



Handbuch

Version 1.0 [deutsch]
22. September 2014

“Neue Dialektometrie mit Methoden der stochastischen Bildanalyse”
DFG-finanziertes Projekt (2008-2014)

Stephan Elspaß
Henrik Hassfeld
Werner König
Daniel Meschenmoser
Simon Pickl
Simon Pröll
Jonas Rumpf
Volker Schmidt
Aaron Spettl
Evgeny Spodarev
Julius Vogelbacher
Raphael Wimmer

<http://www.geoling.net/>

Inhalt

Inhalt	2
Hintergrund / Personen.....	2
Lizenz und Nutzungsbedingungen	2
I. Grundlagen	4
I.1 Was ist <i>GeoLing</i> und wozu kann man es nutzen?	4
I.2 Download und Installation.....	4
I.3 <i>GeoLing</i> starten.....	4
I.4 Eigene Daten einlesen.....	5
I.5 Das Programm anpassen und erweitern.....	15
II. Programmfunktionen	16
II.1 Kategorien	17
II.2 Dichteschätzung	17
II.3 Gruppen.....	22
II.4 Export	22
II.5 Faktorenanalyse.....	23
II.6 Clusteranalyse	26
III. Literatur.....	27

Hintergrund / Personen

GeoLing ist – ebenso wie dieses Handbuch – das Resultat des Forschungsprojekts *Neue Dialektometrie mit Methoden der stochastischen Bildanalyse*, das zwischen 2008 und 2014 durch die Deutsche Forschungsgemeinschaft (DFG) unterstützt wurde. Alle beteiligten Personen waren den folgenden drei Institutionen zugeordnet:

- *Institut für Stochastik* (Universität Ulm)
- *Lehrstuhl für Deutsche Sprachwissenschaft* (Universität Augsburg)
- *Germanistische Sprachwissenschaft* (Universität Salzburg)

Lizenz und Nutzungsbedingungen

Diese Software steht unter der **GNU General Public License v 3.0** (vom 29. Juni 2007). Der volle Text ist unter <https://www.gnu.org/licenses/gpl-3.0> einsehbar.

Trotz sorgfältiger Arbeit und ausführlicher Tests können wir keine Garantie für das Endprodukt gewähren; die Nutzung erfolgt daher auf eigene Verantwortung.

I. Grundlagen

I.1 Was ist *GeoLing* und wozu kann man es nutzen?

Kurz gesagt ist *GeoLing* ein einfach zu bedienendes Programm für statistische Analysen räumlich aufgelöster Daten: Material aus Spracherhebungen kann (mittels Dichteschätzung) in der Form geglätteter Flächenkarten dargestellt werden; wiederkehrende Strukturen in den Daten können ebenso aufgedeckt werden wie Gruppen aus einzelnen Karten, die sich entsprechen.

Das Programm wurde zwar primär für die Analyse sprachlicher Daten entwickelt, eine Anwendung auf beliebige andere Arten räumlicher Daten ist aber ebenfalls möglich.

I.2 Download und Installation

Das Programm ist in Java geschrieben und daher auf den verschiedensten Plattformen lauffähig. Die Java-Version des Rechners sollte nach Möglichkeit aktuell sein; entsprechende Updates lassen sich auf <http://www.java.com> beziehen. Die aktuellste Fassung von *GeoLing* sowie dieses Handbuchs stehen auf <http://www.geoling.net/> als Archivdatei (*geoling.zip*) zum Download bereit. *GeoLing* benötigt mindestens 2 GB Arbeitsspeicher; für flüssiges Arbeiten empfehlen wir außerdem einen dual-core-Prozessor.

GeoLing erfordert keine eigene Installation; es reicht, das Archiv *geoling.zip* nach dem Download zu entkomprimieren. Wichtig ist, dass *GeoLing* in dem Ordner, in dem es sich befindet, volle Schreibrechte hat. Das sollte im Normalfall z.B. im eigenen Benutzerverzeichnis der Standard sein.

Teil der Archivdatei ist auch ein Beispieldatensatz, das Material des *Sprachatlas von Bayerisch-Schwaben* (SBS).¹ Dieses Material (das auch hier im Handbuch zu Anschauungszwecken genutzt wird) war auch der Datensatz, an dem alle Methoden entwickelt und getestet wurden.

I.3 *GeoLing* starten

Das Programm startet durch Ausführen der Datei *geoling.jar*.

Die Dateien *start-geoling-windows.vbs* und *start-geoling-linux.sh* sind Batchdateien für den Start unter Windows respektive Linux/Mac, die es ermöglichen, *GeoLing* mehr Arbeitsspeicher zuzuweisen. Da manche Programmfunktionen (wie etwa das Clustern von Karten, siehe II.6) relativ speicherintensiv sind, ist dies grundsätzlich zu empfehlen. Die Batchdateien können einfach mit einem simplen Texteditor angepasst werden (Abbildung 1). Der Wert in der hervorgehobenen Zeile kann angepasst werden. Der voreingestellte Standardwert beträgt 1400 MB. Auf 32-bit-Systemen sind je nach Systemkonfiguration maximal 1500 MB möglich, auf 64-bit-Systemen sind auch höhere Werte möglich.

¹ Der *Sprachatlas von Bayerisch-Schwaben* beruht auf Erhebungen, die in den 1980er Jahren im Übergangsgebiet der süddeutschen Regierungsbezirke Schwaben, Oberbayern und Mittelfranken durchgeführt wurden (vgl. SBS 1 für weiterführende Informationen).

```

1 Dim objShell
2 Set objShell = WScript.CreateObject("WScript.Shell")
3
4 ' set working directory
5 Set objFSO = CreateObject("Scripting.FileSystemObject")
6 Set objFile = objFSO.GetFile(WScript.ScriptFullName)
7 objShell.CurrentDirectory = objFSO.GetParentFolderName(objFile)
8
9 ' options for Java virtual machine:
10 ' set maximum heap size to allow Geoling to use more memory:
11 ' - default max heap size:
12 ' (empty value, i.e., no parameter, means that Java auto-selects a value
13 ' depending on your total physical memory installed, usually 1/4)
14 Const MAX_HEAP = ""
15 ' - user-defined max heap size:
16 ' (note that at most ~1400m to ~1600m is possible on 32-bit machines)
17 Const MAX_HEAP = "-Xmx1400m"
18
19 ' start application
20 objShell.Run("javaw " & MAX_HEAP & " -splash:geoling-splash-screen.png -jar geoling.jar")
21
22 Set objShell = Nothing
23

```

Ln 18:23 Col 1 Sel 30 844 Bytes ANSI CR+LF INS VBScript

Abbildung 1: Batchdatei zum *GeoLing*-Start mit mehr Arbeitsspeicher (Windows-Variante).

I.4 Eigene Daten einlesen

GeoLing greift auf Daten in Form einer SQL-Datenbank zurück. Als Nutzer hat man die Wahl, entweder selbst eine Datenbank, die auf einem Server läuft, aufzusetzen oder die Erstellung *GeoLing* zu überlassen. Beide Optionen werden im Folgenden vorgestellt.

In beiden Fällen beginnt man damit, im „Datenbank“-Dialogfeld nach dem Start „Neue Datenbank erstellen“ auszuwählen (Abbildung 2).

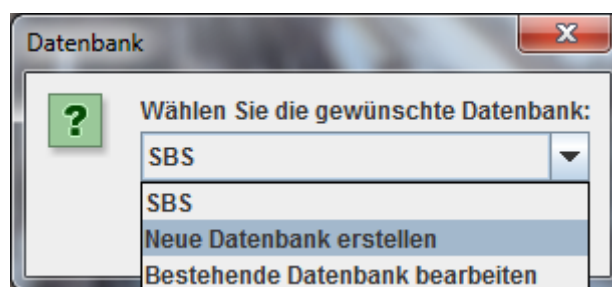


Abbildung 2: Neue Datenbank erstellen.

Im anschließenden Dialogfeld (Abbildung 3) kann man zwischen beiden Optionen wählen: einer eigenständigen SQL-Datenbank auf einem Server oder der in *GeoLing* implementierten SQLite-Variante. Die erste Option hat die Vorteile schnellerer Zugriffs- und Verarbeitungszeiten sowie die Möglichkeit zur Nutzung in einem Netzwerk, ist aber eher für Nutzer zu empfehlen, die bereits Erfahrungen mit Datenbanksystemen aufweisen. Die zweite Option ist für Einsteiger einfacher zu nutzen, geht dafür jedoch mit leichten Geschwindigkeitseinbußen einher.

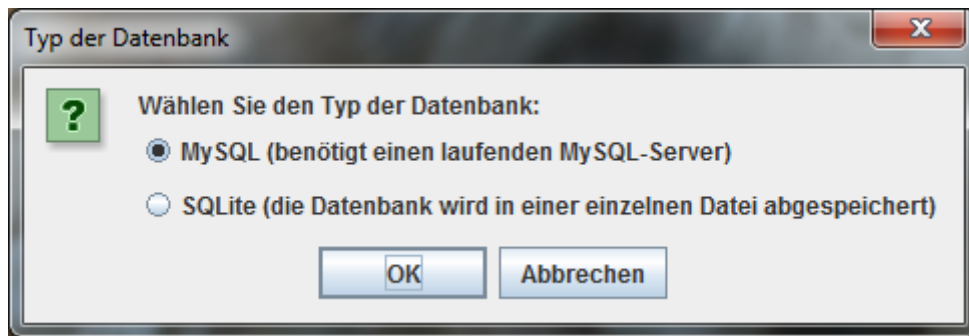


Abbildung 3: Auswahl des Datenbanktyps.

⇒ Erste Option: selbst eine SQL-Datenbank aufsetzen

Wir beginnen mit den nötigen Schritten, um einen Server aufzusetzen. Sollten Sie die zweite Option vorziehen, dann überspringen Sie diesen Abschnitt bitte. (Sollten Sie sich unsicher sein, wählen Sie besser die zweite Option – man kann problemfrei später noch zur ersten Option zurückwechseln.)

a) *MySQL Workbench* installieren

Für Windows-Nutzer ist das Programm *MySQL Workbench* zum SQL-Datenbankmanagement zu empfehlen, das unter <http://dev.mysql.com/downloads/mysql/> frei erhältlich ist.

b) Neue Verbindung aufsetzen

Nach der Installation bitte MySQL Workbench starten. Ein Klick auf das Plus rechts neben „MySQL Connections“ bringt ein Fenster wie in Abbildung 4 hervor, die entsprechend dem Beispiel ausgefüllt werden sollten. Der „Connection Name“ kann frei gewählt werden.

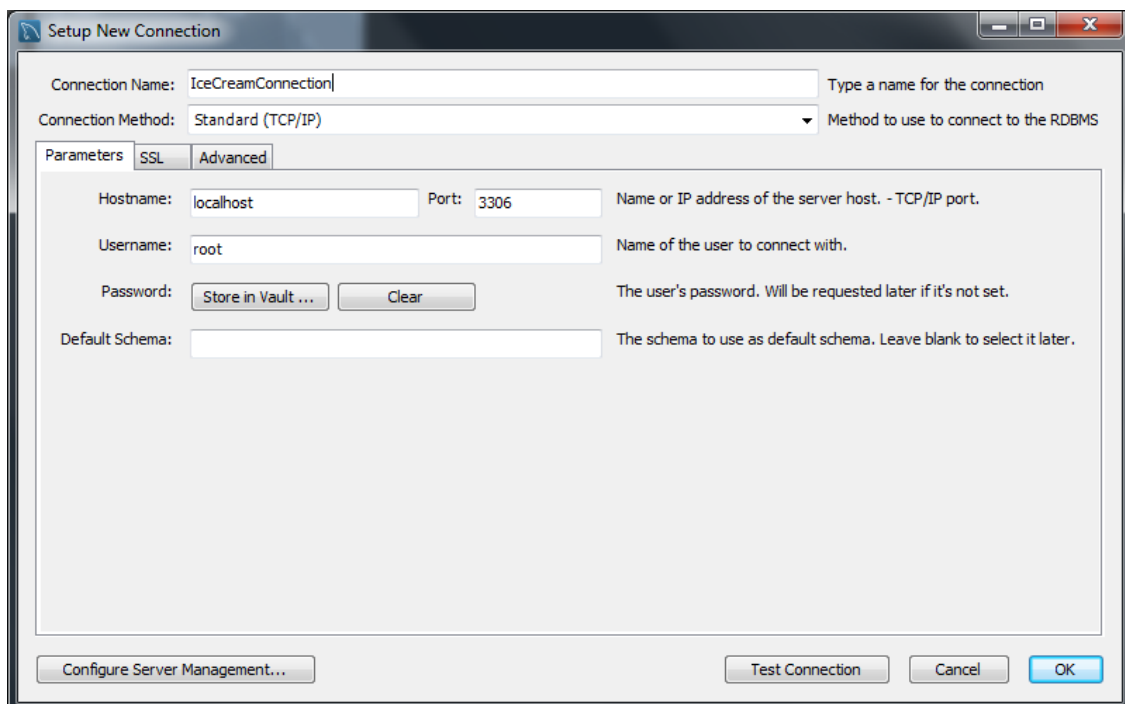


Abbildung 4: Neue Verbindung aufsetzen.

Per Klick auf „Test Connection“ last sich prüfen, ob alle Werte korrekt sind. Nach einem Klick auf „OK“ sollte die neue Verbindung in MySQL Workbench sichtbar sein.

c) Neues Schema erzeugen

Nach einem Klick auf diese neue Verbindung kann man links die Navigationsleiste sehen; bitte klicken Sie oberhalb auf das Symbol für „Create a schema in the connected server“, in Abbildung 5 gelb markiert.

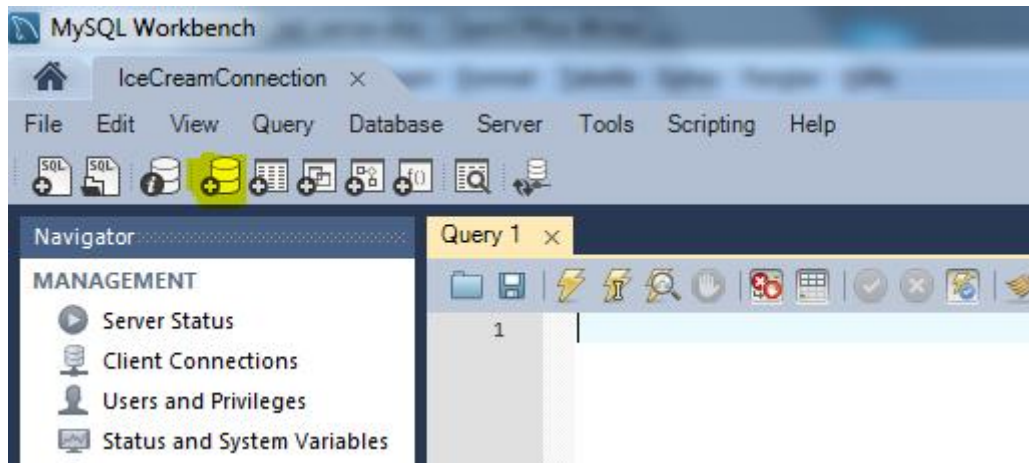


Abbildung 5: Neues Schema erzeugen.

Bitte einen Namen für das Schema eingeben und zuerst „Apply“, dann „Data Import/Restore“ klicken. Dort „Import from Self-Contained File“ wählen, den Pfad zu *dialectometry.sql* wählen und das Schema als „Default Schema to be Imported to“ setzen. Nach dem Import (über „Start Import“) sollten bei einem Klick auf das neue Schema folgende Tabellen vorhanden sein (in Klammern jeweils die dazugehörigen Spalten).

- bandwidths (id, map_id, weights_identification, kernel_identification, distance_identification, estimator_identification, bandwidth)
- border_coordinates (id, border_id, order_index, latitude, longitude)
- borders (id, name)
- categories (id, parent_id, lft, rgt, name)
- categories_maps (id, category_id, map_id)
- configuration_options (id, name, value)
- distances (id, name, type, identification)
- groups (id, name)
- groups_maps (id, group_id, map_id)
- informants (id, location_id, name)
- interview_answers (id, interviewer_id, informant_id, variant_id)
- interviewers (id, name)
- levels (id, name)
- locations (id, name, code, latitude, longitude)
- location_distances (id, distance_id, location_id1, location_id2, distance)
- maps (id, name)
- tags (id, parent_type, parent_id, name, value)
- variants (id, map_id, name)
- variants_mappings (id, variant_id, level_id, to_variant_id)

Falls alle 19 Tabellen vorhanden sind, waren die Installation des lokalen Servers und die Erstellung der Datenbank erfolgreich. Der MySQL-Server muss immer laufen, wenn *GeoLing* ausgeführt werden soll.

⇒ **Zweite Option: *GeoLings* internes System nutzen**

Nach dem Programmstart wählen Sie „Neue Datenbank erstellen“ (Abbildung 2) und darauf „SQLite“ als Datenbanktyp (Abbildung 3). Anschließend werden Speicherort und Name der neuen Datenbank sowie der Ordner, in dem die Ergebnisse ausgegeben werden, festgelegt. Der Name folgt dem Schema DATENBANKNAME.db (mit Dateiendung, vgl. Abbildung 6).

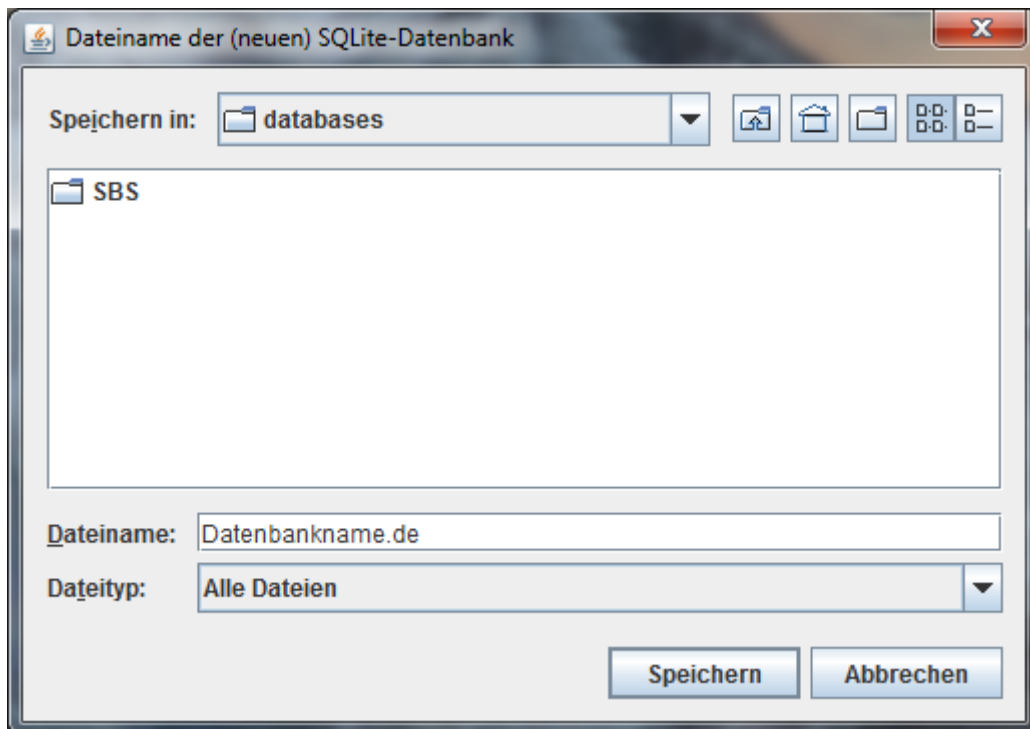


Abbildung 6: Speicherort und Name der Datenbankdatei.

⇒ **Datenimport**

Nun kann der Import eigener Daten erfolgen; dazu sollten die im Folgenden genauer erläuterten Einzeldateien in einem gemeinsamen Ordner liegen, der dem Programm über den Button „Eingabe-Ordner wählen“ zugewiesen wird (Abbildung 7).

Rechts daneben lässt sich auch die Kodierung der Dateien, die man importieren möchte, einstellen: Zur Auswahl stehen UTF-8 und ISO-8859-1 (Abbildung 8).

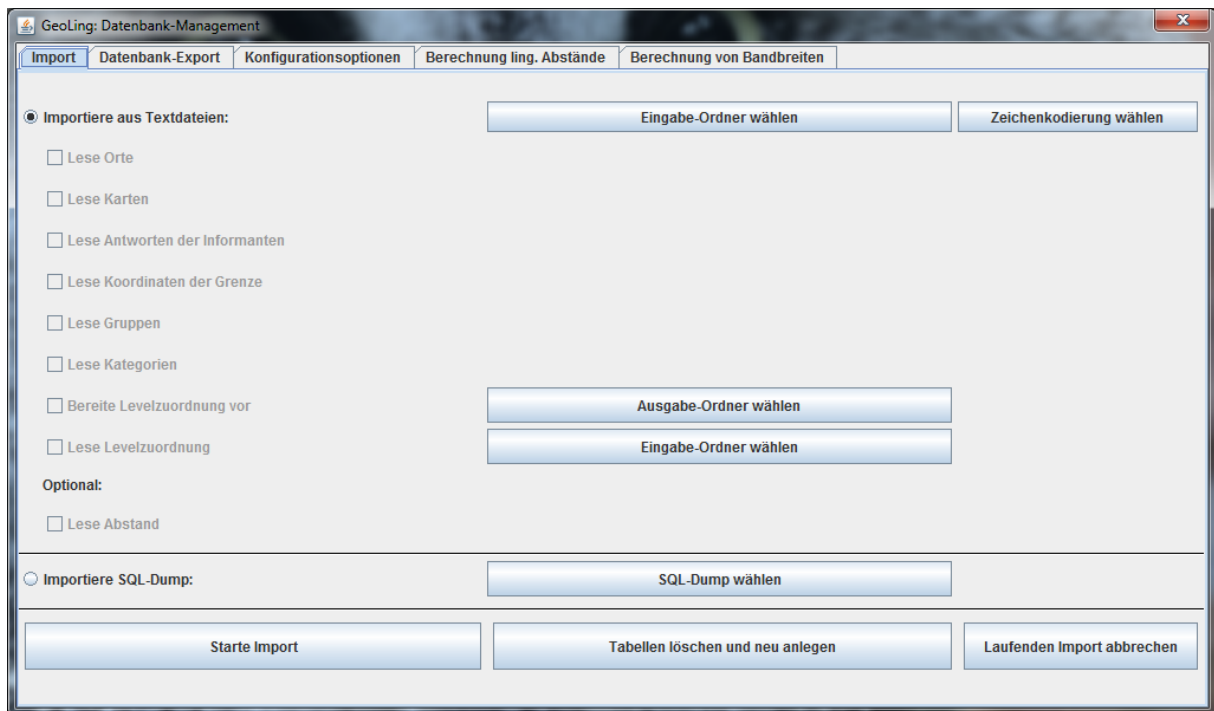


Abbildung 7: Import-Fenster in GeoLing.



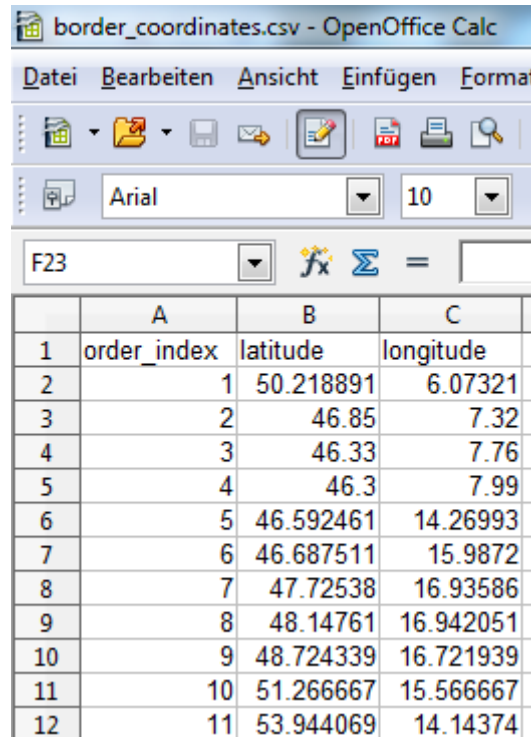
Abbildung 8: Auswahl der Kodierung der Daten.

Anmerkung: Grundsätzlich sollte beachtet werden, dass *GeoLing* als Dezimaltrennzeichen einen Punkt (".") erwartet, kein Komma. Ein Tausendertrennzeichen ist nicht vorgesehen. Die Tabellen werden alle als csv-Dateien gehandhabt, hier ist ein Semikolon (";") das Zeichen, das die einzelnen Felder voneinander trennt. In den Datensätzen, die importiert werden sollen, darf also nirgends ein ";" vorhanden sein. Obwohl zwei verschiedene Zeichenkodierungen möglich sind (ISO-8859-1 und UTF-8), empfiehlt es sich generell, auf Sonderzeichen (wie IPA) möglichst zu verzichten bzw. sie durch Kodierungen, die auf Portabilität optimiert sind (etwa SAMPA), zu ersetzen.

Die erforderlichen Dateien im einzelnen:

border_coordinates.csv

Über diese Tabelle wird die Außengrenze der Grundkarte festgelegt. Die erste Spalte gibt die Reihenfolge der Außenpunkte an, die nachfolgenden beiden Spalten geografische Breite und Länge (Abbildung 9).²



	A	B	C
1	order_index	latitude	longitude
2	1	50.218891	6.07321
3	2	46.85	7.32
4	3	46.33	7.76
5	4	46.3	7.99
6	5	46.592461	14.26993
7	6	46.687511	15.9872
8	7	47.72538	16.93586
9	8	48.14761	16.942051
10	9	48.724339	16.721939
11	10	51.266667	15.566667
12	11	53.944069	14.14374

Abbildung 9: Tabelle für Außengrenze.

Anmerkung: Es ist möglich, mehr als eine Außengrenze anzulegen, daher wird man beim Import jeweils gebeten, der Grenze einen Namen zu geben.

locations.csv

Diese Tabelle beinhaltet die Namen und Koordinaten der Belegorte. Die erste Spalte führt den Namen an, die zweite eine Identifikationskennung (Ortsnummer, Kürzel, etc. – dieser Eintrag kann später direkt in der Karte angezeigt werden), die dritte und vierte geografische Breite und Länge (Abbildung 10). Weder Namen noch auch Koordinaten dürfen mehr als einmal vorkommen.

² In den Beispielabbildungen werden die Tabellen mit *OpenOffice Calc* und die Textdateien mit dem Windows-internen Editor angezeigt, das spezifische Bearbeitungsprogramm spielt aber im Prinzip keine Rolle. Bei der Verwendung von *Microsoft Excel* sollte jedoch darauf geachtet werden, dass die dort oftmals auftretenden automatischen Formatierungen von Zellen (etwa weil ein scheinbares Datum erkannt wurde oder ein "." als Tausendertrennzeichen interpretiert wird) eventuell zu unerwünschten Nebeneffekten führen können.

	A	B	C	D	E
1	name	code	latitude	longitude	
2	Aitrang	1	47.8196541	10.536059	
3	Altenmünster	2	48.4589825	10.5513457	
4	Altheim	3	48.5788648	10.0105589	
5	Asbach	4	48.6941038	10.81004	
6	Auchsesheim	5	48.6908262	10.7806048	
7	Augsburg	6	48.3714407	10.8982552	
8	Biburg	7	48.374965	10.757741	
9	Bobingen	8	48.2746049	10.7738392	
10	Bühl	9	48.4185225	10.1894515	
11	Buxheim	10	47.9978929	10.1336661	

Abbildung 10: Ortsliste.

maps.csv

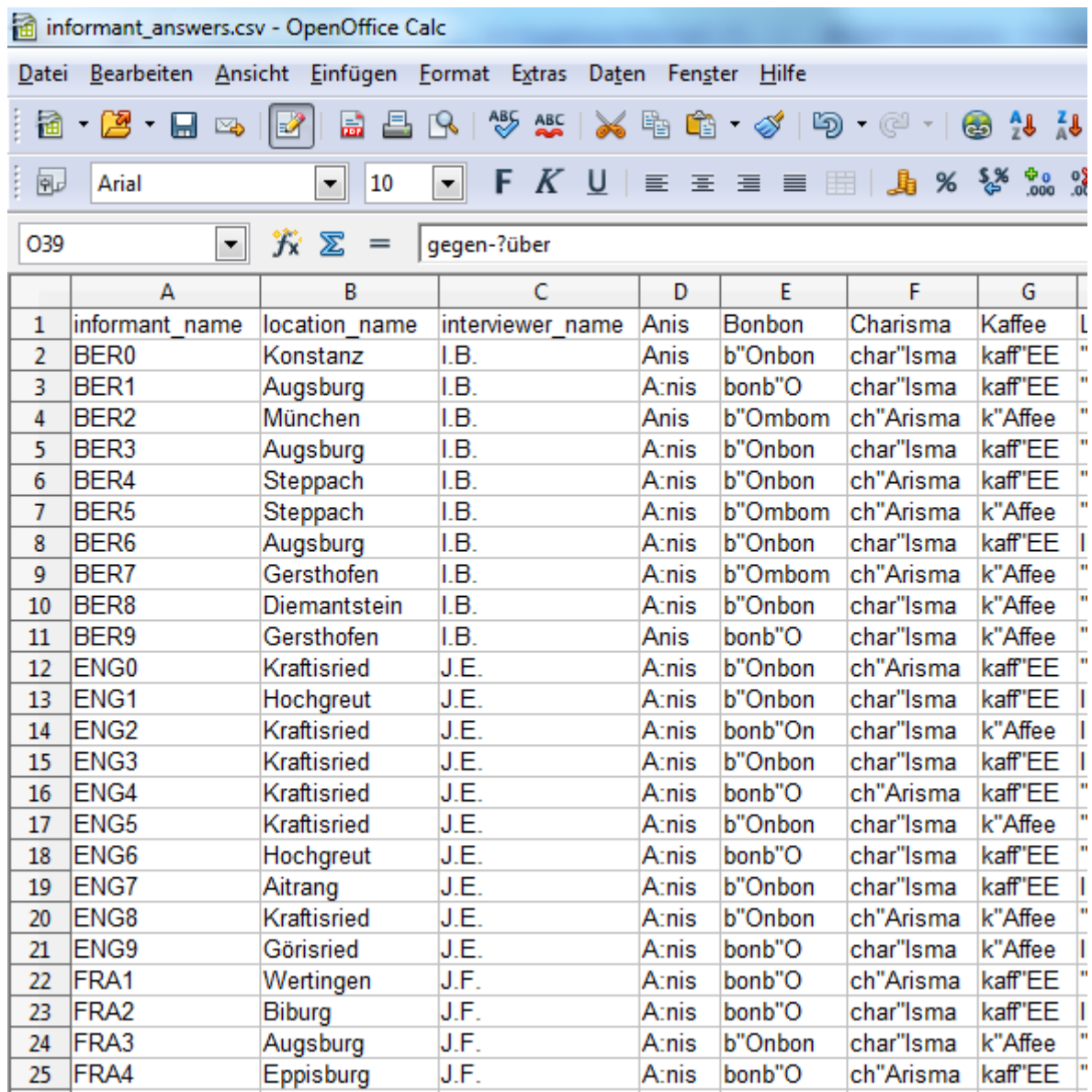
Hier sind die Benennungen der einzelnen Karten (bzw. Variablen) aufgeführt. Dazu bedarf es lediglich einer einzelnen Spalte (Abbildung 11). Auch hier darf kein Name doppelt besetzt sein.

	A
1	name
2	Anis
3	Bonbon
4	Charisma
5	Kaffee
6	LKW
7	Make-up
8	Motor
9	Oblate
10	Verein
11	real
12	Oase
13	gegenüber
14	Beamter
15	Ruine

Abbildung 11: Kartennamen.

informant_answers.csv

Hier werden die Antworten in einer Tabelle der Struktur *Informant* (Zeilen) x *Karte* (Spalten) = *Beleg* (Felder) gebündelt: Die erste Spalte enthält den Namen oder die Sigle des Informanten, die zweite den Ort (Schreibung wie in locations.csv), die dritte Spalte den Explorator (oder ähnliche belegabhängige Zusatzinformationen). Ab der vierten Spalte beginnen die Belege, jede Spalte ist betitelt mit einem der Einträge in maps.csv (also den Karten- bzw. Variablennamen, vgl. Abbildung 12). Treten mehrere Belege von nur einem Informanten an einem Ort auf, können sie gemeinsam in ein Feld eingetragen und durch das Zeichen "|" voneinander abgetrennt werden.



	A	B	C	D	E	F	G
1	informant_name	location_name	interviewer_name	Anis	Bonbon	Charisma	Kaffee
2	BER0	Konstanz	I.B.	Anis	b"Onbon	char"Isma	kaff"EE
3	BER1	Augsburg	I.B.	A:nis	bonb"O	char"Isma	kaff"EE
4	BER2	München	I.B.	Anis	b"Ombom	ch"Arisma	k"Affee
5	BER3	Augsburg	I.B.	A:nis	b"Onbon	char"Isma	kaff"EE
6	BER4	Steppach	I.B.	A:nis	b"Onbon	ch"Arisma	kaff"EE
7	BER5	Steppach	I.B.	A:nis	b"Ombom	ch"Arisma	k"Affee
8	BER6	Augsburg	I.B.	A:nis	b"Onbon	char"Isma	kaff"EE
9	BER7	Gersthofen	I.B.	A:nis	b"Ombom	ch"Arisma	k"Affee
10	BER8	Diemantstein	I.B.	A:nis	b"Onbon	char"Isma	k"Affee
11	BER9	Gersthofen	I.B.	Anis	bonb"O	char"Isma	k"Affee
12	ENG0	Kraftisried	J.E.	A:nis	b"Onbon	ch"Arisma	kaff"EE
13	ENG1	Hochgreut	J.E.	A:nis	b"Onbon	char"Isma	kaff"EE
14	ENG2	Kraftisried	J.E.	A:nis	bonb"On	char"Isma	k"Affee
15	ENG3	Kraftisried	J.E.	A:nis	b"Onbon	char"Isma	kaff"EE
16	ENG4	Kraftisried	J.E.	A:nis	bonb"O	ch"Arisma	kaff"EE
17	ENG5	Kraftisried	J.E.	A:nis	b"Onbon	char"Isma	k"Affee
18	ENG6	Hochgreut	J.E.	A:nis	bonb"O	char"Isma	kaff"EE
19	ENG7	Aitrang	J.E.	A:nis	b"Onbon	char"Isma	kaff"EE
20	ENG8	Kraftisried	J.E.	A:nis	b"Onbon	ch"Arisma	k"Affee
21	ENG9	Görisried	J.E.	A:nis	bonb"O	char"Isma	k"Affee
22	FRA1	Wertingen	J.F.	A:nis	bonb"O	ch"Arisma	kaff"EE
23	FRA2	Biburg	J.F.	A:nis	bonb"O	char"Isma	kaff"EE
24	FRA3	Augsburg	J.F.	A:nis	b"Onbon	char"Isma	k"Affee
25	FRA4	Eppisburg	J.F.	A:nis	bonb"O	ch"Arisma	kaff"EE

Abbildung 12: Antworttabelle.

categories.txt

Mittels dieser Textdatei wird eine Hierarchie der Kartenthemen (z.B. in Bände / Themen / Unterthemen / etc.) erstellt. Jede Zeile enthält eine hierarchisch übergeordnete Ebene und – nach einem Doppelpunkt – die untergeordneten Kategorien, jeweils mit Strichpunkten

voneinander abgetrennt. Untergeordnete Kategorien können selbst wieder als übergeordnete dienen (Abbildung 13). Als Resultat erhält man später in der Benutzeroberfläche die Baumstruktur, die im Kategorien-Reiter (siehe II.1) angezeigt wird.

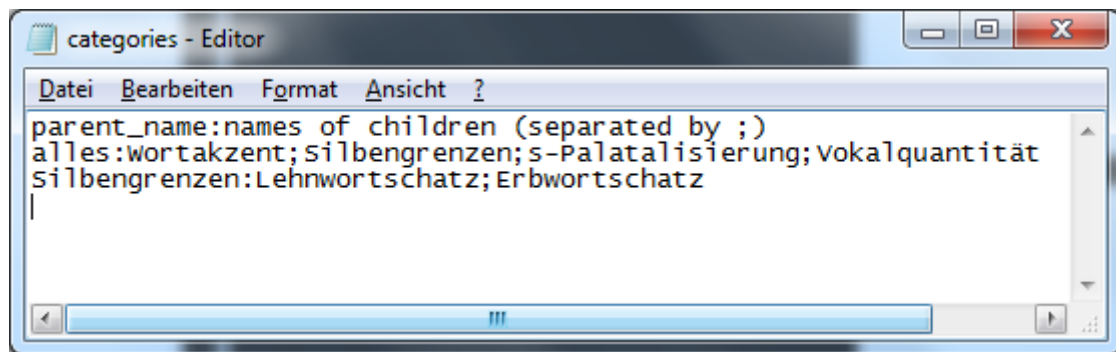


Abbildung 13: Kategorien und Unterkategorien.

categories_maps.txt

Jede Karte muss einer Kategorie zugewiesen sein, wenn man sie später im Kategorien-Reiter anwählen will. Diese Datei regelt die Zuweisung, ebenfalls nach dem oben angeführten Prinzip: Pro Zeile existiert eine Kategorie, nach einem Doppelpunkt folgen die Karten, die ihr zugewiesen werden sollen, jeweils durch einen Strichpunkt voneinander getrennt (Abbildung 14).

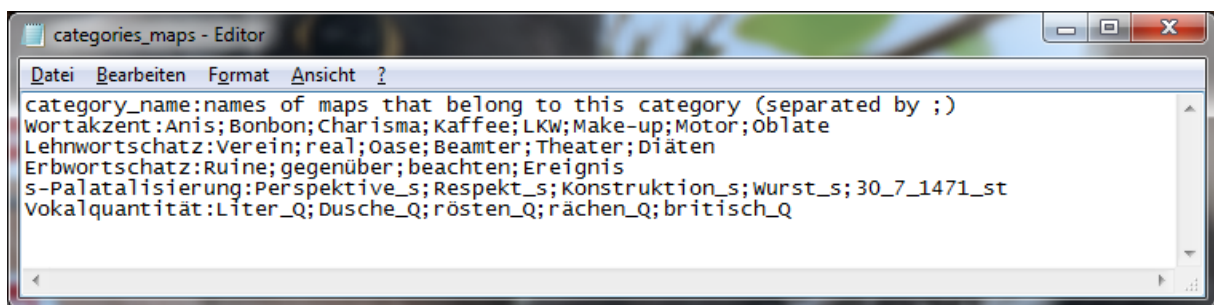


Abbildung 14: Zuweisung von Karten zu Kategorien.

groups_maps.txt

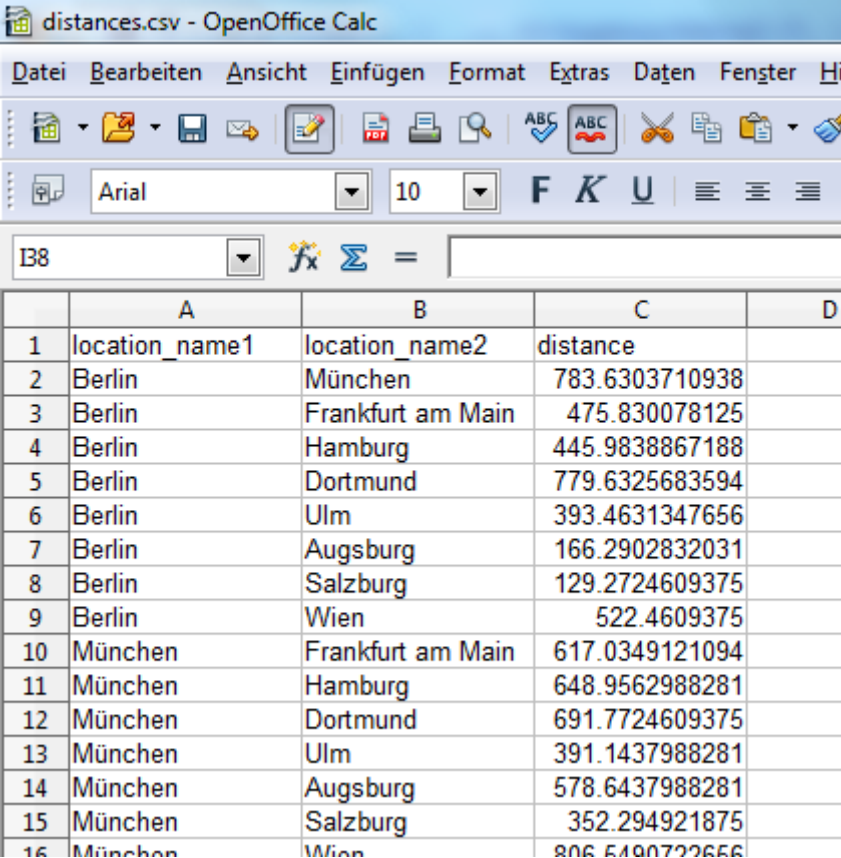
Diese Textdatei dient zur Bündelung einzelner Karten zu Gruppen (zur Erleichterung der Auswertung). Diese Funktion, die nach demselben Prinzip wie die anderen beiden Textdateien zu behandeln ist, ist im Importprozess optional, da Gruppen auch nachträglich per Oberfläche erstellt werden können (siehe II.3).

Wichtig: Bei allen drei Textdateien muss jede Zeile – auch die letzte! – mit einem Enterzeichen abgeschlossen werden, nur so wird sie vom Importsubprogramm korrekt interpretiert.

distances.csv

GeoLing errechnet anhand der Koordinaten in der Ortstabelle die geografischen Distanzen zwischen den Orten selbst – man kann aber optional mittels der Datei *distances.csv* eigene Distanzmaße (etwa die Reisezeit) einpflegen. Diese müssen in einer Tabelle, in der jedes Ortspaar

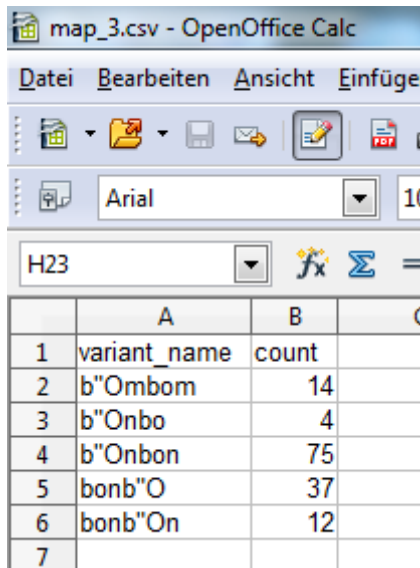
einen Distanzwert zugewiesen bekommt, eingespeist werden. Abbildung 15 zeigt ein Beispiel der notwendigen Formatierung.



	A	B	C	D
1	location_name1	location_name2	distance	
2	Berlin	München	783.6303710938	
3	Berlin	Frankfurt am Main	475.830078125	
4	Berlin	Hamburg	445.9838867188	
5	Berlin	Dortmund	779.6325683594	
6	Berlin	Ulm	393.4631347656	
7	Berlin	Augsburg	166.2902832031	
8	Berlin	Salzburg	129.2724609375	
9	Berlin	Wien	522.4609375	
10	München	Frankfurt am Main	617.0349121094	
11	München	Hamburg	648.9562988281	
12	München	Dortmund	691.7724609375	
13	München	Ulm	391.1437988281	
14	München	Augsburg	578.6437988281	
15	München	Salzburg	352.294921875	
16	München	Wien	806.5190722656	

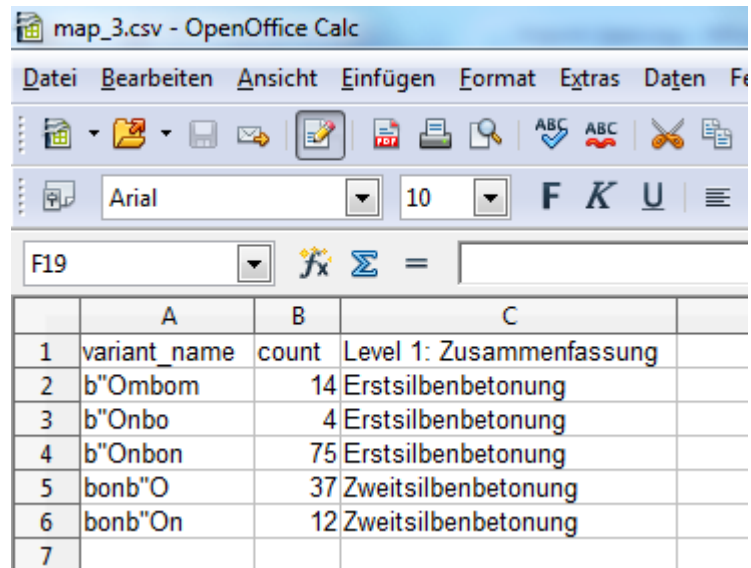
Abbildung 15: Distanzen zwischen Ortspaaren.

Die Funktion „Bereite Levelzuordnung vor“ im Importdialog erstellt für jede Karte (in einem selbstgewählten Ordner) eine Tabelle, die die einzelnen Varianten und ihre Anzahl auflistet (Abbildung 16). Hier können manuell Zusammenfassungen der Belege vorgenommen werden: Zum einen können so z.B. vereinzelt Fehlschreibungen der Varianten zusammengeführt, zum anderen aber auch Abstraktionsebenen über die rohen Belege gebildet werden. Sind die Originalbelege z.B. eng phonetisch transkribiert, enthalten aber auch lexikalische Unterschiede, so kann man hier ein Abstraktionslevel erzeugen, das die lautliche Variation ignoriert und eine rein lexikalische Karte ermöglicht. Jede zusätzliche Spalte kann eine individuelle Zusammenfassung beinhalten; die erste Zeile legt den Namen für das Abstraktionsniveau fest (Abbildung 17). Sollte man keine Zusammenfassung der Rohdaten benötigen, reicht es, die erste Spalte zu kopieren und als dritte Spalte einzufügen.



	A	B
1	variant_name	count
2	b"Ombom	14
3	b"Onbo	4
4	b"Onbon	75
5	bonb"O	37
6	bonb"On	12
7		

Abbildung 16: Leere Zuweisungstabelle.



	A	B	C
1	variant_name	count	Level 1: Zusammenfassung
2	b"Ombom	14	Erstsilbenbetonung
3	b"Onbo	4	Erstsilbenbetonung
4	b"Onbon	75	Erstsilbenbetonung
5	bonb"O	37	Zweitsilbenbetonung
6	bonb"On	12	Zweitsilbenbetonung
7			

Abbildung 17: Zuweisungstabelle mit einem Abstraktionslevel.

Diese Zuweisungen werden im Importdialog mittels „Lese Levelzuordnung“ wieder eingelesen.

Sollte man die Datenbank zu einem späteren Zeitpunkt überarbeiten oder erweitern wollen, ist der Zugriff auf den Einleseprozess jederzeit über „Bestehende Datenbank bearbeiten“ beim Programmstart möglich (Abbildung 18).

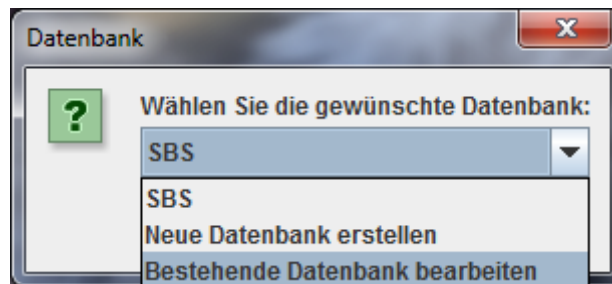


Abbildung 18: Dialogfeld zur Datenbankauswahl.

I.5 Das Programm anpassen und erweitern

Wie es die GPL 3-Lizenz vorsieht, steht auch der vollständige Quellcode von *GeoLing* zur Verfügung. Damit ist es sowohl möglich, eigene Programmbestandteile hinzuzufügen, als auch die bestehenden Teile nach Bedarf anzupassen.

II. Programmfunktionen

Nach dem Programmstart erscheint die Auswahlmöglichkeit, mit welcher Datenbank gearbeitet werden soll (Abbildung 19).

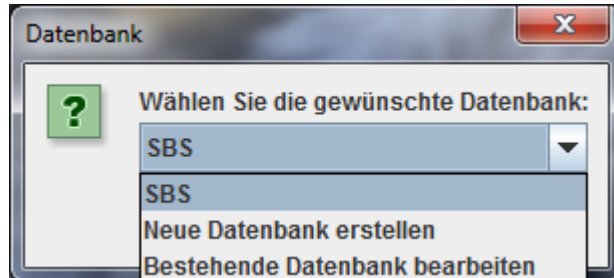


Abbildung 19: Auswahl der Datenbank.

Die Datenbank des SBS (Sprachatlas von Bayerisch-Schwaben) ist zu Demonstrationszwecken bereits enthalten. Von diesem Auswahlfeld aus ist es auch möglich, eine neue Datenbank zu erstellen oder eine bestehende zu ändern (siehe I.4, *Eigene Daten einlesen*).

Wählt man eine Datenbank, so öffnet sich das Grundfenster *GeoLings* (Abbildung 20). Drei Registerkarten – *Kategorien*, *Gruppen* und *Export* – sind oben links wählbar, während rechts unten drei Schaltflächen – „Öffne gewählte Karte...“, „Öffne Faktorenanalyse...“ und „Öffne Clusteranalyse...“ – bei Bedarf weitere Registrierkarten öffnen. Jede dieser Optionen wird in den folgenden Abschnitten dieses Handbuchs erläutert.

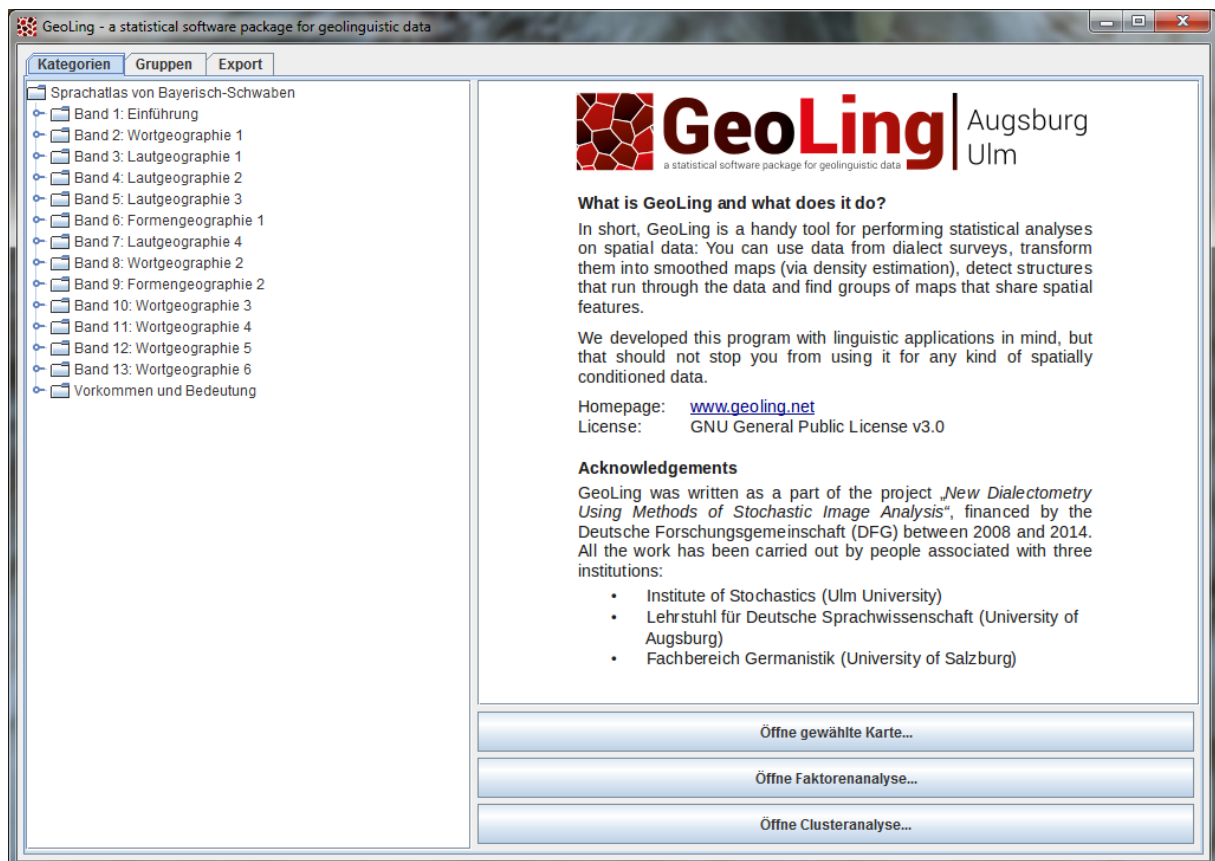


Abbildung 20: Grundfenster in GeoLing.

II.1 Kategorien

Hier werden die Inhalte der Datenbank in einer Baumstruktur angezeigt. In Abbildung 21 sieht man einen Ausschnitt aus der SBS-Datenbank: Wie in der Printausgabe gibt es 13 Bände, jeder davon ist in Kategorien unterteilt, die jeweils die einzelnen Karten enthalten.

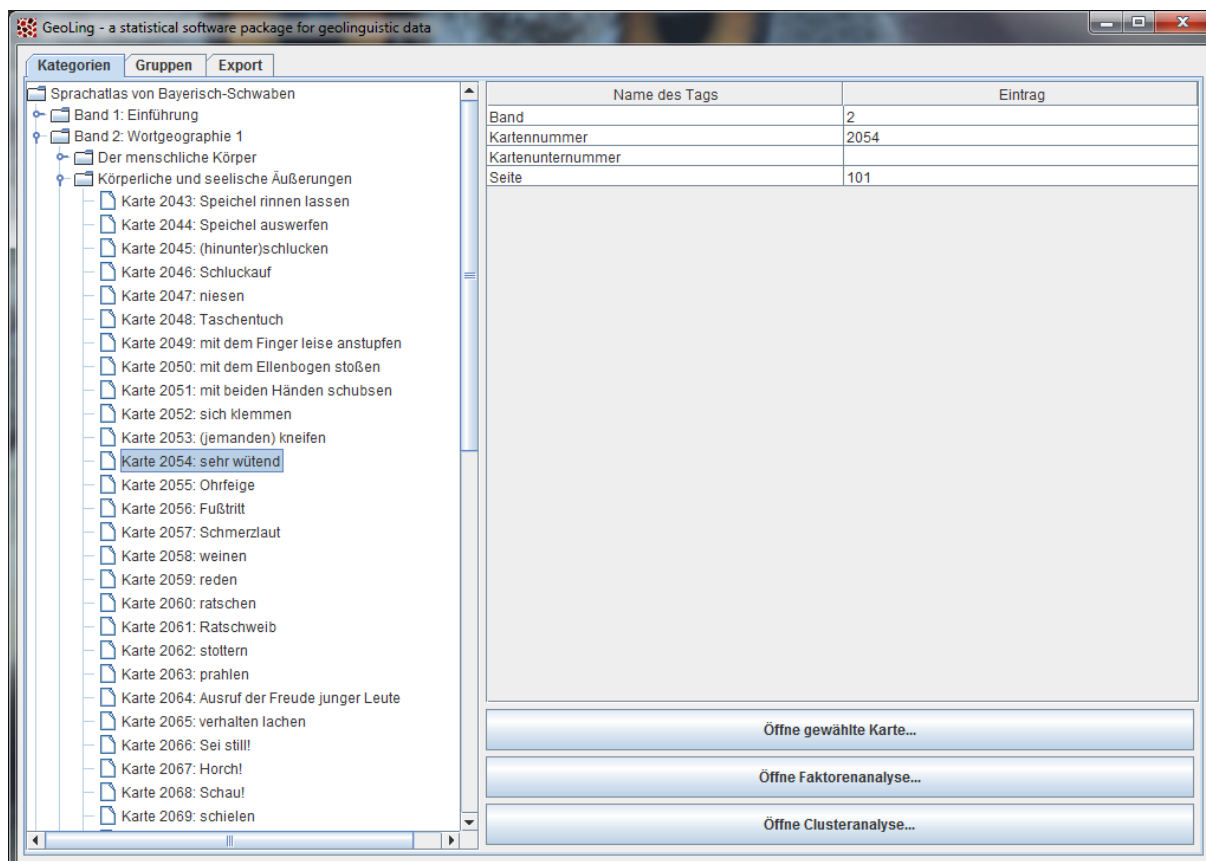


Abbildung 21: Registrierkarte Kategorien.

Bei einem Klick auf den Eintrag einer Karte (Abbildung 21) werden im rechten Fensterteil zusätzliche Informationen angezeigt; ein Doppelklick (oder die Schaltfläche „Öffne gewählte Karte...“) erzeugt eine neue Registrierkarte zur Dichteschätzung.

II.2 Dichteschätzung

Hier lassen sich die punktuellen, diskreten Informationen eines Datensatzes in eine fließende Flächenkarte umwandeln. Bevor wir tiefer in die technischen Details einsteigen, lohnt sich ein kurzer Blick darauf, was der Zweck einer solchen Uminterpretation der Daten ist.

Motivation

Die Grundmotivation für die Nutzung einer Dichteschätzung von Dialektdaten führt zurück zur Situation der Datenerhebung: Alle erhobenen Daten sind statistisch gesehen zunächst Stichproben – nicht mehr und nicht weniger. Zwar wird seit jeher von Seiten der Dialektologie viel getan, um Störfaktoren in den Daten zu minimieren und möglichst passende Gewährspersonen zu finden, aber dennoch besteht wohl kein ernsthafter Zweifel daran, dass es so etwas wie eine maximal repräsentative Stichprobe in der Praxis nicht gibt. Zeichnen wir aber eine Punktsymbolkarte, basierend auf punktuell erhobenen Daten, implizieren wir just dieses; bei einer dichtegeschätzten Flächenkarte ist dies jedoch nicht der Fall.

Wenn wir davon ausgehen, dass unsere Daten Stichproben sind, dann folgt daraus, dass sie eine gewisse Fehlermarge enthalten. Hätte man andere – oder mehr – Gewährspersonen gehabt, wäre die Erhebung an einem anderen Tag und einer anderen psychischen Verfassung durchgeführt worden usw., so wären die Antworten eventuell anders ausgefallen. Um dies zu kompensieren, können wir jedem Datenpunkt seine Wahrscheinlichkeit – die Intensität – zuordnen und diese an Stelle des Originalbelegs kartieren. Um diese Wahrscheinlichkeit bestimmen zu können, zieht man die Nachbarbelege zu Rate, ausgehend von der Überlegung, dass Nähe oftmals mit linguistischer Ähnlichkeit korreliert. (Wie genau wir Nähe definieren können, werden wir später noch aufgreifen.) Geglättete Flächenkarten sind daher nicht einfach nur eine grafische Spielerei, sondern sie bieten eine probabilistische Sicht auf die zugrundeliegenden Daten.³ Dies stellt eine Reinterpretation der Daten dar, keine Modifikation: Die Dichteschätzung ändert das „Gewicht“ eines Datums nicht, es „verteilt“ es nur in der Fläche.

Parameter

Kehren wir nun zur Programmoberfläche zurück. In Abbildung 22 sieht man die Maske zur Berechnung einer Flächenkarte, auf der alle relevanten Parameter bestimmt werden können.

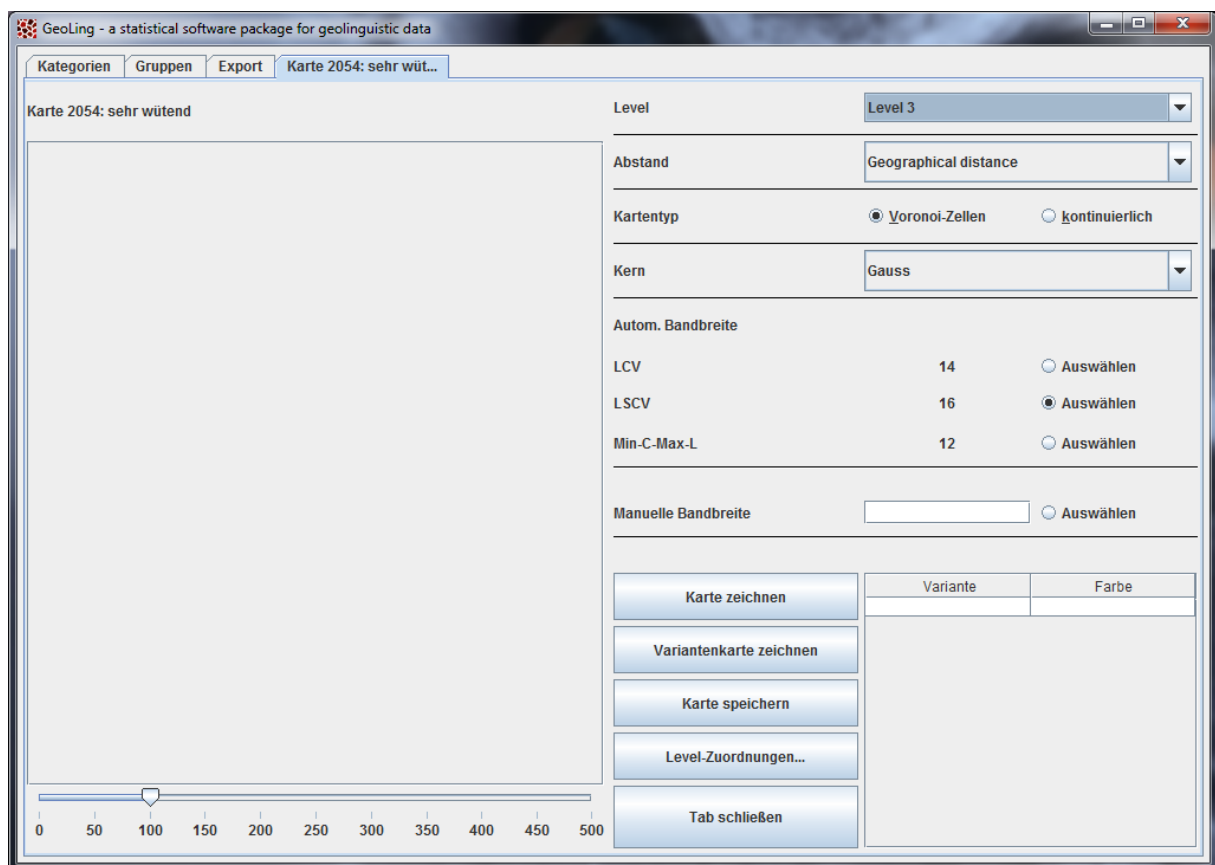


Abbildung 22: Maske für die Dichteschätzung einer Karte.

³ Der gesamte Prozess wird durch RUMPF / PICKL / ELSPAB / KÖNIG / SCHMIDT (2009) genauer erläutert. Anwendungen sind z.B. RUMPF (2010), PICKL / RUMPF (2011, 2012), PICKL (2013a, b) und PRÖLL (2013a, b) zu entnehmen. SCHERRER (2011) und SCHERRER / RAMBOW (2010a, b) nutzen das grundlegende Verfahren für Systeme zur automatisierten Übersetzung.

Im Folgenden werden alle Parameter von oben nach unten aufgeführt:

Parameter	Beschreibung
Level	Hier kann der Abstraktionsgrad der genutzten Daten bestimmt werden. Die SBS-Datenbank hat drei Level, die in ansteigendem Maße phonetische und morphologische Feinheiten ignorieren bzw. zusammenfassen (PICKL 2013b: 75–78 äußert sich dazu eingehender). Die Levelerzeugung für eigene Daten wird in Abschnitt I.4 behandelt.
Abstand	Dieses Auswahlfeld legt die Art des Abstandsmaßes zwischen den Erhebungsorten fest. In der SBS-Datenbank sind geografischer und linguistischer Abstand verfügbar: ⁴ Eigene Abstandsmaße können problemlos eingelesen werden (siehe auch hierzu I.4).
Kartentyp	Es können zwei verschiedene Arten von Karten erzeugt werden. Für die eine wird die Karte in Voronoizellen unterteilt; jeder Punkt der Kartenfläche wird dem nächsten Erhebungspunkt zugewiesen. Die andere ist kontinuierlich; jeder Wert wird auf einem engmaschigen Gitter interpoliert (das ist nur bei geografischem Abstand möglich, vgl. ausführlicher PRÖLL 2013b: 58–64).
Kern	Hier kann die Art des Kerns gewählt werden (Gauss, K_3 oder Epanechnikov), der für die Glättung eingesetzt werden soll. Der Kern ist für das Resultat der Dichteschätzung weniger entscheidend als andere Parameter (vgl. JANERT 2010: 20 bzw. konkret für unsere Situation RUMPF 2010: 49–63), kann aber dennoch einen Unterschied bewirken.
Autom. Bandbreite	Die drei Optionen dieses Abschnitts stehen für drei verschiedene Methoden, die Bandbreite automatisch zu optimieren: <ul style="list-style-type: none"> - LCV (likelihood cross-validation) - LSCV (least-squares cross-validation) - Min-C-Max-L, ein Algorithmus, der über Komplexität und Homogenität kosten-/nutzenoptimierend arbeitet (vgl. ausführlich PICKL 2013a).
Manuelle Bandbreite	Durch Eingabe einer Zahl kann hier selbst eine Bandbreite gewählt werden; dadurch werden die automatischen Werte umgangen.

Visualisierung

Nach erfolgter Parameterwahl bewirkt ein Klick auf „Karte zeichnen“, dass die Dichteschätzung ausgeführt und eine entsprechende Karte gezeichnet wird. Abbildung 23 zeigt ein Beispiel aus der SBS-Datenbank. Verschiedene Farben repräsentieren verschiedene dominante Varianten – mit „dominant“ ist hier gemeint, dass eine Variante eine höhere Intensität als alle anderen an einem Ort hat.

Die erzeugten Karten sind interaktiv: Bewegt man den Mauszeiger zu einem Ort, wird ein kleines Infofeld mit weiteren Informationen zum Ort bzw. zu den Intensitäten angezeigt.

⁴ PICKL / SPETTL / PRÖLL / ELSPAB / KÖNIG / SCHMIDT (2014) enthält einen systematischen Vergleich der Ergebnisse mittels geografischen und linguistischen Abständen.

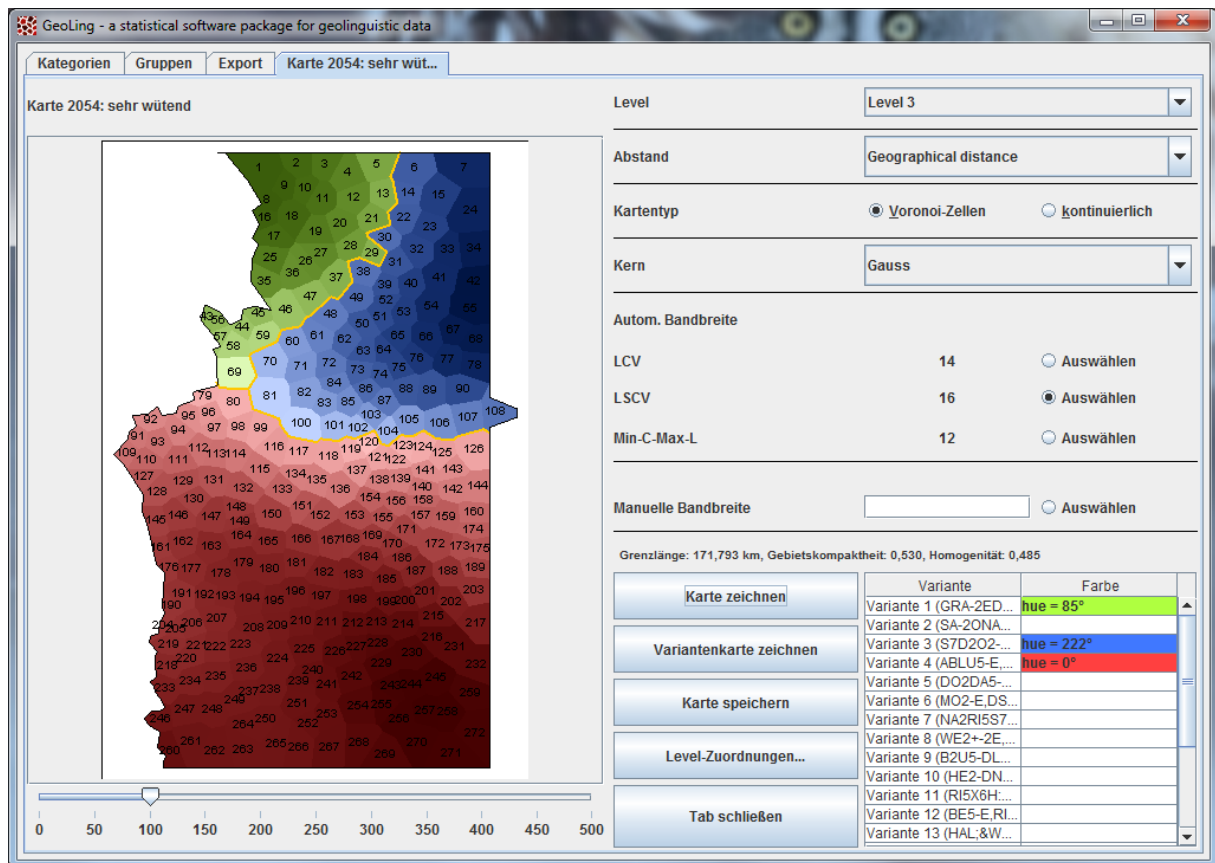


Abbildung 23: Flächenkarte.

Raumstatistische Werte

Zwischen den Parametern und den unteren Bedienelementen werden drei Werte angezeigt, die in RUMPF / PICKL / ELSAB / KÖNIG / SCHMIDT (2009) eingeführt wurden:

- Grenzlänge (oder Komplexität)
- Kompaktheit
- Homogenität

Diese Werte können für verschiedenste Zwecke genutzt werden: PICKL (2013a,b) und PRÖLL (2013b) zeigen nur ein paar der möglichen Interpretationen, die mittels statistischer Tests zu diesen Werten möglich sind.

Export

Per Klick auf „Karte speichern“ stehen mehrere Exportfunktionen zur Auswahl. Die Karte selbst kann als Vektorgrafik im EPS-Format (Encapsulated PostScript) oder als Rastergrafik im PNG-Format (Portable Network Graphics) gespeichert werden. Dazu sind aber auch die rohen Werte der Berechnungen direkt in der Form einer XML-Datei (Extensive Markup Language) exportierbar.

Einzelvarianten

In den Feldern rechts unten sind alle Varianten aufgeführt, die im aktuellen Datensatz vorhanden sind. Per Doppelklick auf die Varianten ist es möglich, ihre Farbzweisung zu ändern (Abbildung 24).

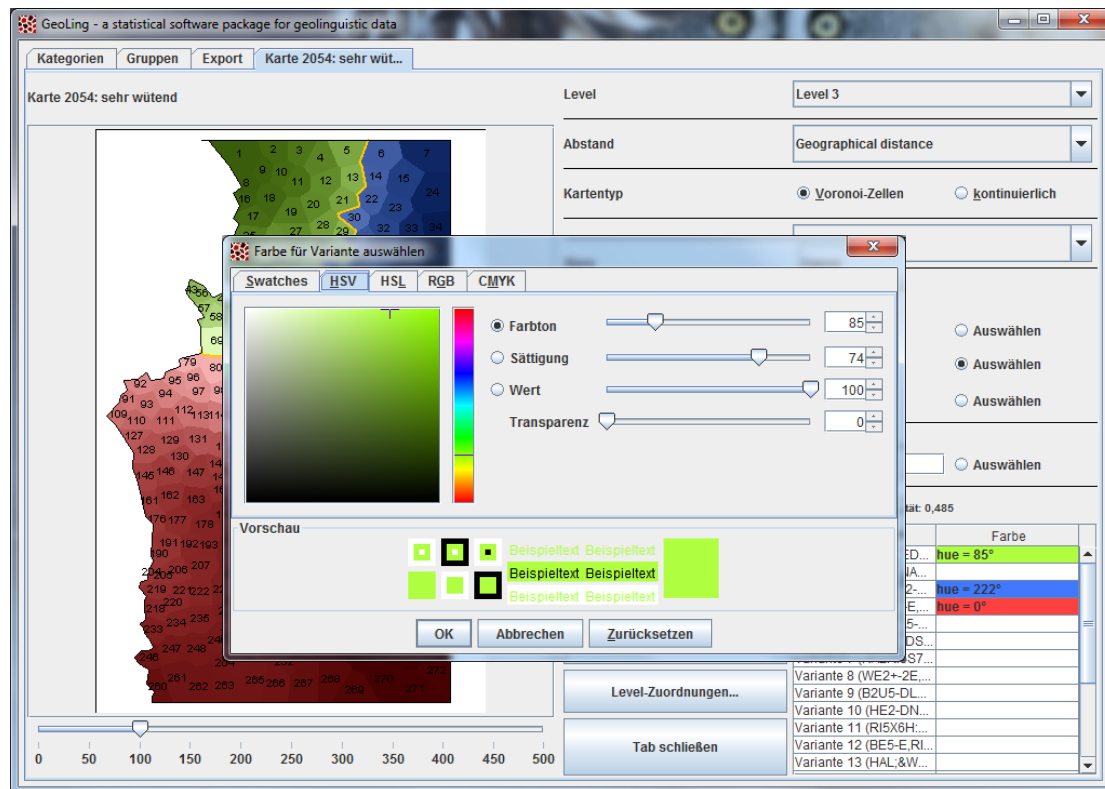


Abbildung 24: Wechseln der Farbe einer Variante.

Mit einem einzelnen Klick auf eine Variante und einem weiteren auf „Variantenkarte zeichnen“ wird eine Karte erzeugt, die die Intensität nur dieser einzelnen Variante visualisiert (egal ob sie dominant ist oder nicht). Standardmäßig ist eine solche Karte blau (Abbildung 25). Klickt man stattdessen nicht auf die Variante, sondern auf das Farbfeld rechts von ihr, kann man sie per „Variantenkarte zeichnen“ auch in der Farbe, die sie in der Gesamtkarte hatte, visualisieren.

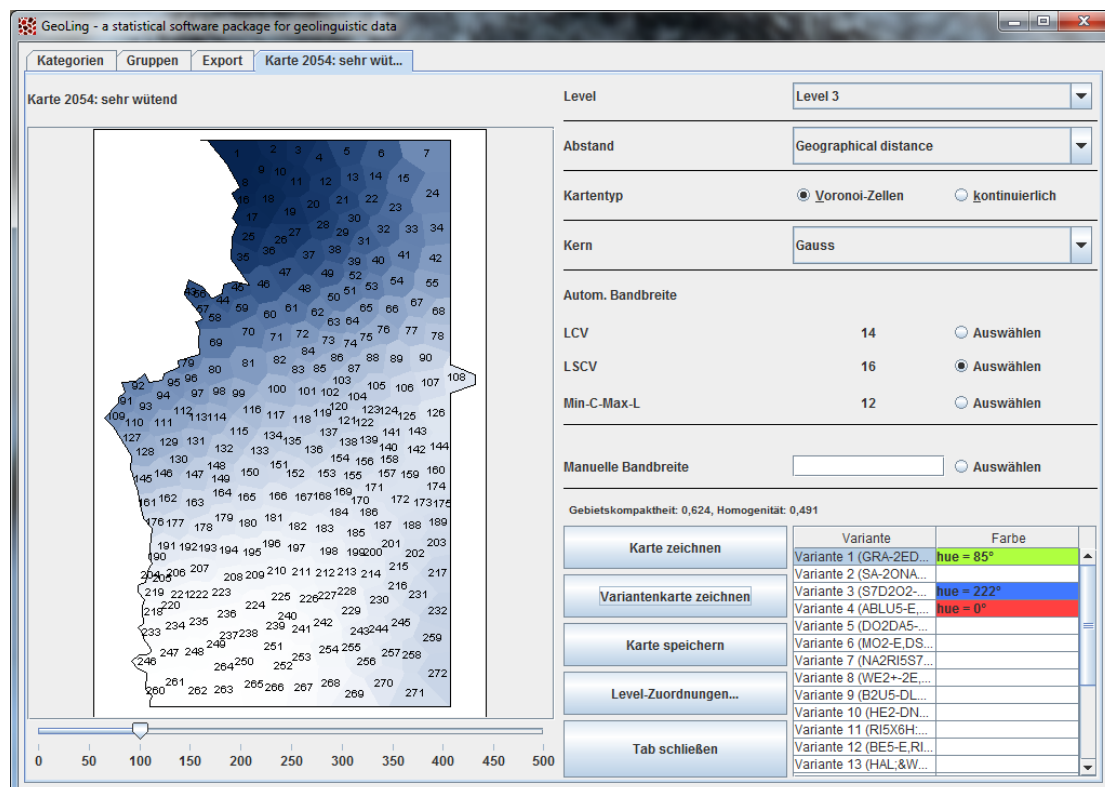


Abbildung 25: Einzelvariante.

II.3 Gruppen

Diese Registrierkarte bietet die Möglichkeit, Karten zu Gruppen zusammenzufassen; diese Gruppen können als Basis für weitere Schritte genutzt werden, wie etwa eine Clusteranalyse aller Karten zum Vokalismus oder eine Faktorenanalyse der Karten zum semantischen Feld ‘Getreide’. Abbildung 26 zeigt als Beispiel eine Gruppe aller SBS-Karten zur Morphologie. Im linken Fenster sind alle gespeicherten Gruppen aufgelistet (inklusive der Anzahl der ihnen zugeordneten Karten). Ein Klick auf eine Gruppe zeigt ihren Inhalt im rechten Fenster.

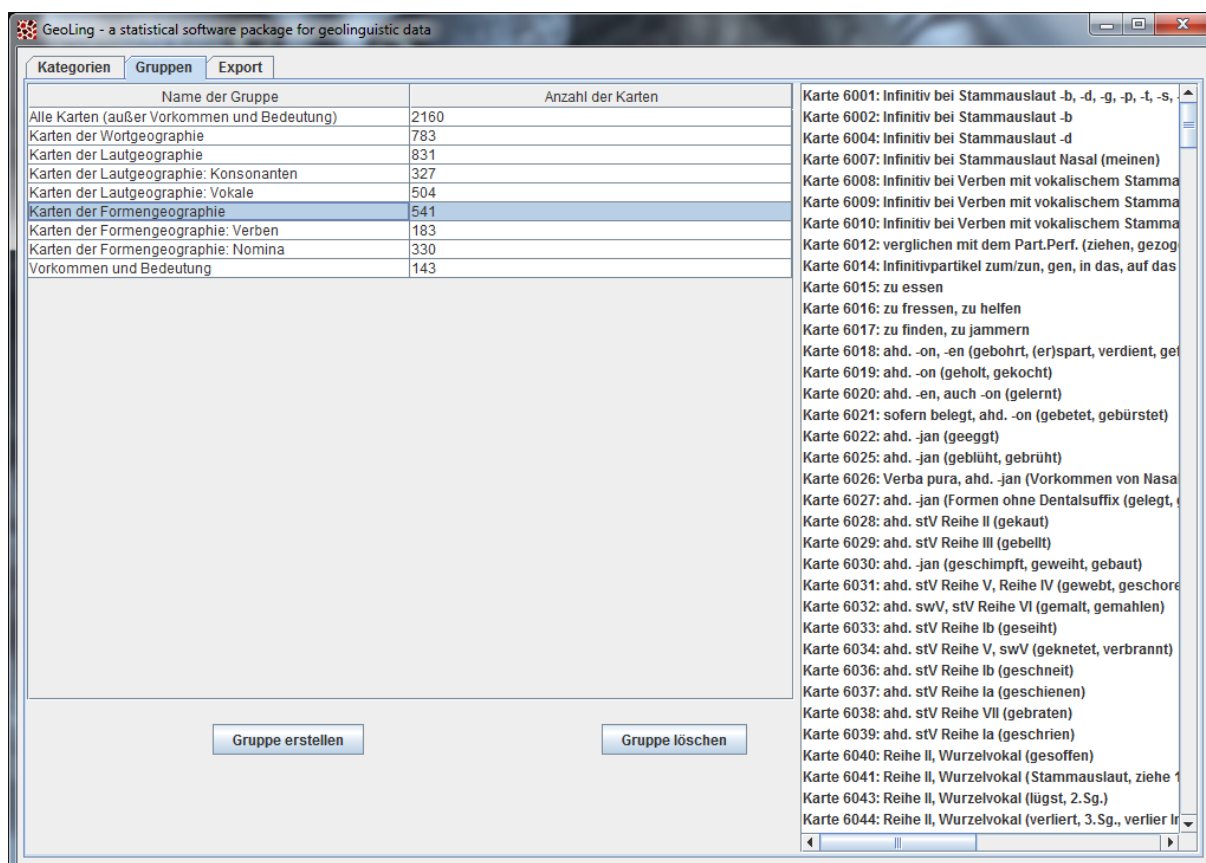


Abbildung 26: Gruppen-Registrierkarte mit einer ausgewählten Gruppe (links) und ihrem Inhalt (rechts).

Neue Gruppen können durch Gruppe erstellen erzeugt werden – so öffnet sich eine Registrierkarte, mittels der man einzelne Karten per Checkboxen auswählen kann.

II.4 Export

Will man mehrere Karten nutzen, wäre es unökonomisch, jede einzeln erzeugen und exportieren zu müssen. Zu diesem Zweck gibt es die Export-Registrierkarte, die jeweils ganze Gruppen (siehe II.3) ausgeben kann.

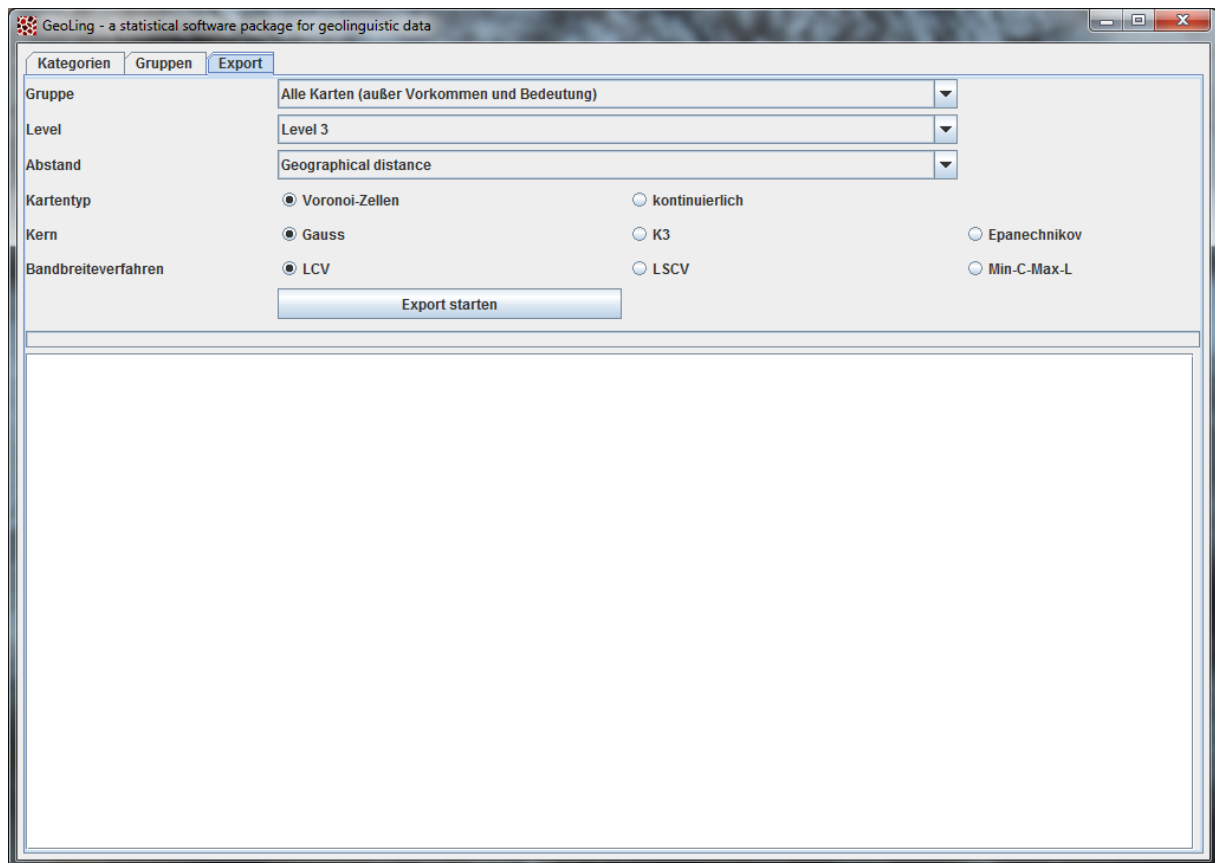


Abbildung 27: Exportfunktion.

Hier kann man eine vordefinierte Gruppe (siehe II.3) ausgewählt, die gewünschten Parameter (siehe II.2) für die Einzelkarten eingestellt sowie ein Ausgabeverzeichnis bestimmt werden; das Programm generiert dann für jeden Datensatz die Karte als EPS- und PNG-Datei sowie eine XML-Datei mit den rohen Werten. Dazu wird auch pro Gruppenexport eine CSV-Tabelle (mit gängiger Bürosoftware problemlos zu öffnen) ausgegeben, die eine Liste aller Karten der Gruppe, ihrer Komplexität, Kompaktheit und Homogenität sowie der Anzahl ihrer dominanten Varianten enthält.

II.5 Faktorenanalyse

Über eine weitere Schaltfläche auf dem Startbildschirm (vgl. II) lässt sich eine Registrierkarte für Faktorenanalysen öffnen.

Hintergrundinformationen zur Faktorenanalyse

Die Faktorenanalyse ist ein Verfahren zur Dimensionsreduktion in Datensätzen (ursprünglich für persönlichkeitspsychologische Anwendungen entwickelt, vgl. WOTAWA 1996: 813). Sie soll die Gesamtvariation in Form einer möglichst geringen Anzahl gemeinsamer Muster interpretieren; so lassen sich große Teile der Originalvariation durch Kookkurrenzen sehr kompakt beschreiben. Nach unserem Kenntnisstand erfolgte die erste variationslinguistische Anwendung der Faktorenanalyse durch WERLEN (1984), durch BIBER (1985, 1991) wurde sie in der Korpuslinguistik gängig.

Parameter

Über die Oberfläche sind drei Parameter direkt zugänglich (Abbildung 28):

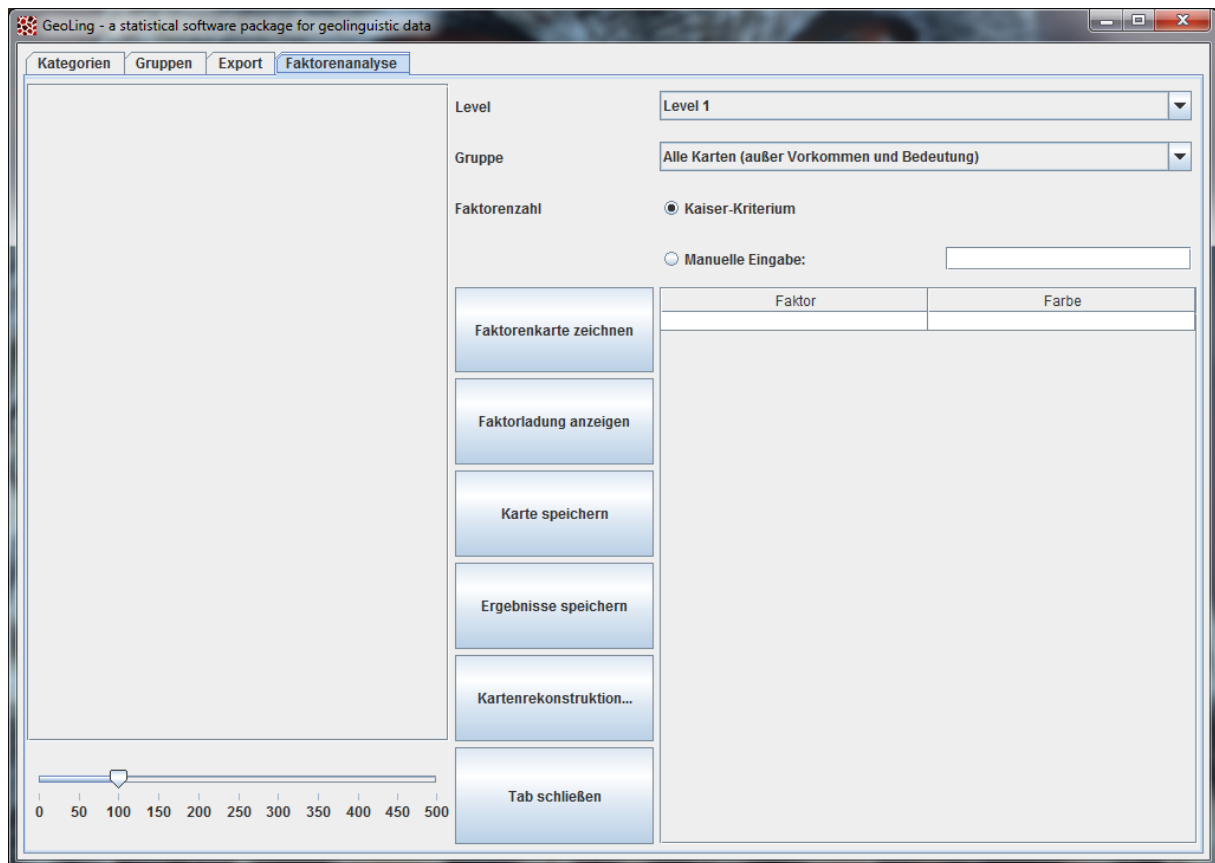


Abbildung 28: Maske für Faktorenanalyse.

Der erste ist wiederum das Abstraktionsniveau. Für die Faktorenanalyse empfiehlt es sich dabei, möglichst nahe an den Rohdaten zu bleiben, also Level 1 zu wählen. Über „Gruppe“ (vgl. II.3) wird das Korpus an Karten gewählt, das analysiert werden soll. Zuletzt last sich die Anzahl an Faktoren, die aus den Daten extrahiert werden sollen, bestimmen. Prinzipiell gibt es keine „richtige“ oder „falsche“ Anzahl an Faktoren. Trotzdem haben sich Vorgehensweisen etabliert, um den nötigen Entscheidungsprozess zu optimieren; in *GeoLing* implementiert ist das *Kaiser-Guttman-Kriterium*.⁵ Es ist aber auch über die Textbox möglich, manuell eine Faktorenzahl vorzugeben.⁶

Visualisierung

Mittels Klick auf Karte zeichnen wird die Faktorenanalyse ausgeführt und eine entsprechende Karte gezeichnet. Abbildung 29 zeigt ein typisches Resultat, eine Faktorenkarte zu allen SBS-Karten, die die Nominalmorphologie betreffen. Auch hier ist die Karte interaktiv: Befindet sich der Mauszeiger über einem Ort, wird das lokale Zusammenspiel der Faktoren in einer Infobox aufgeschlüsselt.

Auf der rechten Seite erscheint eine Liste aller extrahierten Faktoren, jeweils mit dem Prozentsatz der Gesamtvariation, die durch sie abgedeckt wird. Die Schaltflächen „Ergebnisse speichern“ und „Karte speichern“ ermöglichen es, die Rohwerte als XML-Datei und die Karte als EPS-Datei zu speichern.

⁵ Das Kaiser-Guttman-Kriterium berücksichtigt nur Faktoren mit einem Eigenwert größer als 1.

⁶ Generell wird das Kaiser-Guttman-Kriterium eher als conservative Lösung angesehen, weil es dazu neigt, eine relativ hohe Anzahl an (schwächeren) Faktoren zuzulassen (und damit die Originaldaten weniger stark zusammenzufassen). Unsere Anwendungen (PICKL 2013b, PRÖLL 2013, PRÖLL / PICKL / SPETTL i.E.) zeigen aber, dass aus linguistischer Perspektive auch schwache Faktoren stimmig interpretiert werden können.

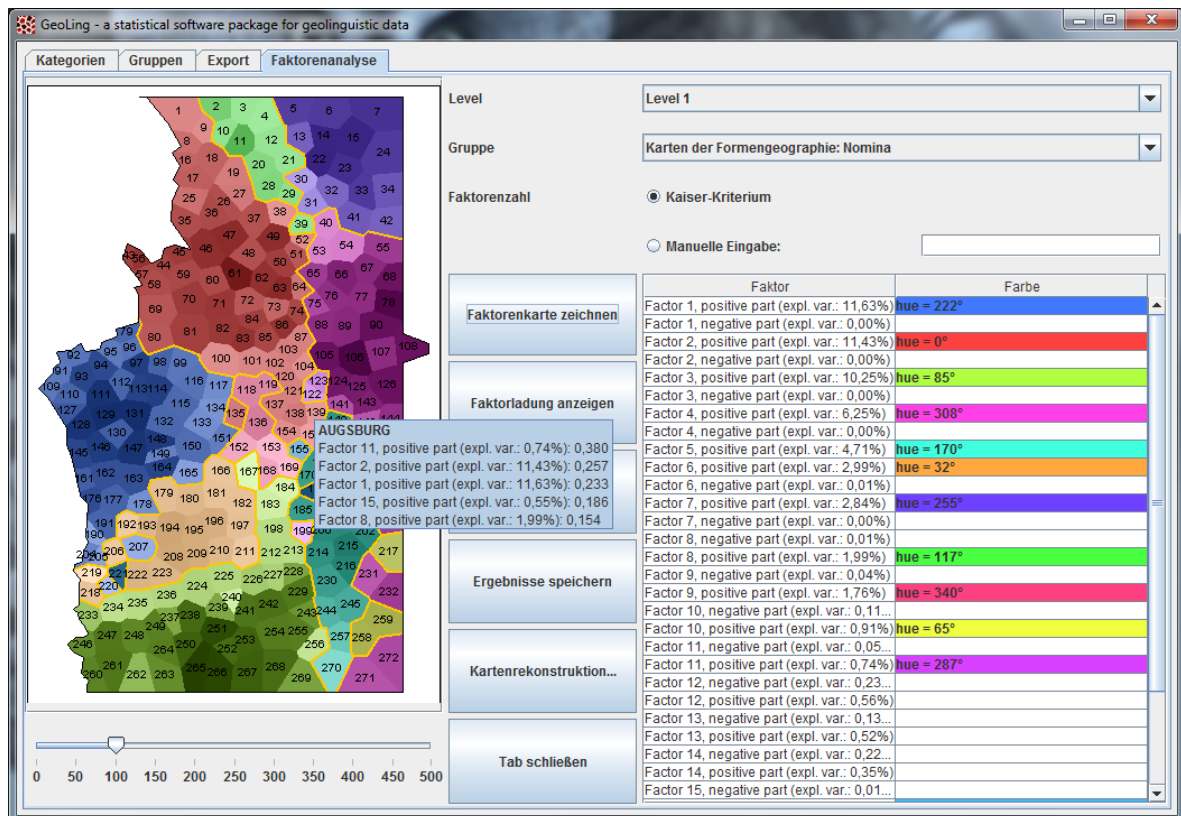


Abbildung 29: Beispiel: Faktorenanalyse der SBS-Karten zur Nominalmorphologie.

Wie schon im Abschnitt zur Dichteschätzung für einzelne Varianten beschrieben (siehe II.2) kann man auch einzelne Faktoren anklicken und per „Faktorladung anzeigen“ visualisieren (Abbildung 30).

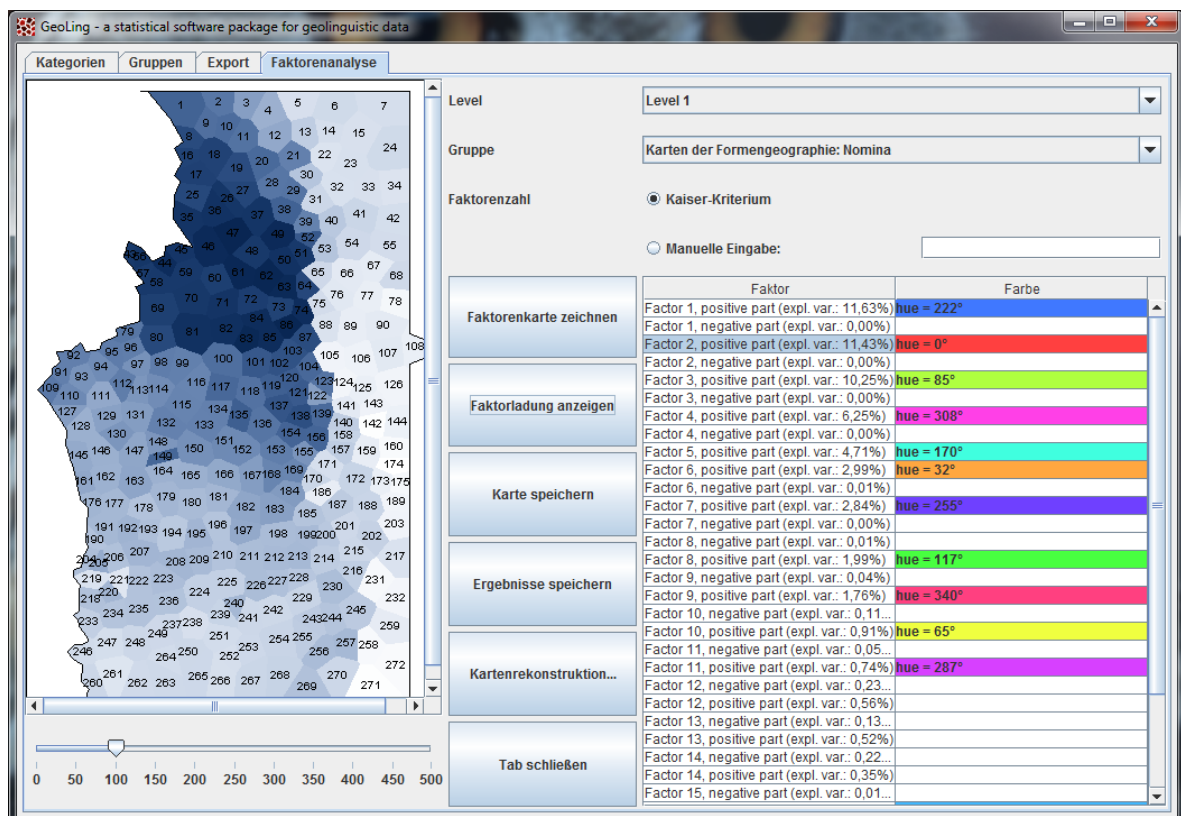


Abbildung 30: Beispiel: Einzelfaktor der obigen Faktorenanalyse.

Zusätzlich ist es im Anschluss an eine Faktorenanalyse möglich, einzelne Karten nicht per Dichteschätzung, sondern über die individuellen Faktorgewichte der Faktorenanalyse zu rekonstruieren. So werden die Anteile der Variation, die von der Faktorenanalyse als nicht korrelierend betrachtet und daher aussortiert worden sind, auch in Einzelkarten ausgeklammert. Zugänglich ist dies über die Schaltfläche „Kartenrekonstruktion...“, die eine entsprechende neue Registerkarte öffnet (Abbildung 31).

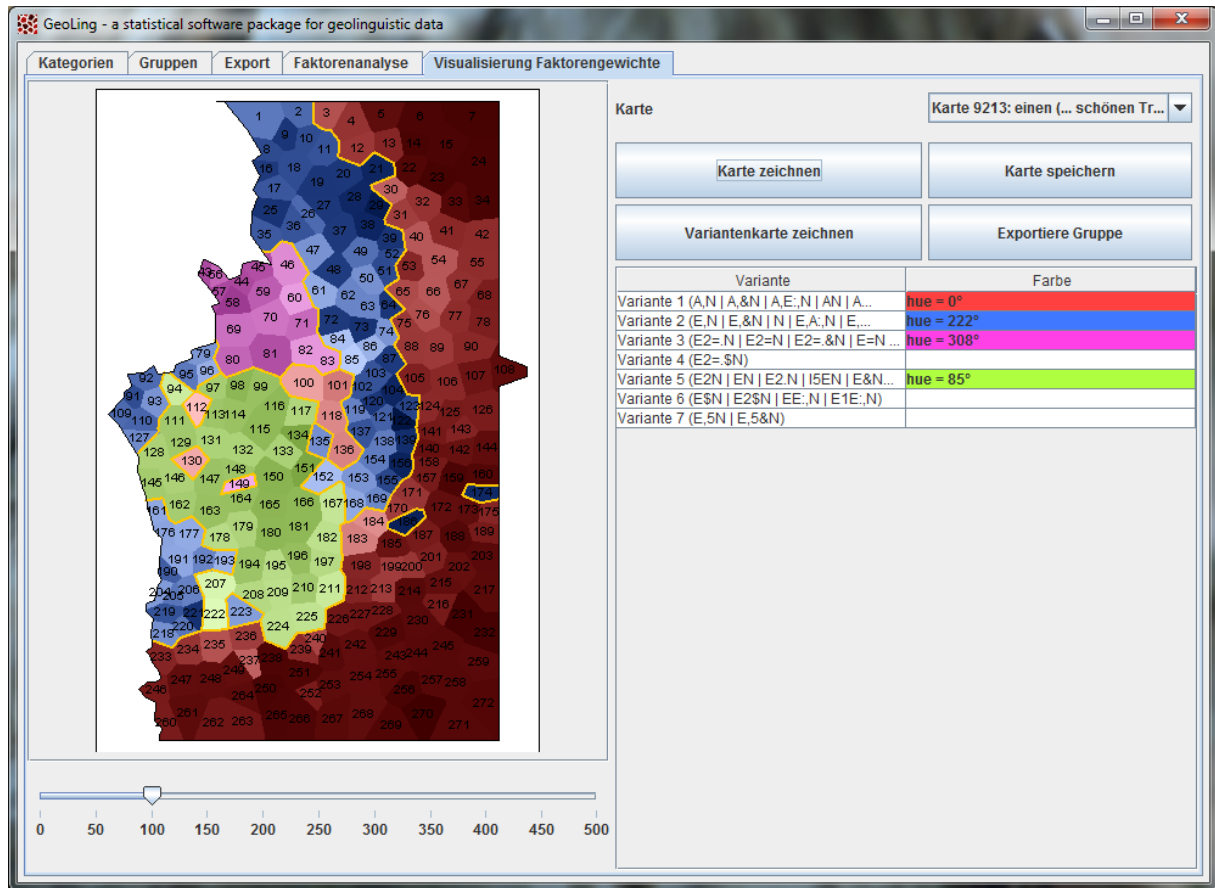


Abbildung 31: Rekonstruktion einer Karte mittels Faktoren.

II.6 Clusteranalyse

Über diese Registerkarte ist die Durchführung von Clusterverfahren möglich. Zwar ist Clustern schon länger ein Bestandteil des üblichen dialektometrischen Werkzeugkastens, *GeoLing* bietet aber eine gänzlich neue Perspektive: Es werden keine Dialekte (bzw. Orte) geclustert, sondern einzelne Karten. So lassen sich Gruppen von Karten finden, deren Raumbild sich ähnelt.

Kurz zum Prinzip des Clusters: Das menschliche cognitive System ist grundsätzlich erstaunlich gut darin, Muster in Reizen zu finden – vorausgesetzt, dass einige grundlegende psychologische Bedingungen erfüllt sind und die Datenmengen nicht zu groß sind. Will man aber z.B. ermitteln, welche der über zweitausend SBS-Karten sich wie stark ähneln, sind diese Voraussetzungen nicht gegeben; nur eine „objektive“, maschinelle Methode könnte dies leisten. Genau dies wird durch Clustern möglich.

Zum Clustern bedarf es a) paarweiser Distanzen (bzw. Ähnlichkeiten) zwischen den Karten sowie b) eines Algorithmus, der die Karten basierend auf diesen Werten nach Ähnlichkeit

gruppiert. Das Programm stellt dafür drei verschiedene Typen der Quantifizierung von „Abständen“ zwischen Karten bereit:

1. Abstandsvektoren, die auf einer Unterteilung der Kartenfläche in Sektoren beruhen
2. Helligkeitswerte
3. Abstandsvektoren, die auf Kovarianzwerten basieren

Als Algorithmen stehen hierarchisches Clustern, K-Means und Fuzzy C-Means zur Verfügung.

Unsere Publikationen zur Clusterung von Karten konzentrieren sich auf zwei Kombinationen von Abstandsmaßen zwischen Karten und Algorithmus: Die eine nutzt Sektoren und Helligkeit kombiniert mit hierarchischem Clustern (RUMPF / PICKL / ELSPAß / KÖNIG / SCHMIDT 2010); die anderen nutzen unscharfes Clustern (Fuzzy C-Means) von Kovarianzwerten (MESCHENMOSER / PRÖLL 2012a; PRÖLL 2013a, b). Das Programm erlaubt es jedoch, alle möglichen Kombinationen von Datenrepräsentation und Algorithmus verwenden zu können.

III. Literatur

- BIBER, Douglas (1985): Investigating macroscopic textual variation through multifeature / multidimensional analyses. In: *Linguistics* 23/2, 337–360.
- BIBER, Douglas (1991): *Variation across speech and writing*. Cambridge: Cambridge University Press.
- MESCHENMOSER, Daniel / PRÖLL, Simon (2012a): Using fuzzy clustering to reveal recurring spatial patterns in corpora of dialect maps. In: *International Journal of Corpus Linguistics* 17/2, 176–197.
- MESCHENMOSER, Daniel / PRÖLL, Simon (2012b): Automatic detection of radial structures in dialect maps: determining diffusion centers. In: *Dialectologia et Geolinguistica* 20, 71–83.
- PICKL, Simon (2013a): Lexical Meaning and Spatial Distribution. Evidence from Geostatistical Dialectometry. In: *Literary and Linguistic Computing* 28/1, 63–81.
- PICKL, Simon (2013b): *Probabilistische Geolinguistik. Geostatistische Analysen lexikalischer Distribution in Bayerisch-Schwaben*. Stuttgart: Steiner.
- PICKL, Simon / RUMPF, Jonas (2011): Automatische Strukturanalyse von Sprachkarten. Ein neues statistisches Verfahren. In: GLASER, Elvira / SCHMIDT, Jürgen Erich / FREY, Natascha (Hrsg.): *Dynamik des Dialekts – Wandel und Variation. Akten des 3. Kongresses der Internationalen Gesellschaft für Dialektologie des Deutschen (IGDD)*. (Zeitschrift für Dialektologie und Linguistik, Beihefte. Band 144.) Stuttgart: Steiner, 267–285.
- PICKL, Simon / RUMPF, Jonas (2012): Dialectometric Concepts of Space: Towards a Variant-Based Dialectometry. In: HANSEN, Sandra / SCHWARZ, Christian / STOECKLE, Philipp / STRECK, Tobias (Hrsg.): *Dialectological and folk dialectological concepts of space*. Berlin: Walter de Gruyter, 199–214.
- PICKL, Simon / SPETTL, Aaron / PRÖLL, Simon / ELSPAß, Stephan / KÖNIG, Werner / SCHMIDT, Volker (2014): Linguistic Distances in Dialectometric Intensity Estimation. In: *Journal of Linguistic Geography* 2/1, 25–40.
- PRÖLL, Simon (2013a): Detecting Structures in Linguistic Maps – Fuzzy Clustering for Pattern Recognition in Geostatistical Dialectometry. In: *Literary and Linguistic Computing* 28/1, 108–118.
- PRÖLL, Simon (2013b): *Raumvariation zwischen Muster und Zufall. Geostatistische Analysen am Beispiel des Sprachatlas von Bayerisch-Schwaben*. Dissertation, Universität Augsburg.
- PRÖLL, Simon (2014): Stochastisch gestützte Methoden der Dialektdifferenzierung. In: HUCK, Dominique (Hrsg.): *Dialekte im Kontakt. Beiträge zur 17. Arbeitstagung zur alemannischen*

- Dialektologie; Université de Strasbourg vom 26.–28. Oktober 2011. Stuttgart: Steiner, 229–242.
- PRÖLL, Simon / PICKL, Simon / SPETTL, Aaron (i.E.): Latente Strukturen in geolinguistischen Korpora. Erscheint in: ELEMENTALER, Michael / HUNDT, Markus (Hrsg.): Deutsche Dialekte. Konzepte, Probleme, Handlungsfelder. 4. Kongress der Internationalen Gesellschaft für Dialektologie des Deutschen (IGDD), Kiel vom 13.–15. September 2012. Stuttgart: Steiner.
- RUMPF, Jonas (2010): Statistical Models for Geographically Referenced Data. Applications in Tropical Cyclone Modelling and Dialectology. Dissertation, Universität Ulm.
- RUMPF, Jonas / PICKL, Simon / ELSPAß, Stephan / KÖNIG, Werner / SCHMIDT, Volker (2009): Structural analysis of dialect maps using methods from spatial statistics. In: Zeitschrift für Dialektologie und Linguistik 76/3, 280–308.
- RUMPF, Jonas / PICKL, Simon / ELSPAß, Stephan / KÖNIG, Werner / SCHMIDT, Volker (2010): Quantification and Statistical Analysis of Structural Similarities in Dialectological Area-Class Maps. In: Dialectologia et Geolinguistica 18, 73–100.
- SBS = KÖNIG, Werner (Hrsg.) (1996–2009): Sprachatlas von Bayerisch-Schwaben. Band 1–14. Heidelberg: Winter.
- SCHERRER, Yves (2011): Morphology Generation for Swiss German Dialects. In: MAHLOW, Cerstin / PIOTROWSKI, Michael (Hrsg.): Systems and Frameworks for Computational Morphology: Second International Workshop, SFCM 2011, Zurich, Switzerland, August 26, 2011, Proceedings, 130–140.
- SCHERRER, Yves / RAMBOW, Owen (2010a): Natural Language Processing for the Swiss German dialect area. In: PINKAL, Manfred / REHBEIN, Ines / SCHULTE IM WALDE, Sabine / STORRER, Angelika (Hrsg.): Semantic Approaches in Natural Language Processing. Proceedings of the Conference on Natural Language Processing 2010, 93–102.
- SCHERRER, Yves / RAMBOW, Owen (2010b): Word-based dialect identification with georeferenced rules. In: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, MIT, Massachusetts, USA, 9.–11. October 2010, 1151–1161.
- VOGELBACHER, Julius (2011): Statistische Analyse von Wortvarianten des Sprachatlas von Bayerisch-Schwaben. Diplomarbeit, Universität Ulm.
- WERLEN, Erika (1984): Studien zur Datenerhebung in der Dialektologie. Stuttgart: Steiner.