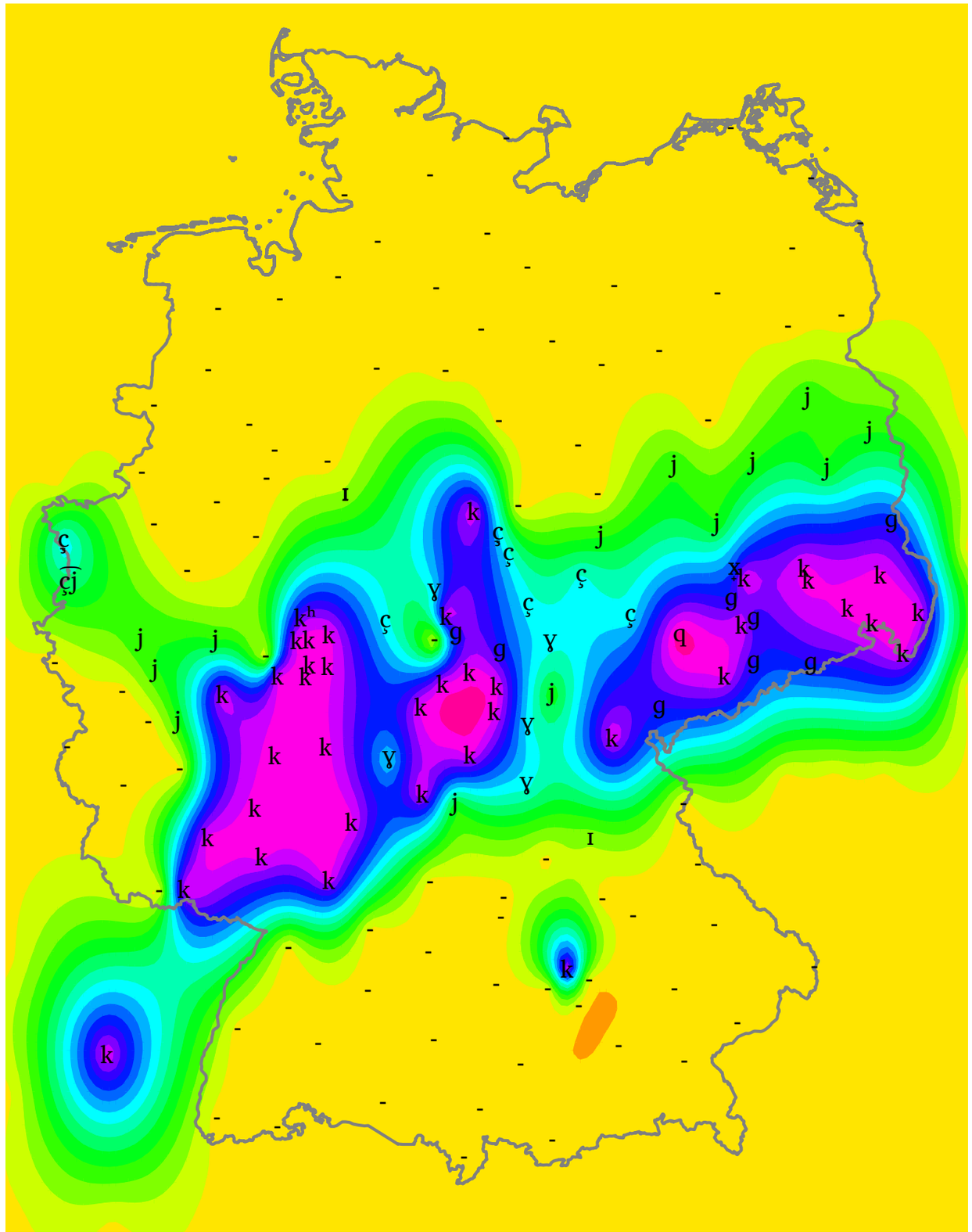
gestorben



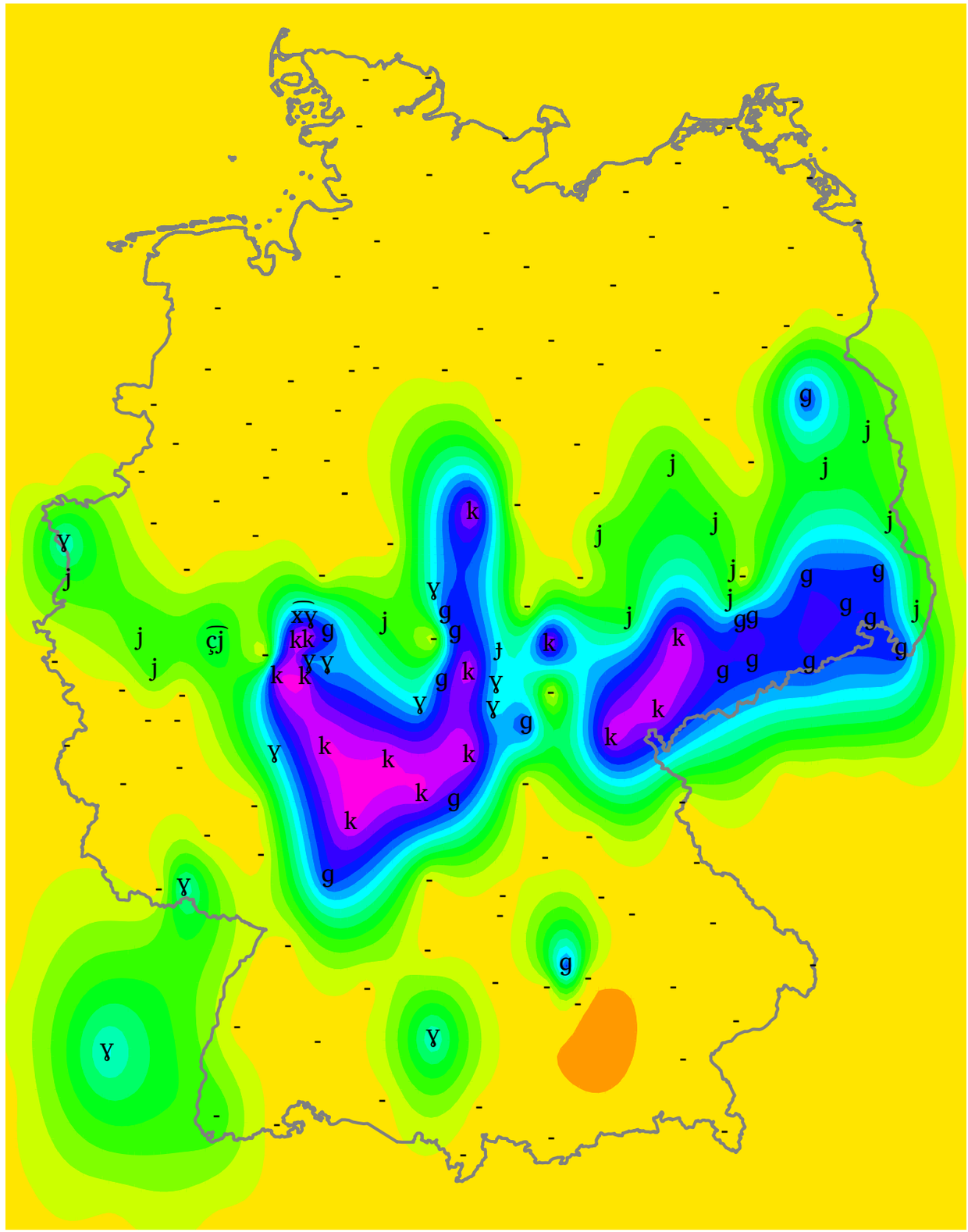
gebrannt



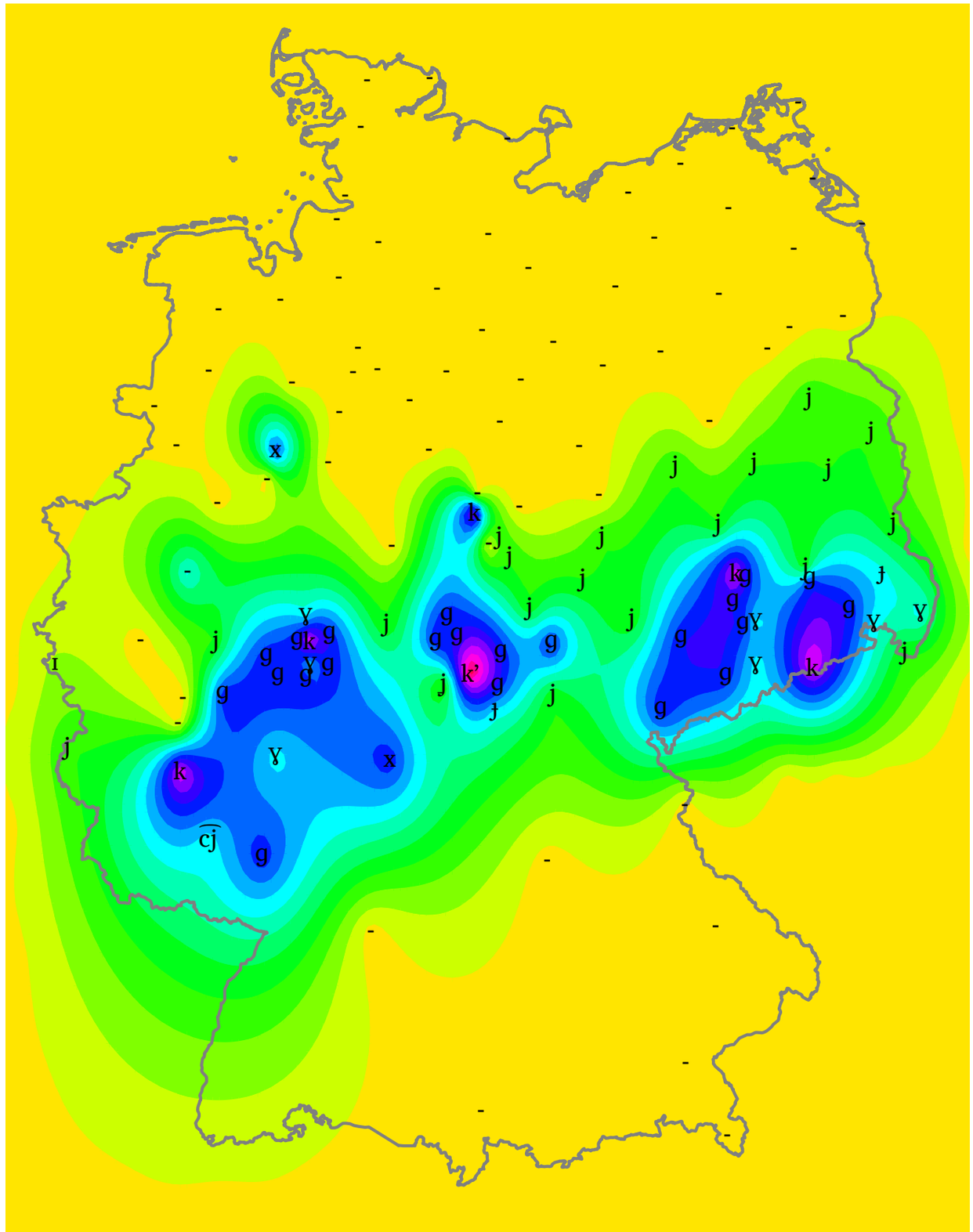
gebracht



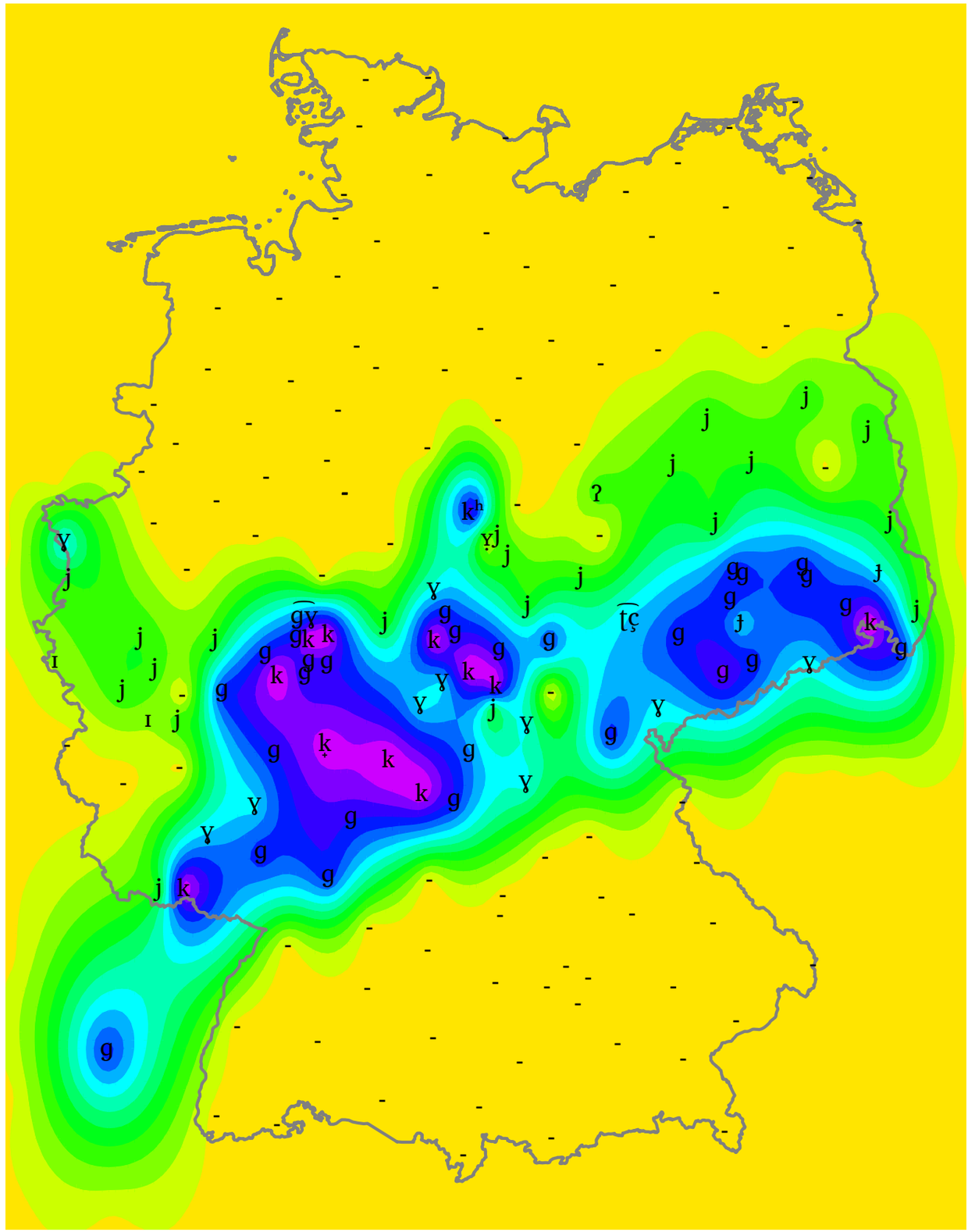
geblieben



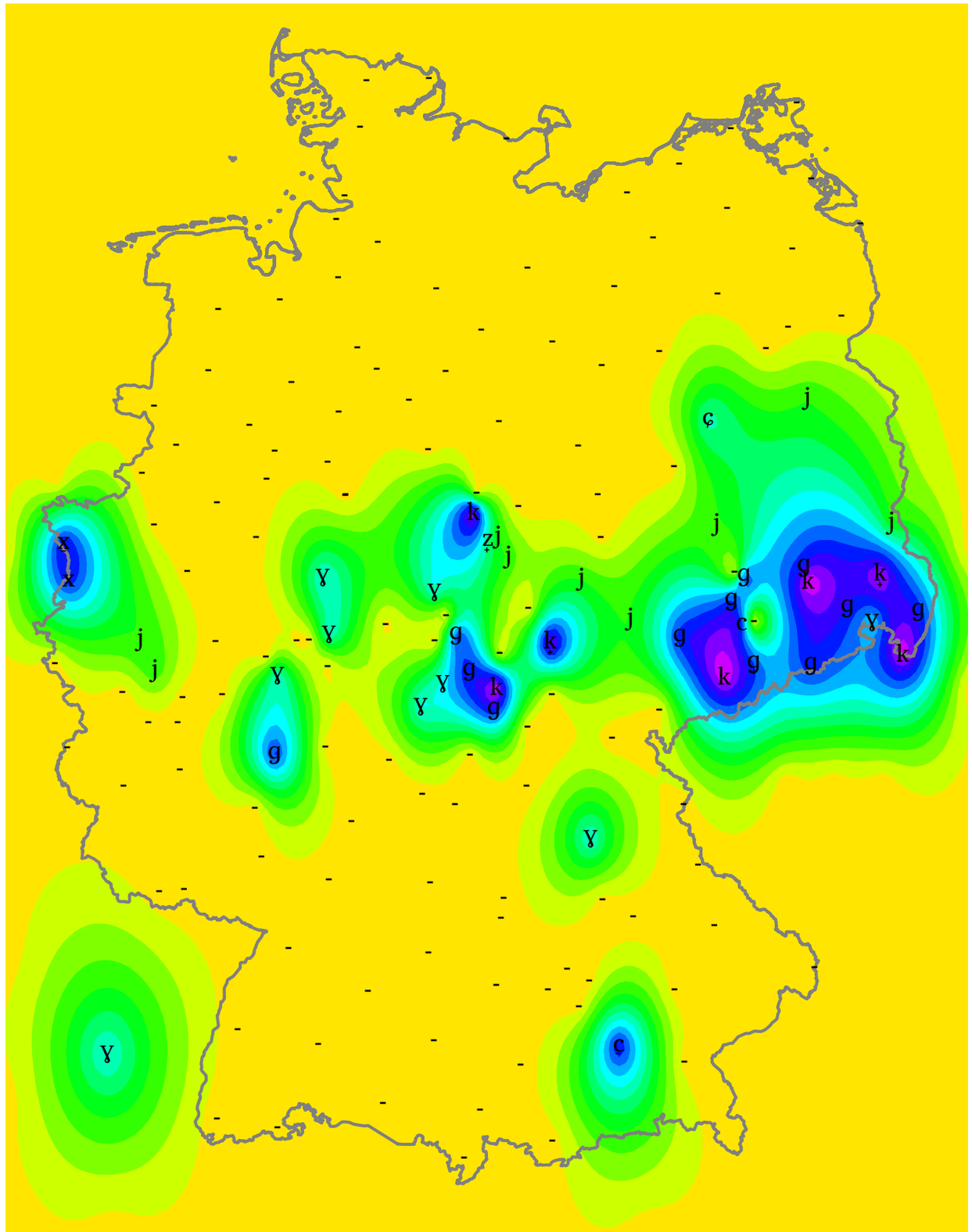
eingebrochen



gekannt



gekommen

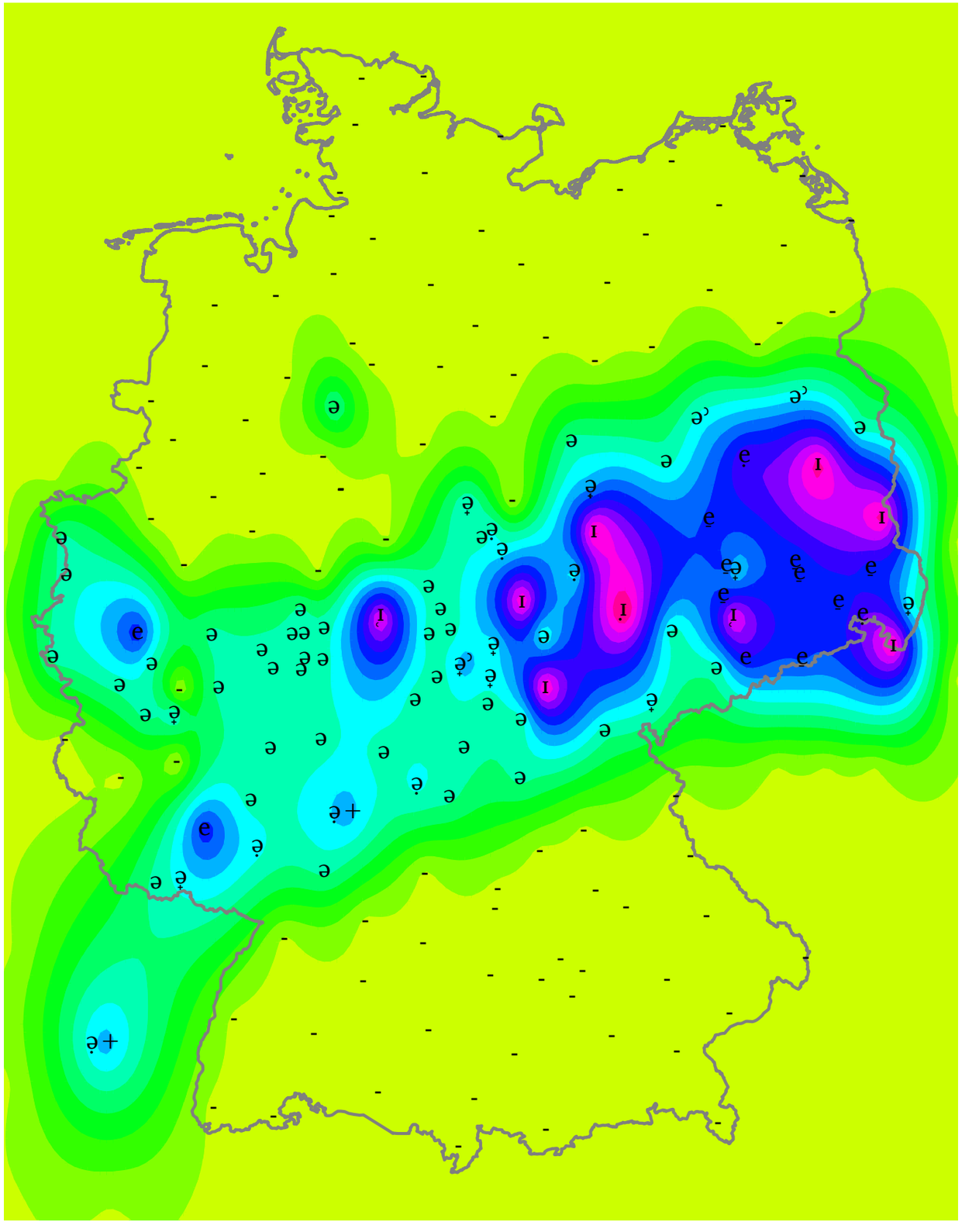


Example 2

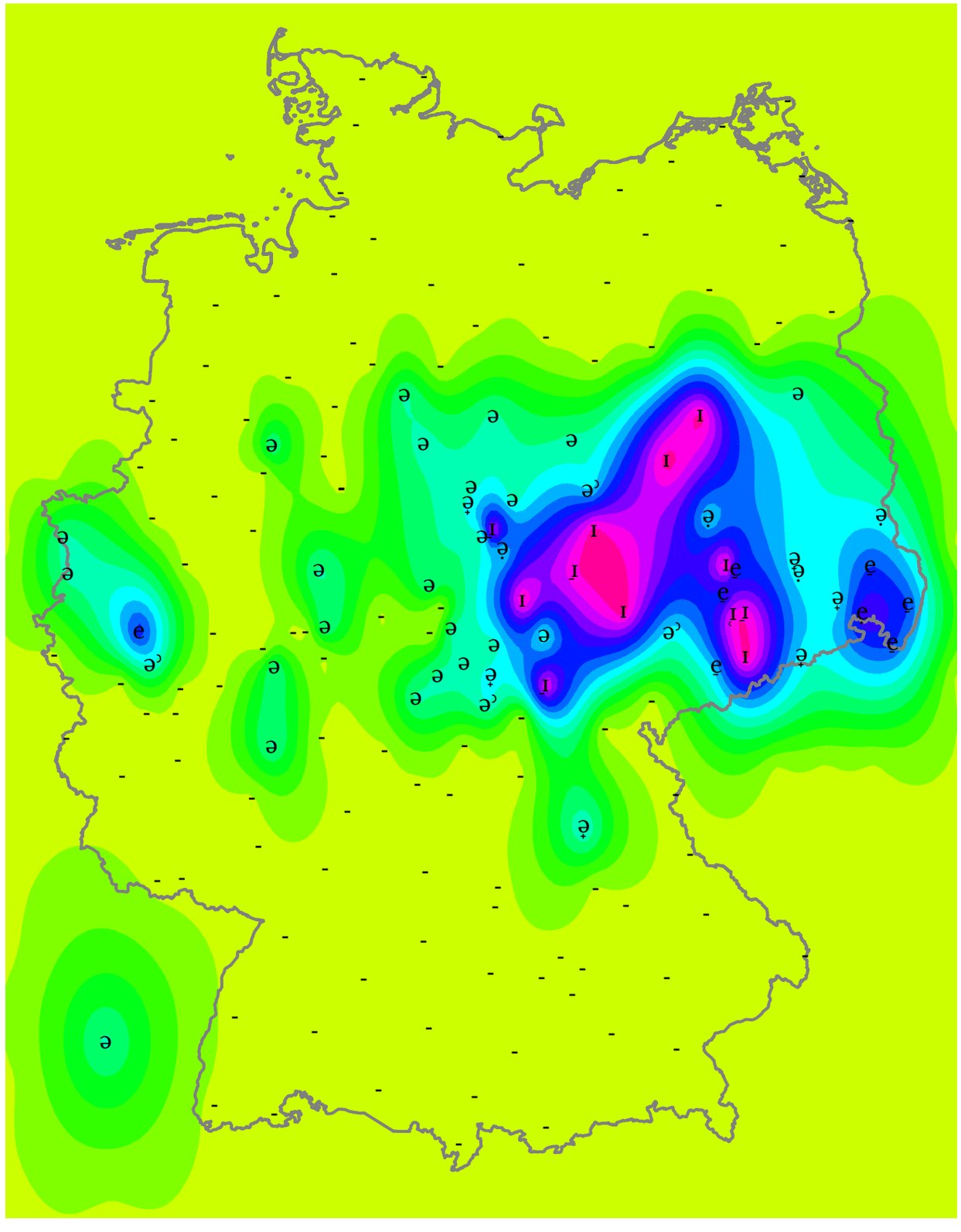
- **Vowel** in German Perfect-prefix ge-
- Ordered to ‘strength’, visualised as ‘height’
- (CLUSTER) (NULL) ə̹ ə̄ ə əˁ ə̟ ə̗ əˀ ə̟ˀ ə̗ˀ

  ə+ ə̗+ ə̟ˀ ə̗ˀ ē e ēˀ ẹ œˁ œ ɛ̄ ɛ̣ I̹ Iˀ Ī I̗ I

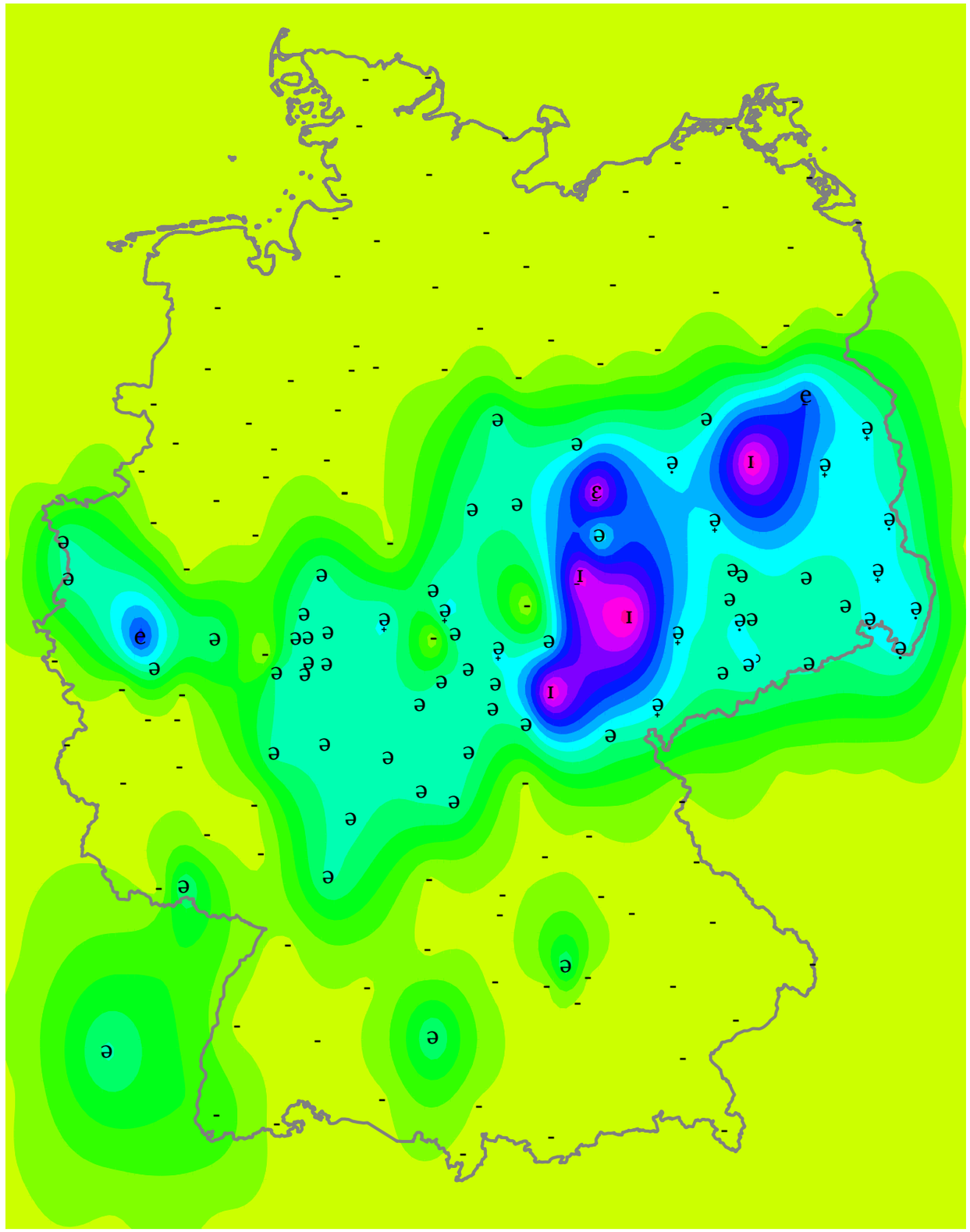
gekannt



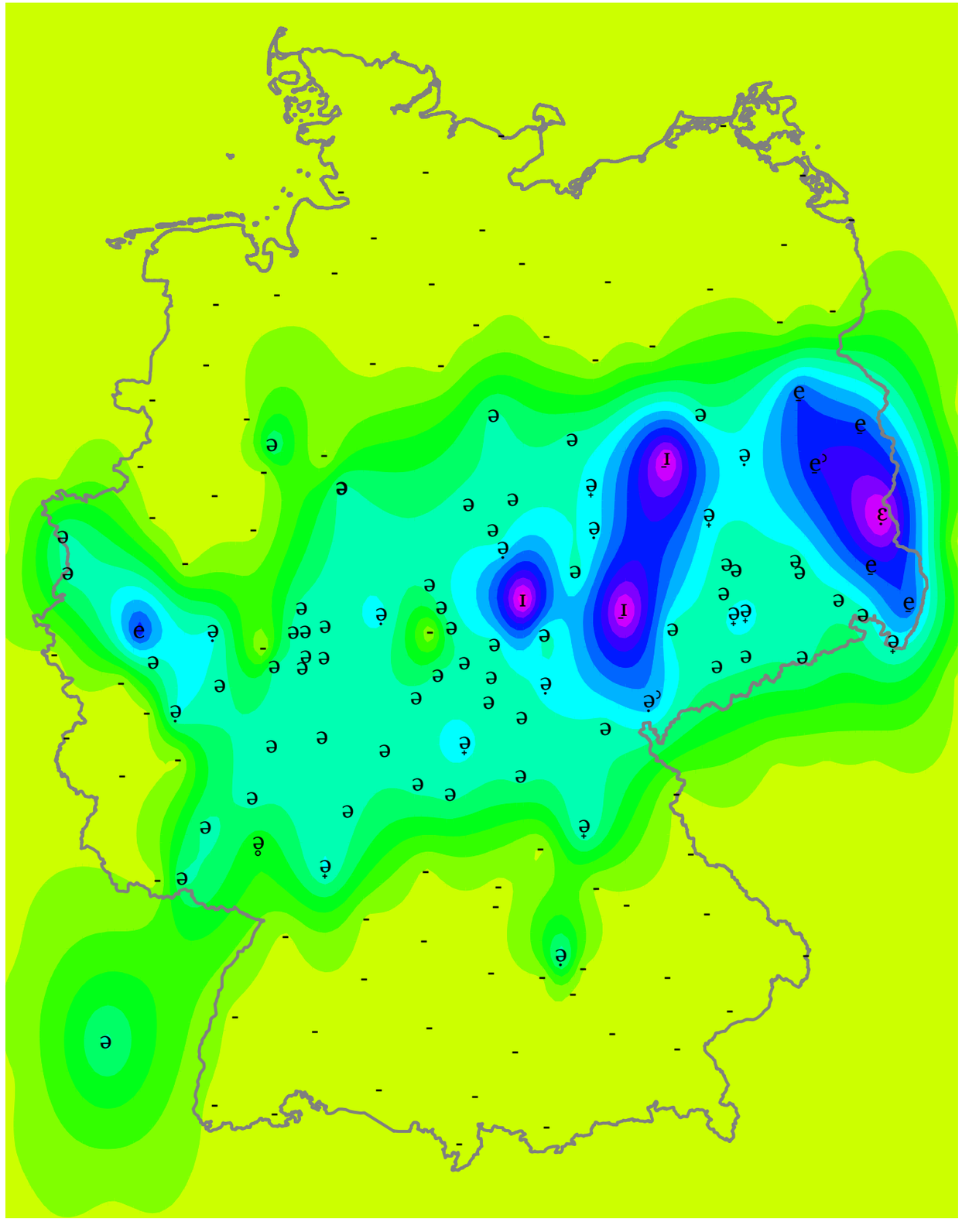
gekommen



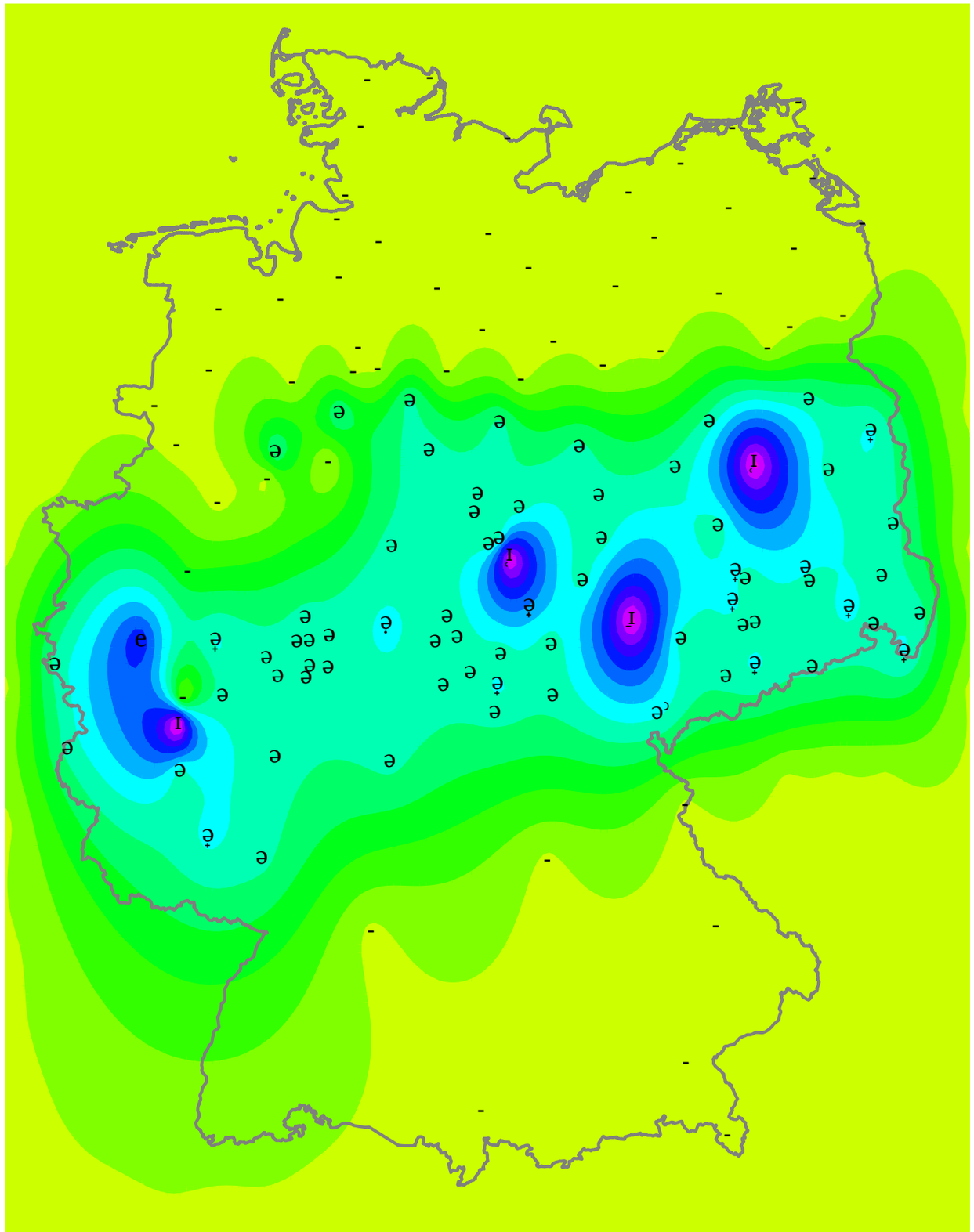
geblieben



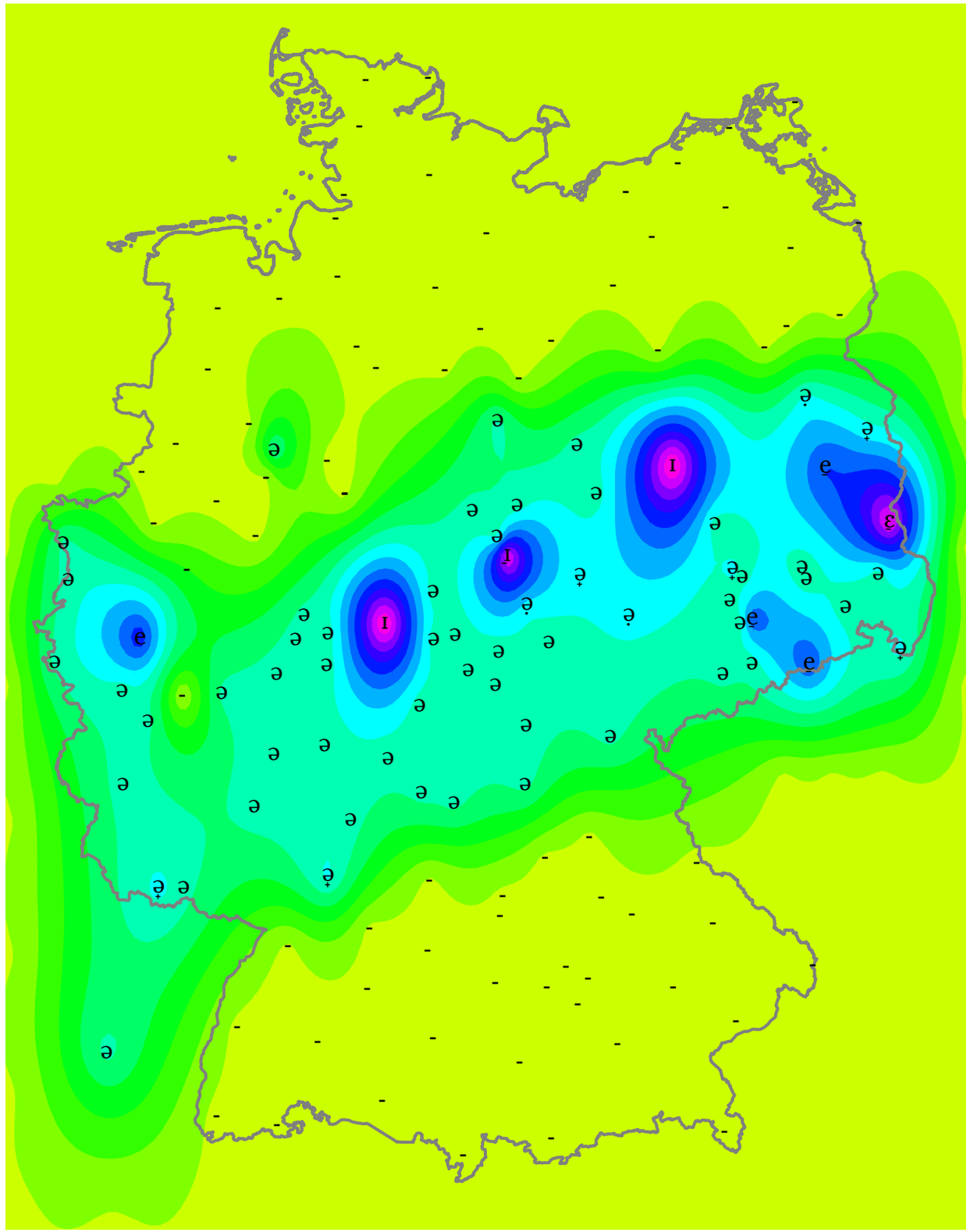
gebracht



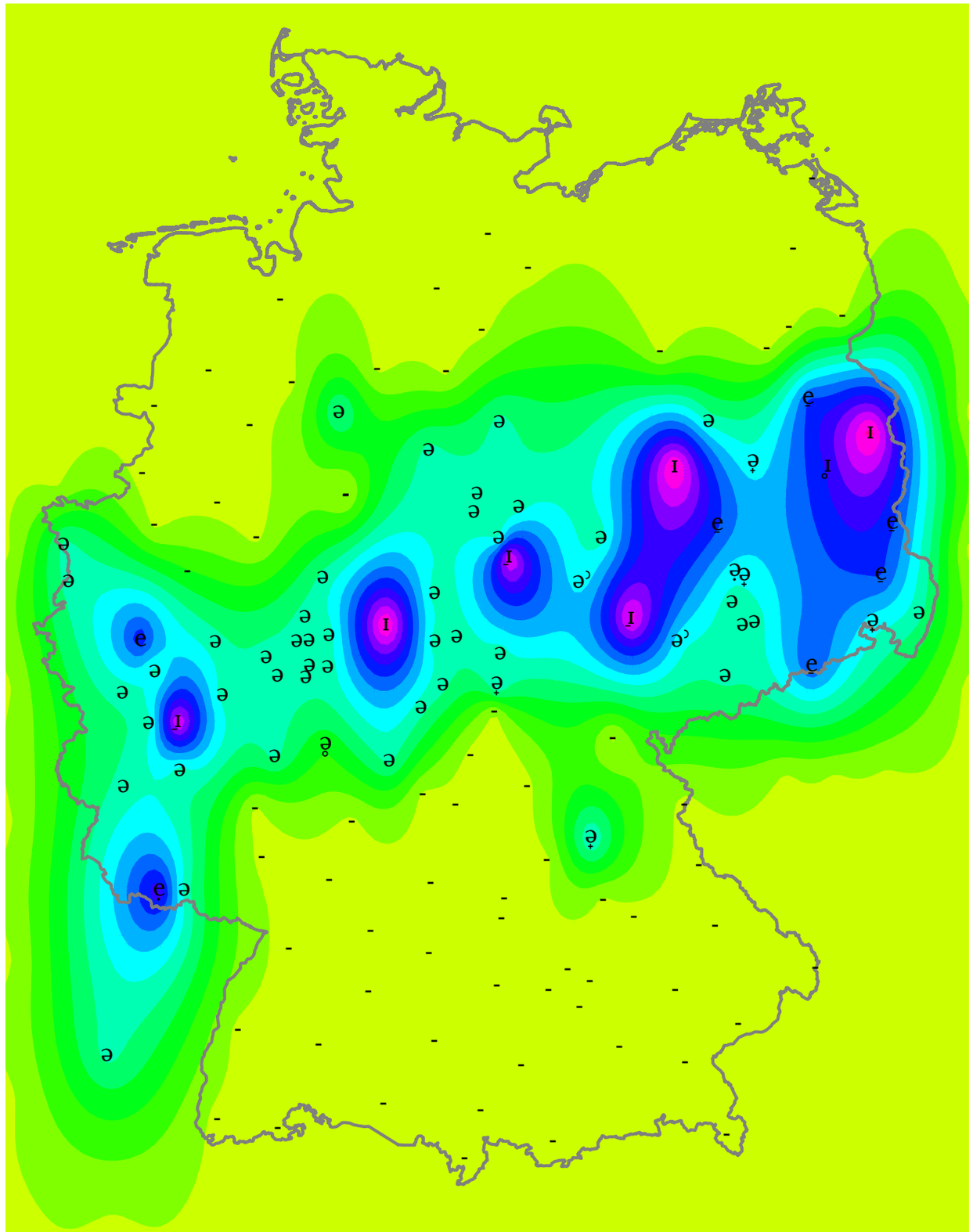
eingebrochen



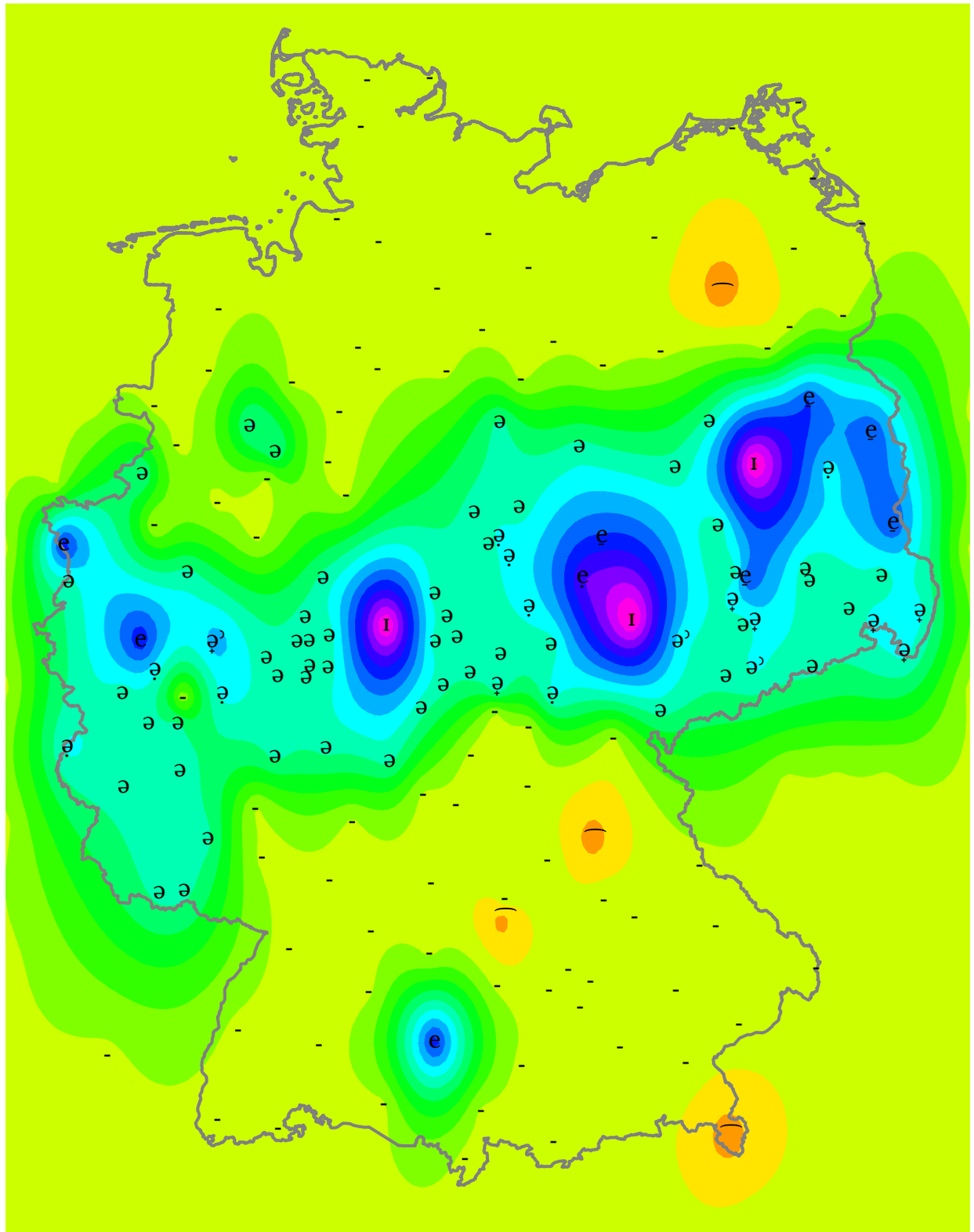
gebrannt



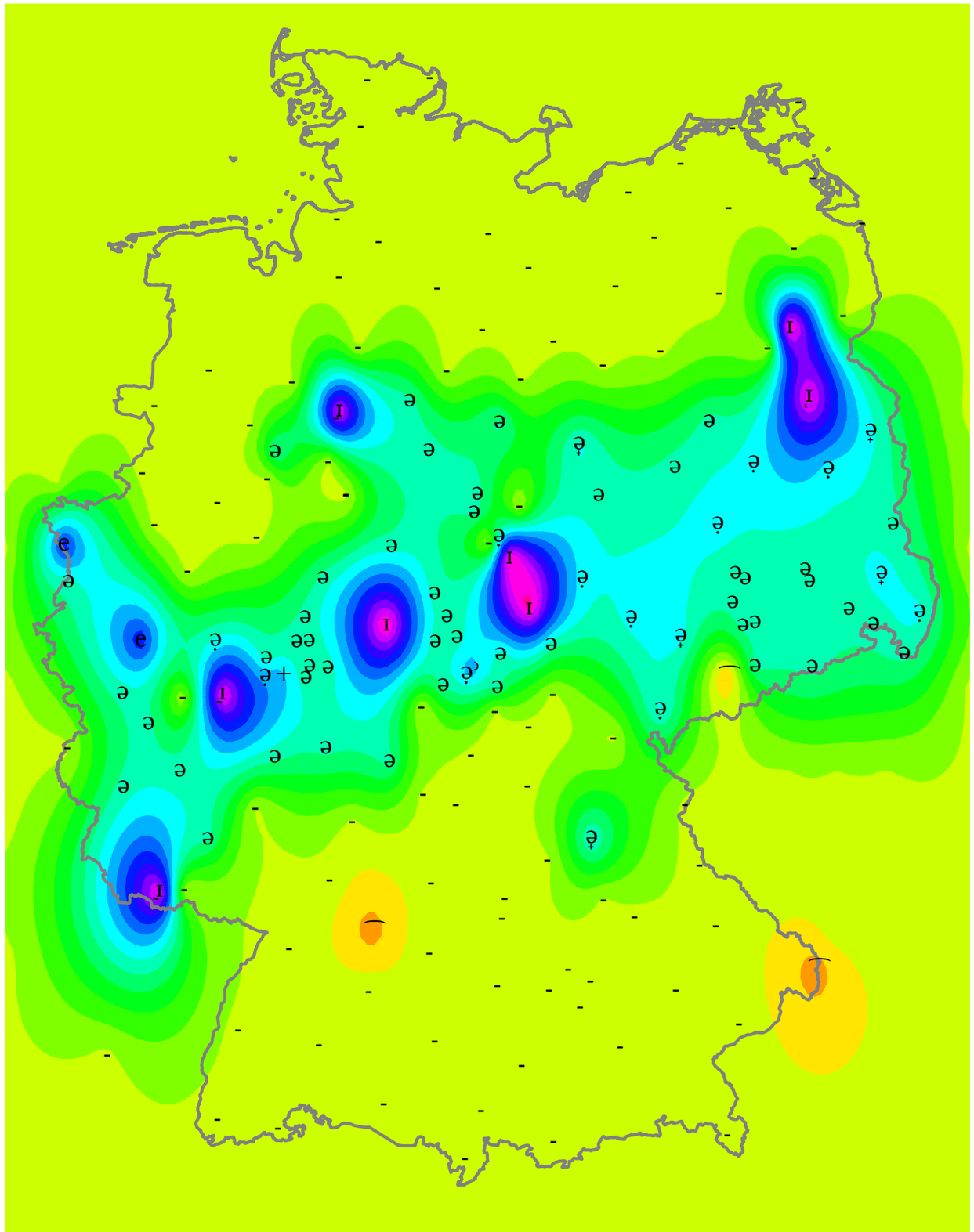
gestohlen



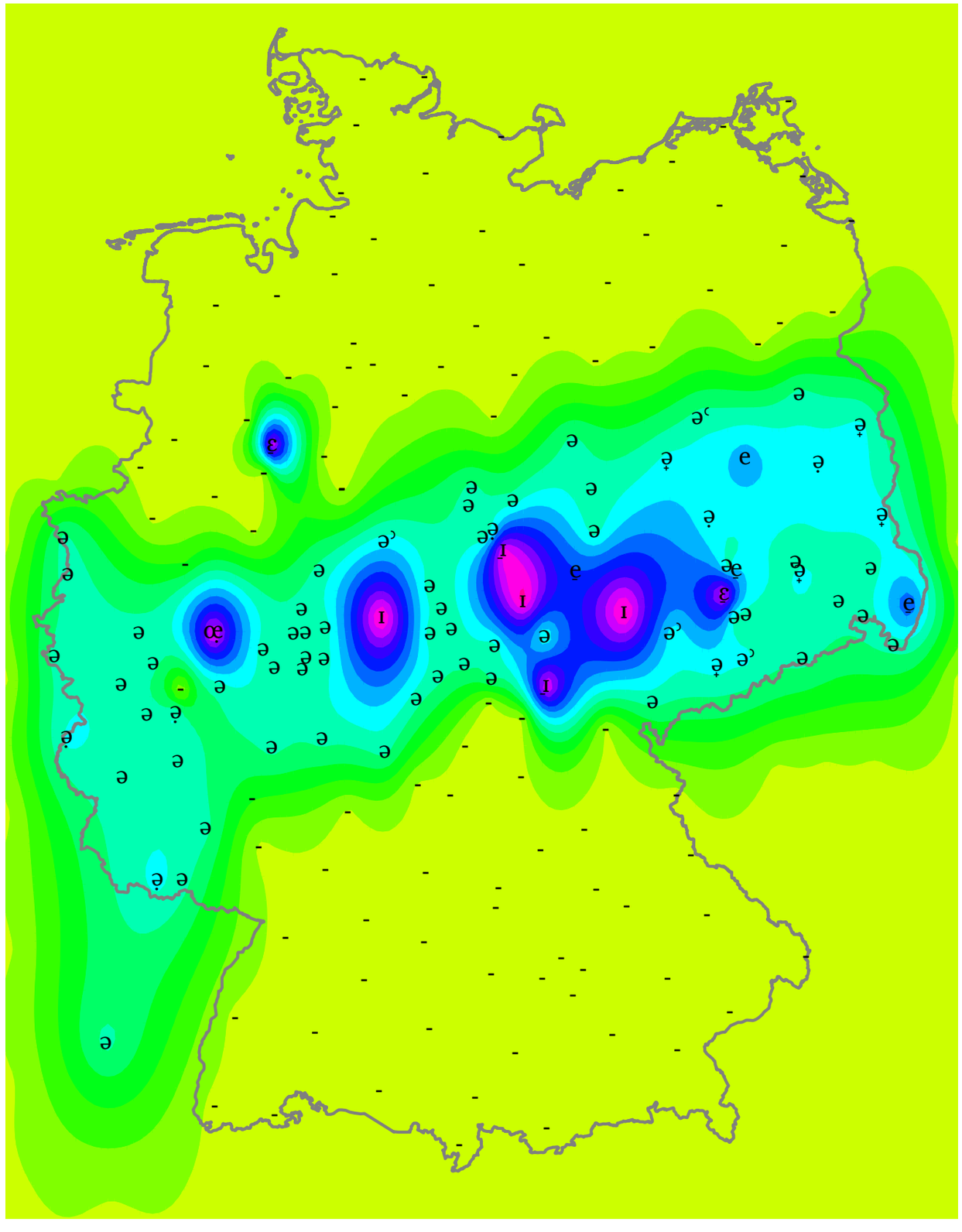
gestorben



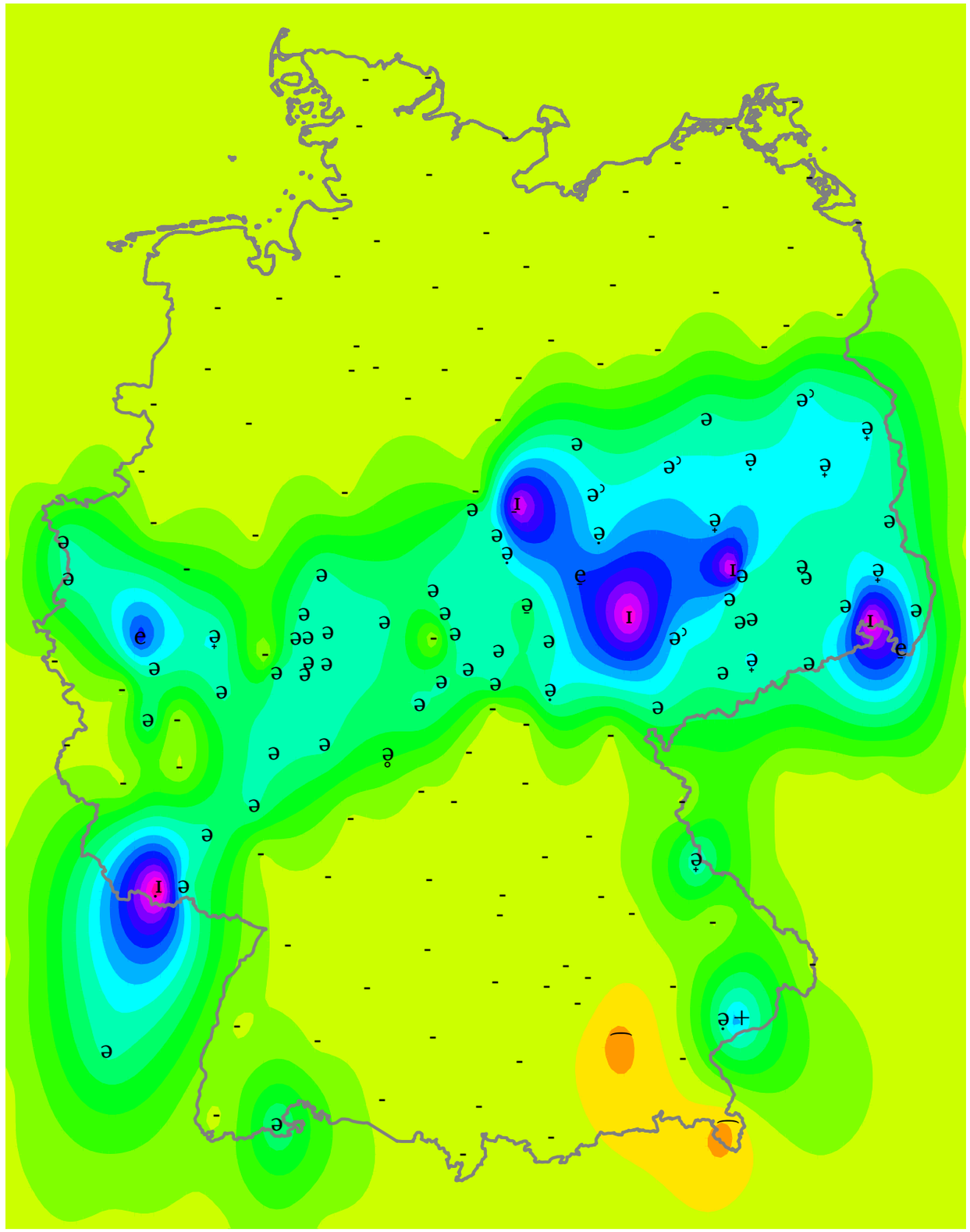
eingeschlafen



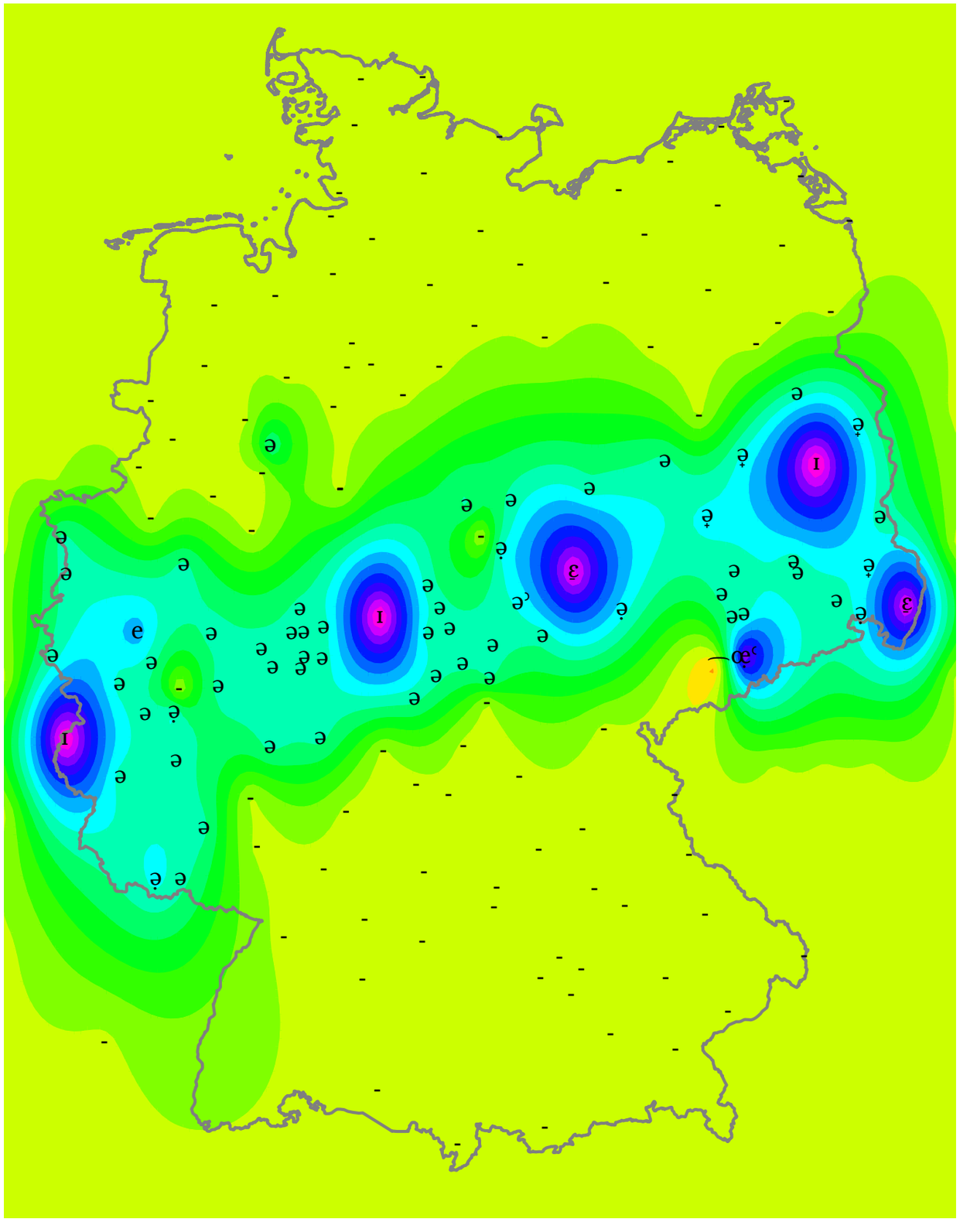
gefahren



gefunden



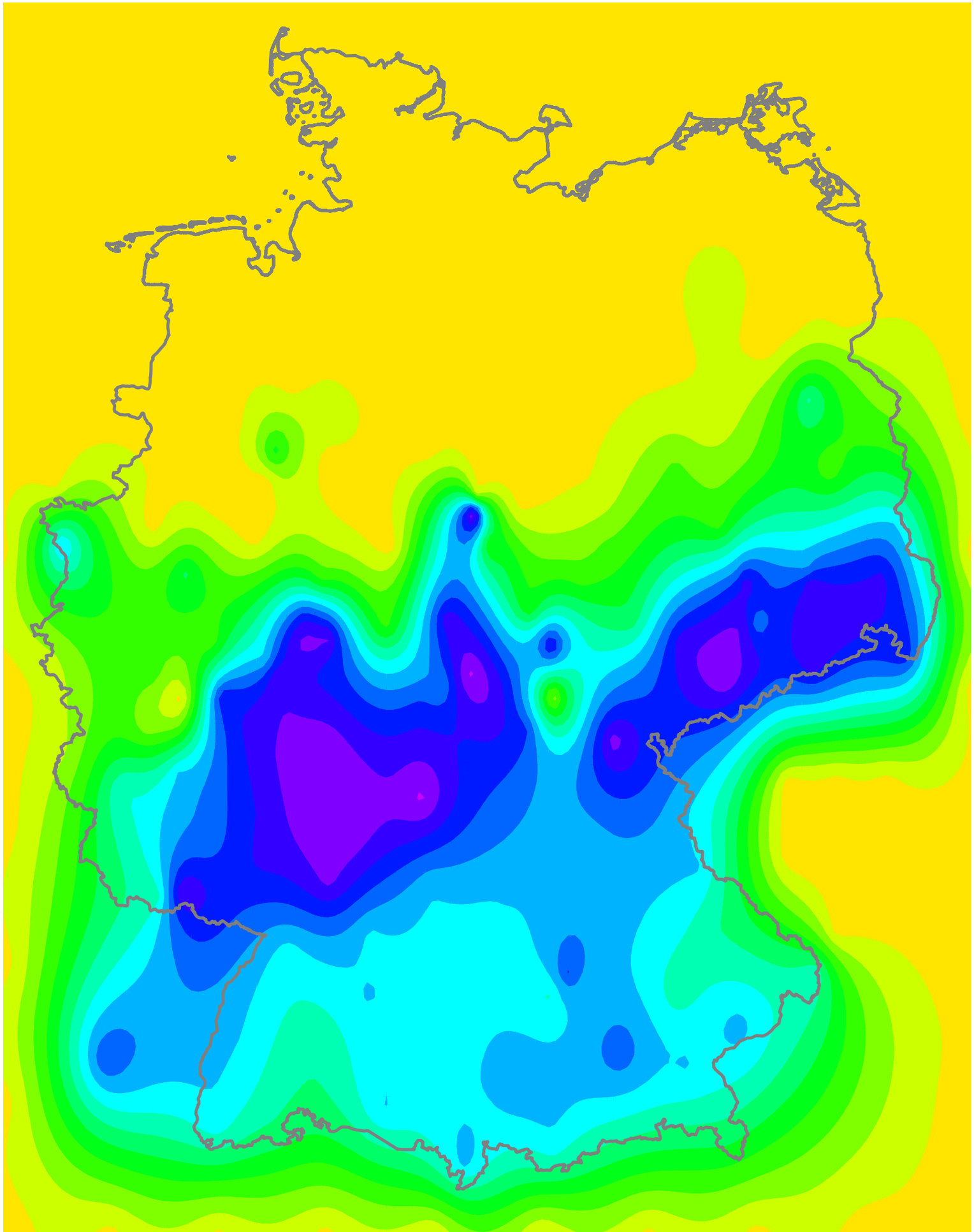
gefallen



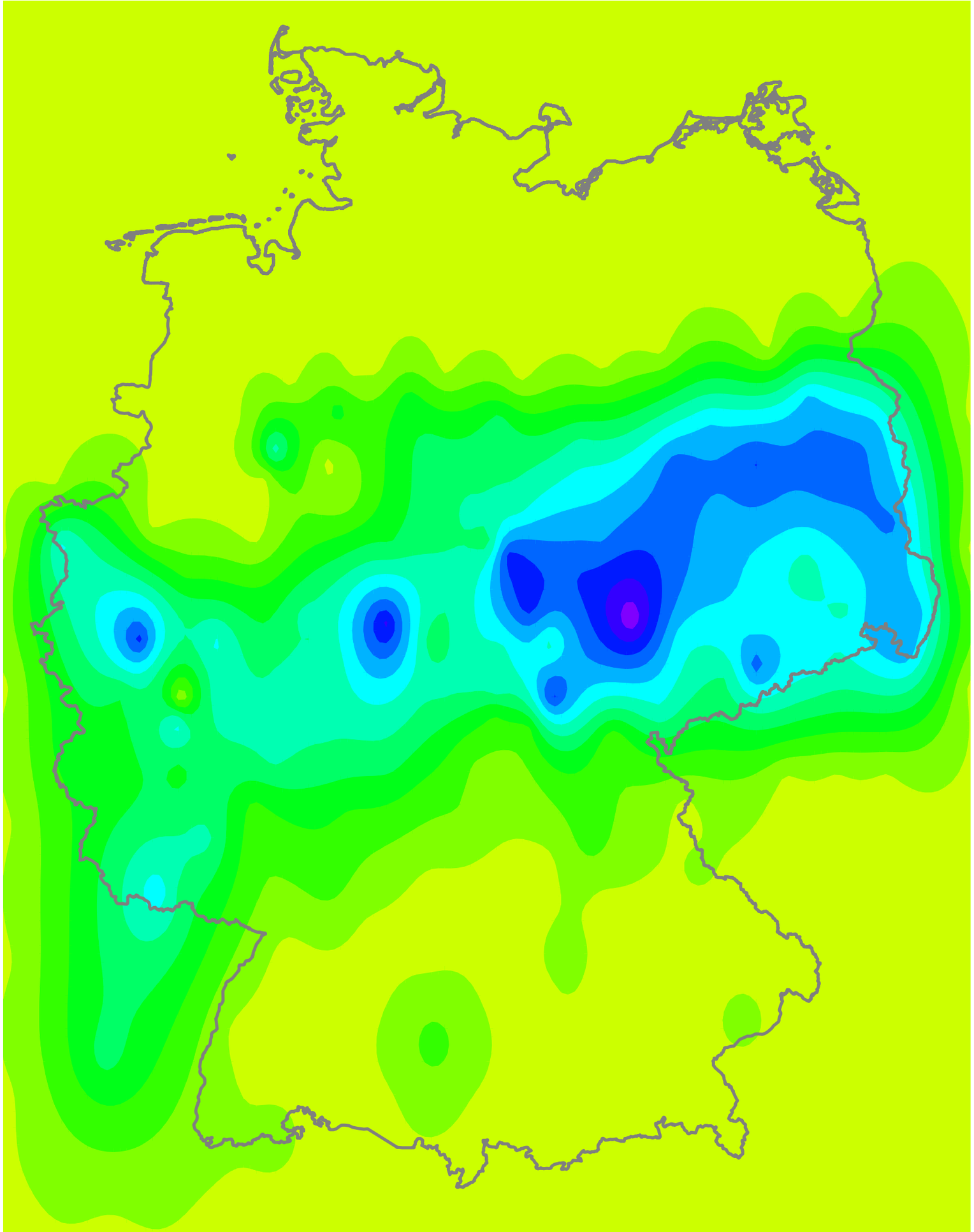
Averaging

- Combining all consonants and all vowels into a single map

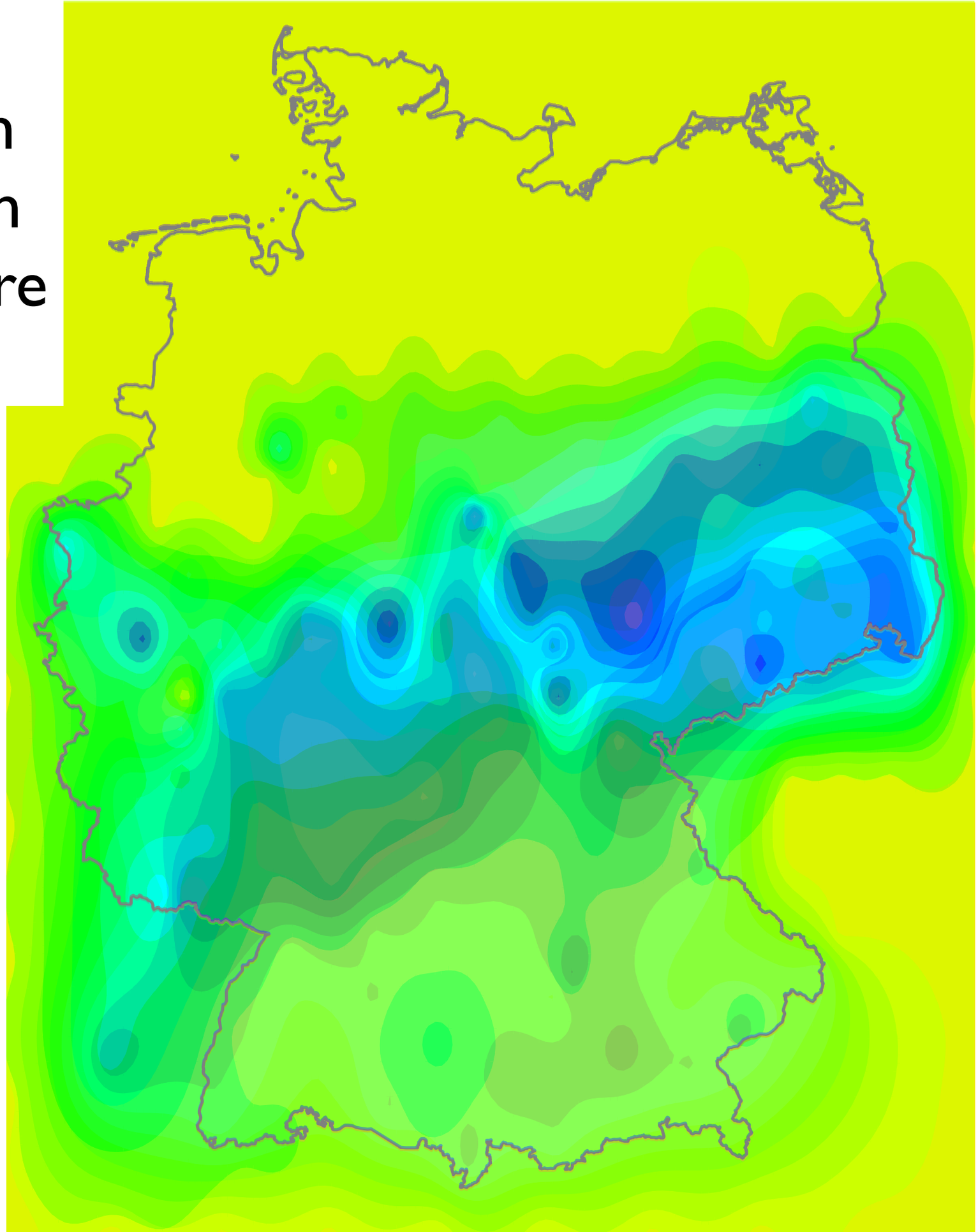
Consonants
of ge- prefix
(averages over
14 lexemes)



Vowels
of ge- prefix
(averages over
14 lexemes)

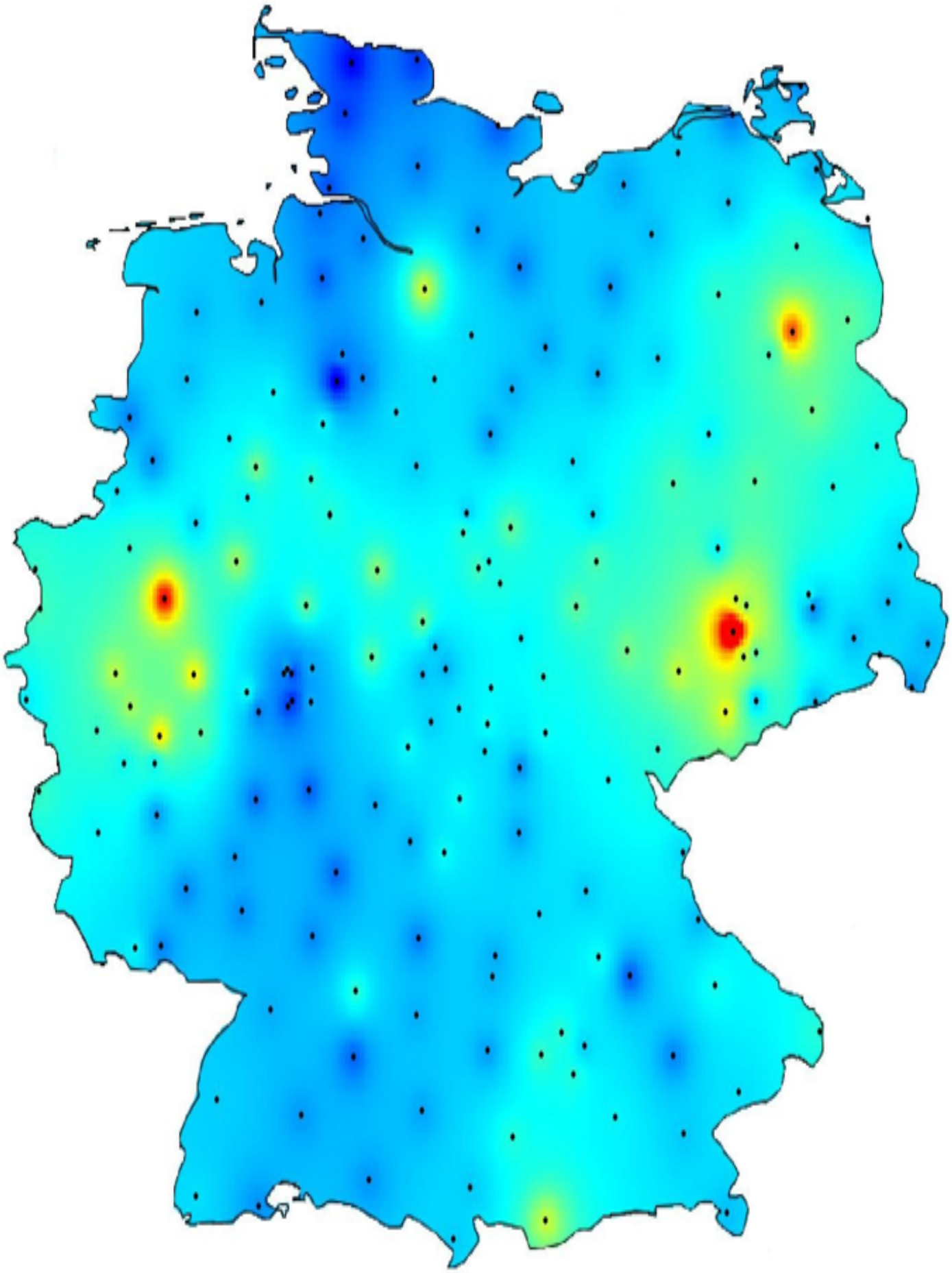


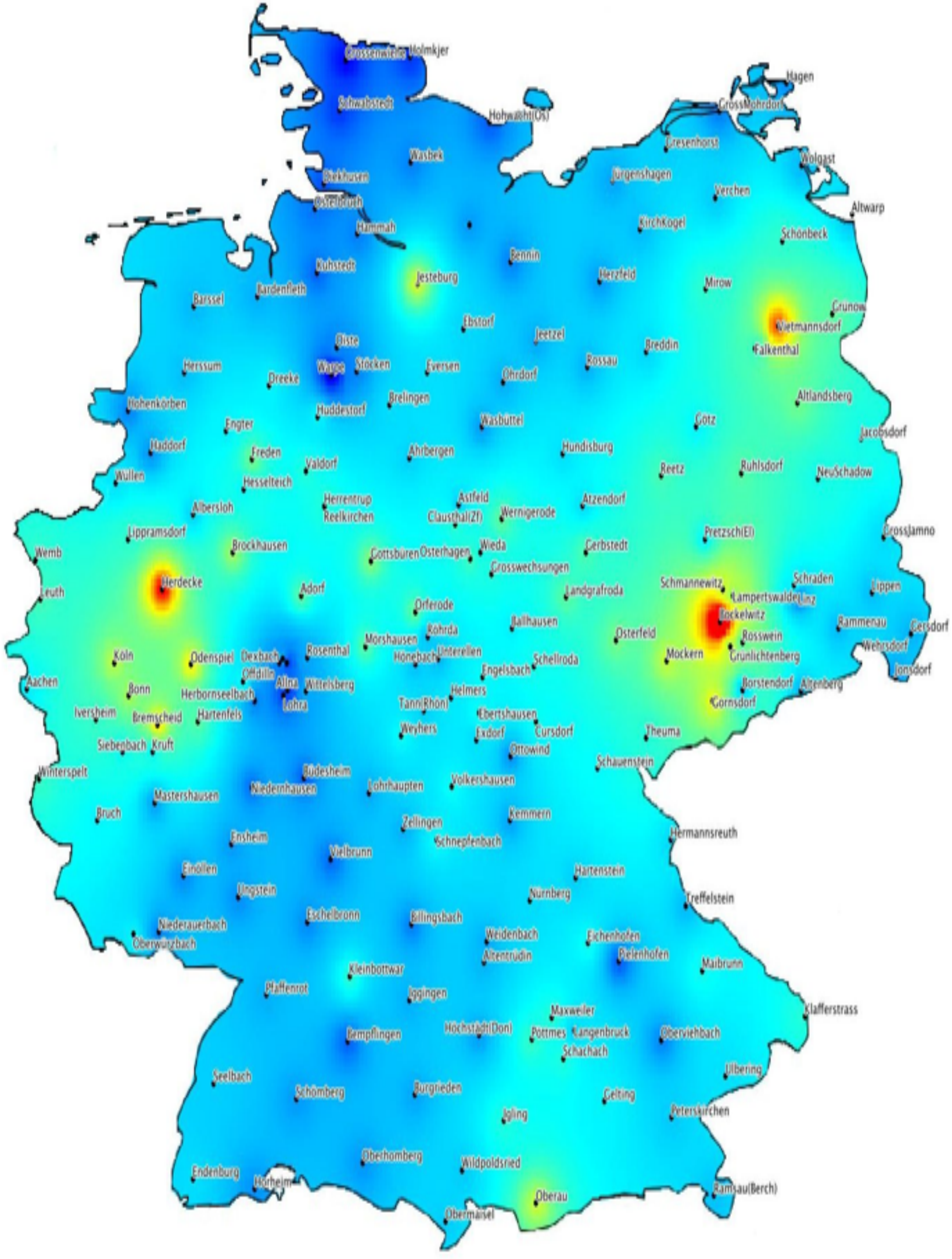
Vowels are ‘strong' in an area in which consonants are ‘weakening’



Looking at Regularity

- Idea: do not look at the actual forms, but at the regularity of the changes between locations
- Practice: compare all neighbouring locations, and establish how regular the correspondences are
- Aggregate all irregularity per location:
  - blue: regular correspondences to neighbours
  - yellow/red: irregular correspondences to neighbours





Grossenwiehe
Holmkjer
Hagen
Schwabstedt
GrossMohrdorf
Hohwacht(Os)
Wasbek
Gresenhorst
Wolgast
Jurgenshagen
Bekhusen
Verchen
Altwarp
Hammah
KirchKogel
Schönbeck
Bennin
Kuhstedt
Jesteburg
Herzfeld
Mirow
Barssel
Bardenfleth
Grunow
Ebstorf
Wietmannsdorf
Jeetzel
Falkenthal
Oiste
Breddin
Herssum
Stöcken
Eversen
Rossau
Dreeke
Ohrdorf
Altlandsberg
Hohenkörben
Brelingen
Huddestorf
Wasbüttel
Götz
Engter
Jacobsdorf
Haddorf
Freden
Hundisburg
Ahrbergen
Ruhlsdorf
Valdorf
Reetz
NeuSchadow
Wullen
Hesselteich
Astfeld
Herrentrup
Atzendorf
Albersloh
Reelkirchen
Wernigerode
Clausthal(Zf)
Grossjamno
Pretzsch(El)
Lippramsdorf
Brockhausen
Wieda
Gerbstedt
Cottsbüren
Osterhagen
Wemb
Crosswechsungen
Schraden
Herdecke
Schmannewitz
Lippen
Landgrafroda
Adorf
Lampertswalde
Leuth
Linz
Orferode
Cockelwitz
Rammenau
Ballhausen
Rohrda
Gersdorf
Osterfeld
Morshausen
Rosswein
Wehrsdorf
Rosenthal
Unterellen
Köln
Odenspiel
Dexbach
Mockern
Grunlichtenberg
Hönebach
Schellroda
Jonsdorf
Engelsbach
Aachen
Borstendorf
Altenberg
Allna
Wittelsberg
Bonn
Herbornseelbach
Helmers
Lohra
Gornsdorf
Tann(Rhön)
Ebertshausen
Iversheim
Bremscheid
Hartenfels
Weyhers
Exdorf
Cursdorf
Theuma
Siebenbach
Kruft
Ottowind
Schauenstein
Winterspelt
Büdesheim
Lohrhaupten
Volkershausen
Niedernhausen
Mastershausen
Kemmern
Bruch
Zellingen
Hermannsreuth
Ensheim
Schnepfenbach
Vielbrunn
Hartenstein
Eindlen
Nürnberg
Ungstein
Treffelstein
Eschelbronn
Billingsbach
Niederauerbach
Weidenbach
Eichenhofen
Oberwürzbach
Altentrüdin
Pielenhofen
Maibrunn
Kleinbottwar
Pfaffenrot
Iggingen
Maxweiler
Klafferstrass
Hochstadt(Don)
Pöttmes
Langenbruck
Oberviehbach
Bempflingen
Schachach
Ulbering
Seelbach
Schömberg
Burgrieden
Gelting
Peterskirchen
Igling
Oberhomberg
Wildpoldsried
Endenburg
Horheim
Ramsau(Berch)
Oberau
Obermaisel

Working with Survey Data

- Data is never perfectly organised (errors, disagreements about coding)
- Instead of complaining: correct errors, or make your own version of the coding !
- The git-approach: share the data, let others make a fork/clone, possible merging the variants later

The git-approach

- git is a versioning control system: keeping track of who changed what
- You can run it privately on your own computer, or use services like github or bitbucket
- Text-based: so use simple formats! (CSV in UTF-8, NFC, LF)