Geoling 2.0 – Ein aktueller Bericht aus der Werkstatt der webbasierten Sprachgeographie

Thomas Krefeld & Stephan Lücke, Ludwig-Maximilians-Universität München

Vorgestellt werden die folgenden vier Projekte: *Atlante sintattico della Calabria* (AsiCa; <www.asica.gwi.uni-muenchen.de>); *Audio-Atlas siebenbürgisch-sächsischer Dialekte* (ASD; <http://www.asd.gwi.uni-muenchen.de/?projektinfo=true>); *Atlante linguistico digitale dell'Italia e della Svizzera meridionale* (AdIS; <www.adis.gwi.uni-muenchen.de>); *Metropolitalia* (<www.metropolitalia.org>). Diese sehr unterschiedlich fortgeschrittenen Unternehmungen haben zwar jeweils eigene Zielsetzungen; ihnen allen liegt jedoch eine gemeinsame Konzeption digitaler Sprachgeographie zu Grunde, deren wissenschaftliche und technische Prinzipien in diesem Beitrag entwickelt werden.

1 Ausgangspunkt

Die sprachwissenschaftliche Forschung befindet sich in einer Übergangsphase. Ihre Rahmenbedingungen haben sich seit der medialen Revolution so grundlegend verändert, dass es notwendig ist, die etablierten Forschungstraditionen systematisch zu überdenken und in mancherlei Hinsicht neu zu justieren. Hier sind nun vor allem die Bereiche gefordert, die mit der Erhebung und Aufbereitung empirischer Daten zu tun haben. Exemplarisch ist die Situation der Sprachgeographie, also der Subdisziplin, die seit ihrer Begründung zu Beginn des 19. Jahrhunderts (Krefeld 2007) in systematischer Erhebung gesprochener Sprachdaten fundiert ist. Eine kurze Standortbestimmung wollen wir ausgehend von unseren im Folgenden genannten Projekten versuchen; dabei ist die Darstellung vor dem Hintergrund des Kongresses vor allem auf die Georeferenzierungsproblematik zugespitzt (Abschnitt 3.3.). Gleichzeitig soll die teils sehr detaillierte Beschreibung als Einladung verstanden werden, sich diese Konzeption zu eigen zu machen und weiterzuentwickeln. Das wäre unbedingt im Sinne einer wünschenswerten, wenn nicht notwendigen Standardisierung der digitalen Sprachgeographie.

1.1 Atlante sintattico della Calabria (AsiCa; → www.asica.gwi.uni-muenchen.de)

Dieser Atlas zielt auf die syntaktischen Besonderheiten des Kalabresischen;[1] er wurde von 2004-2006 und 2007-2008 von der DFG gefördert. Erfasst werden acht Ortsdialekte, wobei in jedem Dialekt jeweils mehrere, nämlich in der Regel acht Informanten ein und derselben Familie aufgenommen wurden, die teils in Italien, teils in Deutschland leben; darunter sind stets zwei Generationen und beide Geschlechter vertreten (**www.asica.gwi.uni-muenchen.de/index.php?informanti=1**). Die Datenbasis umfasst ca. 400.000 Tokens, die teils in semispontanen Interviews (so genannten *etnotesti*), teils in der Übersetzung von 54 Beispielsätzen erhoben und in strukturierter Form in eine Datenbank eingespeist wurden. Es handelt sich um ein charakteristisches Werk der eingangs erwähnten Übergangsphase: Obwohl die ursprüngliche Konzeption eine traditionelle Publikation vorsah, wurden alle Ergebnisstufen seit 2006 ganz konsequent in einem Online-Portal zugänglich gemacht; die Dokumentation ist multimedial, insofern die Daten in akustischer und (wenngleich noch nicht vollständig) in transkribierter Ver-

[1] Vgl. dazu allgemein Krefeld/Lücke 2008; eine erste Auswertung gibt Salminger 2009; eine Detailstudie findet sich in Krefeld 2007a.

sion abgerufen werden können. Alle transkribierten Materialien sind tokenisiert, lemmatisiert und alle Tokens/Lemmata sind in ihrem jeweiligen Kontext einsehbar. Einstweilen werden nur ausgewählte Korpusdaten kartographisch präsentiert.

1.2 Audioatlas Siebenbürgisch-sächsischer Dialekte (ASD; → www.asd.gwi.uni-muenchen.de)

Diesem Atlas liegt älteres, aber niemals aufgeschlüsseltes oder gar publiziertes Material zu Grunde, das zwischen 1968 und 1973 von rumänischen Germanisten der Universitäten Bukarest, Hermannstadt und Klausenburg auf Tonband aufgenommen wurde; das Projekt wird von 2011-2014 vom Beauftragten der Bundesregierung für Kultur und Medien (BKM) gefördert. Zugänglich sind insgesamt über 350 Stunden Audio-Material (2212 Dateien, 141 GB wav, 11 GB mp3), von dem derzeit (Januar 2014) knapp 260[2] Stunden auch in transkribierter Version vorliegen. Ähnlich wie im AsiCa handelt es sich dabei zum einen um elizitierte Beispielsätze, die berühmten Wenkersätze, die in 139 Ortsdialekte übersetzt wurden, und zum anderen um Spontanmaterial; fast in jedem Ort wurden mehrere Informanten sehr unterschiedlichen Alters erfasst (insgesamt 1428 Personen). Einstweilen werden wie beim AsiCa nur ausgewählte Korpusdaten kartographisch präsentiert; im Unterschied zum AsiCa ist allerdings derzeit bereits eine sehr differenzierte onomasiologische, bzw. 'ontologische' Aufschlüsselung des Spontanmaterials verfügbar; aktuell erfolgt das exhaustive linguistische Tagging des Wenkersatzmaterials. Es ist daher die Einrichtung einer benutzergesteuerten Kartographierungsfunktion absehbar; sie wird einerseits die Verbreitung einzelner Varianten zeigen und andererseits alle erfassten Varianten in quantitativ-dialektometrischer Form darstellen. Im Unterschied zu anderen dialektometrischen Studien erfolgt die Quantifizierung in vollkommener Transparenz, denn es wird genau dokumentiert, auf welchen Merkmalen die jeweils dargestellte Ähnlichkeit zwischen frei wählbaren Bezugs und Vergleichspunkten beruht.

1.3 Atlante linguistico digitale dell'Italia e della Svizzera meridionale (AdIS; → www.adis.gwi.uni-muenchen.de)

Dieses Projekt befindet sich in seiner Startphase; es strebt eine mindestens partielle Tiefendigitalisierung des *Sprach- und Sachatlas Italiens und der Südschweiz* (AIS) von Karl Jaberg und Jakob Jud (1928-40) an. Das Originalmaterial wurde weitestgehend mit einem umfangreichen Fragebuch an 416 Orten elizitiert und in Gestalt von 1681 Karten, 20 Konjugationstabellen sowie einem Indexband (1960) publiziert. Derzeit sind rund 40 Karten abrufbar, die auf einer 'händischen' Übertragung der gedruckten Originaldaten in eine Datenbank basieren. Damit werden also gewissermaßen kartographische Daten in ein teils bereits getaggtes digitales Korpus verwandelt. Es besteht die Absicht, die georeferenzierten sprachlichen Daten mit georeferenzierbaren, nicht sprachlichen Daten aus unterschiedlichen Bereichen, wie etwa der Archäologie und Siedlungsgeschichte zu verbinden, um zu einer induktiven Kartographie kultureller Räume zu gelangen.

1.4 Metropolitalia (→ www.metropolitalia.org)

Dieses Portal ist insofern innovativ und experimentell, als es konsequent die Crowdsourcing-Optionen des Web 2.0 (*social software*) auszuloten versucht. Es bietet eine Spiel-

[2] Summe der Länge der transkribierten Einzeldateien

oberfläche im Sinne eines *game with a purpose* (GWAP) mit dem Zweck, Sprachdaten und sprachbezogene Metadaten zu sammeln. Angestrebt ist der Aufbau eines pluridimensionalen Observatoriums, das die aktuelle räumliche Variation des Italienischen, sowohl im Blick auf die Dialekte, wie – vor allem – im Blick auf das Regionalitalienische kartographisch abbildet.

1.5 Verba Alpina (→www.verba-alpina.gwi.uni-muenchen.de)

Dieses Projekt steckt in der Antragsphase und existiert bislang nur als Konzept; vorgesehen ist eine großräumige Stratigraphie des Alpenraums im Spiegel seiner Mehrsprachigkeit. Die technischen und methodologischen Ergebnisse der Projekte (i)-(iv) laufen hier zusammen, denn es werden retrodigitalisierte Daten aus den verfügbaren Atlanten und weitere georeferenzierbare Daten aus Wörterbüchern mit neuerhobenen („crowdgesourceten") Daten in einer Datenbank und einer gemeinsamen kartographischen Oberfläche kombiniert. Hinzu kommt Datentransfer aus anderen aktuellen Projekten zum Alpenraum, so dass die sprachgrenzüberschreitenden Areale der spezifischen Alpenwörter (Flora, Fauna, Gelände und Ethnographie) mit großer Genauigkeit abgebildet werden können.

Die genannten Unternehmungen unterscheiden sich zwar im Blick auf die dokumentierten Sprachen und die primären sprachwissenschaftlichen Ziele; sie setzen jedoch zwei identische Prinzipien voraus, nämlich einerseits die Verankerung der Sprachgeographie in einer mehrdimensionalen Varietätenlinguistik und andererseits ihre informationstechnologische Modellierung. Eine Letztbegründung beider Prinzipien kann hier nicht geleistet werden, aber im Blick auf das Rahmenthema der Tagung sollen vor allem die durchaus gravierenden methodologischen Implikationen der Online-Publikation und georeferenzierten Online-Kartierung herausgearbeitet werden.

2 Sprachgeographie als mehrdimensionale Varietätenlinguistik

Die Bestimmung der Sprachgeographie als Varietätenlinguistik meint, dass es letztlich um Varietäten, d.h. um Dialekte geht. Diese Feststellung ist keineswegs trivial, denn man darf nicht vergessen, dass Varietäten als solche der direkten Beobachtung nicht zugänglich sind; es handelt sich ja dabei um Abstraktionen über kookkurrierende Varianten, die wiederum Ausprägungen von Variablen sind, welche sich der unmittelbaren Wahrnehmung durch die Wissenschaftler ebenfalls entziehen: Variablen sind funktional definiert und hängen daher vom jeweiligen Beschreibungsmodell ab; so kann, zum Beispiel, *mu* in:

(1) kalabresisch von San Pietro a Maida: *nu juarnu vulissa* ***mu*** *tornu a lu paisi miu* 'un giorno vorrei ritornare al mio paese' (Informant Spi2mIQ1 – ein 16jähriger Schüler)[3]

als Variante einer Konjunktion (Variable) oder in generativer Sicht als Komplementierer (Variable) gefasst werden; für die mit der zweiten Variable identifizierte Kategorie hat die traditionelle Grammatik keine Entsprechung.

[3] Karte: **www.asica.gwi.uni-muenchen.de/index.php?atlante=1&quest=1&frage=31**; Beleg unter: **www.asica.gwi.uni-muenchen.de/index.php?atlante=1&frage=31&qd=Q&details=Spi2mI** (Zugang aus Gründen des Schutzes von Persönlichkeitsrechten passwortgeschützt).

Die eigentlich beobachtbaren empirischen Basisdaten der Varietätenlinguistik sind also weder Variablen, noch Varietäten, sondern nicht standardkonforme Varianten. Solche Varianten sind 'markiert' d.h., sie verweisen auf außersprachlich bestimmte Dimensionen der Variation (den Ort, die soziale Schicht, die Generation, womöglich das Geschlecht und die Situation), oder sie werden durch das gewählte Medium (gesprochen, geschrieben, computervermittelt) konditioniert. Darüber hinaus unterliegen sie auch der individuellen Wahl des Individuums, so dass die varietätenlinguistische Interpretation grundsätzlich schwierig ist. Das Umfeld von Beispiel (1) ist lehrreich; es stammt aus dem Material des AsiCa, für den Informanten zweier verschiedener Generationen und beider Geschlechter befragt wurden (**www.asica.gwi.uni-muenchen.de/index.php?informanti=1**). Im Blick auf den Stimulus von (1), 'un giorno vorrei ritornare al mio paese', liefern die beiden Vertreter der jüngeren Generation in San Pietro a Maida zwei verschiedene syntaktische Lösungen; man vergleiche die folgende, standardnähere Konstruktion mit subordiniertem Infinitiv (und ohne Konjunktion bzw. Komplementierer):

(2) kalabresisch von San Pietro a Maida: *nu juarnu vorisse* ***turnare*** *a lu paisə miu* (Informantin Spi2wIQ1 – eine 17jährige Schülerin)[4]

Die Informanten von (1) und (2) sind fast gleich alt und beide Sekundarschüler; allerdings unterscheiden sie sich im Geschlecht, denn die Äußerung (2) stammt von einer Sprecherin. Eine genauere Analyse der genannten Sprecherin Spi2wIQ1 zeigt allerdings, dass ihr Sprachverhalten im Bezug auf die genannte Struktur keineswegs konsistent ist, denn sie benutzt in Verbindung mit demselben Verb (*volere*) durchaus auch die Konstruktion ohne Infinitiv und mit der Konjunktion *mu*:

(3) *nom volia* ***mu*** *t͡ʃi lu diku* 'non volevo dirglielo' (Informantin Spi2wIQ1)[5]

Eine ausgeprägte, wie es scheint zufällige Variation dieser individuellen Sprecherin zeigt auch eine Gesamtauswertung aller relevanten Inputsätze des Fragebogens.

Variation in der Unterordnung eines Verbs mit der Konjunktion *mu* bei zwei gleichaltrigen Sprechern (Ø = Gebrauch des Infinitivs)	
Spi2mIQ1 (16 Jahre, männlich)	**Spi2mIQ1 (17 Jahre, weiblich)**
F4: Per lavarsi è dovuto uscire fuori.	
mu si llava eppu mu neʃʃa horə	Ø
F10: Comincia a piovere .	
ntʃigna mu kjov	*kumintʃa mu kjova*
F13: Maria se n'è andata senza salutarmi.	
Maria si nda jiu sentsa mu mi salut	Ø
F14: Mio nonno andava a pescare sempre di mattina ...	
'nannuma jia mu peʃka sempe de matin	*'nannuma jia mu piʃka sempre la matina*
F15: Prima di mangiare lavati le mani.	

[4] Karte: **www.asica.gwi.uni-muenchen.de/index.php?atlante=1&quest=1&frage=31**; Beleg unter: **www.asica.gwi.uni-muenchen.de/index.php?atlante=1&frage=31&qd=Q&details=Spi2wI**.

[5] Karte: **www.asica.gwi.uni-muenchen.de/index.php?atlante=1&quest=1&frage=10**; Beleg unter: **www.asica.gwi.uni-muenchen.de/index.php?atlante=1&frage=10&qd=Q&details=Spi2wI**.

prima mu mandʒi lavati li mani	Ø
F16: Si vergogna di uscire di casa .	
si vergogna mu neʃʃa dâ kas	Ø
F17: Mi piace andare in giro con la bicicletta.	
mi piatʃi mu vaju in dʒiru ku la britʃikett	*mi piatʃe mu vaju ddʒirandu ku la bitʃikletta*
F20: Ho dovuto far venire il medico.	
Ø	*eppi mu hattsu mu vena lu miadiku*
F24: Ho sentito strillare qualcuno.	
ntisii . mə . tisi gridari ankunu	Ø
F25: Andavo a lavorare alle sei di mattina.	
jia mu lavoru alli sei dâ mattina	*jia mu lavoru le sei de la matina*
F27: Giuseppe non smette di fumare.	
Peppe nun fina nom fina mu humi	*Ddʒuseppe nom fina mu huma*
F28: Abbiamo cercato di arrangiarci.	
# tʃerkammu mu n arrandʒammu	*tʃerkammə mu n arrandʒamu*
F29: Sai nuotare bene?	
sa mmu nati buanu ?	Ø
F31: Lascialo mangiare.	
ɖɖassalu mu manddʒ	*ɖassalu mu manddʒa*
F33: Non volevo dirglielo.	
nun vulia mu ntʃi lu ddik	*nom volia mu tʃi lu diku*
F43: Gianni mi ha chiesto se volevo scendere in Calabria quest'anno.	
Ddʒanni mi kjese se volia mu ʃindu n Kalabbria …	*Ddʒanni mi kjese si vogliu mu vaju in Kalabbria…*
F45: È salito sull'albero per cogliere i fichi.	
sagghjiu supra l alberu mu koggja li fhiku	Ø
F50: Domani vado alla posta per spedire la lettera.	
domani vaju a la posta mu ʃpediʃʃu la lettərə	Ø
F54: Sono troppo stanco/a per uscire stasera.	
su ʈroppu stanku mu niaʃʃu stasira	Ø
total: 18 (*mu*) : 1 (Infinitiv)	**total**: 10 (*mu*) : 9 (Infinitiv)

Abbildung 1: AsiCa-Korpus – synoptische Darstellung des variierenden Gebrauchs der Konjunktion *mu* zur Unterordnung eines Verbs bei zwei gleichaltrigen Sprechern

Soll man aber ausgehend von diesem Befund darauf schließen, dass jüngere Frauen in San Pietro a Maida grundsätzlich eine stärker italianisierte Varietät des Ortsdialekts sprechen? Wie es scheint, ist die Annahme, der Gebrauch einer Variante sei grundsätzlich und in verlässlicher Weise indexikalisch in Bezug auf außersprachliche Gegebenheiten wie hier im Sinne einer 'diasexuellen Kovariation' auf dieser Datenbasis problematisch; vermutlich würde auch eine Vermehrung der SprecherInenn keine wirklich Klarheit bringen. Vielmehr lässt sich der varietätenlinguistische 'Wert' einer Variante im aktuellen Gebrauch ausschließlich auf Grundlage von Sprachproduktionsdaten nicht zuverlässig ermitteln. Denn der kommunikative Mehrwert einer Variante besteht in den Wissensbeständen, oder: mentalen Repräsentationen, die der Sprecher mit ihnen assoziiert. Es wäre zwar naiv, diesem individuellen ('subjektiven') Sprecherwissen objektive Gültigkeit zuzusprechen; nichtsdestoweniger steuert es den Sprachgebrauch des Individuums und womöglich seine Tendenz, sich an andere SprecherInnen zu akkomodieren. Die Varietätenlinguistik im Allgemeinen und die Dialektologie im Speziellen müssen daher systematisch auch Perzeptionsdaten, genauer: Auto- und Heteroperzeptionsdaten

erheben.[6] Unerlässlich sind Perzeptionsdaten, um mehrfache Markierungen ('gesprochen' + 'dialektal' + 'sozial' usw.) und Markierungsverschiebungen bzw. –verluste zu erfassen (im genannten Beispiel wäre etwa die Markiertheit von *mu* bzw. des Infinitivs zu klären). Es wäre also zu fragen, ob *mu* unter Gleichaltrigen über das Diatopische hinaus als typisch 'männlich' markiert ist, und ob der Infinitiv inzwischen als unauffällige diatopische Variante akzeptiert wird usw. Entsprechende Hinweise hat die traditionelle Dialektologie nur sporadisch aufgenommen, wenn sie von ihren meistens singulären Informanten spontan geäußert wurden; punktuell interessante, aber nur mühsam zu findende Beispiele geben die Legenden des AIS. Auch die Inputdaten der oben erwähnten Projekte (i)-(iii) und (v) geben dergleichen nicht her; es ist allerdings möglich und unbedingt wünschenswert, sie auf Grundlage der Online-Publikationen zukünftig mit entsprechenden Perzeptionsdaten anzureichern.

3 Sprachgeographie als Informationstechnologie

3.1 Allgemeine technische Aspekte

Die meisten der in Zusammenarbeit zwischen der romanischen Sprachwissenschaft und der IT-Gruppe Geisteswissenschaften (ITG; <**www.itg.uni-muenchen.de**>) der LMU entwickelten Projekte auf dem Gebiet der Geolinguistik mussten mit vergleichsweise geringer personeller und finanzieller Ausstattung realisiert werden. Dieser Umstand führte zwangsläufig zur Wahl kostengünstiger technischer Mittel und effizienter wissenschaftlicher Methoden, die sich in der Folge im praktischen Einsatz sehr bewährt haben.

Abgesehen von Metropolitalia, basieren alle aus der genannten Kooperation hervorgegangenen Projekte auf dem Zusammenspiel einer MySQL-Datenbank mit einem PHP-Modul. Während Letzteres für die Erzeugung von HTML-Seiten zuständig ist, die über das Internet an theoretische beliebig viele Clients ausgeliefert werden, befinden sich die eigentlichen Projektdaten in der MySQL-Datenbank. Das PHP-Modul greift auf die Datenbank zu und bindet die Projektdaten variabel, d.h. in Abhängigkeit von der Anfrage des Nutzers im Internet, in die HTML-Seiten ein:

[6] Die Literatur zur perzeptiven Linguistik ist mittlerweile stark angewachsen; zahlreiche Angaben finden sich in den grundlegenden Monographien von Postlep 2010, und Purschke 2011 sowie die Beiträge in Krefeld/Pustka (Hrsg.) 2010.

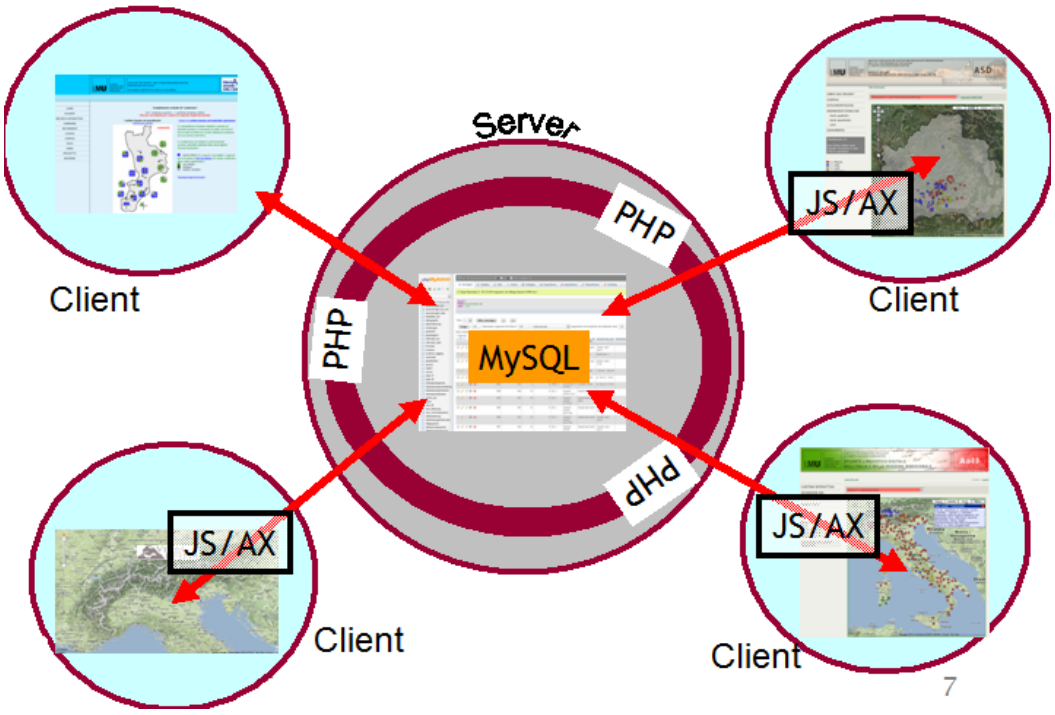

Abbildung 2: Client-Server-Prinzip

Für die dynamische Veränderung des Inhalts der interaktiven Karten kommt clientseitig außerdem die sog. Javascript(JS)/AJAX(AX)-Technologie zum Einsatz, die es erlaubt, nur die Daten vom Server nachzuladen, die für die Einbindung neuer Inhalte in die Grundkarte erforderlich sind, ohne dass die komplette Webseite neu geladen werden muss.

Das vorgestellte Konzept ist auf der Client-Seite (also beim Nutzer im Internet) weitgehend plattformunabhängig, d.h. die Inhalte der Projektseiten können gleichermaßen auf Windows-, Macintosh- und Unix-Rechnern dargestellt werden. Gewisse Einschränkungen bestehen aktuell noch hinsichtlich der einzusetzenden Browser, da die Anpassung an die verschiedenen Programme sehr zeitaufwendig ist. Manche Browser stellen die Inhalte z.B. mit fehlerhaftem Layout dar oder haben Probleme mit der Wiedergabe eingebundener Sounds. Eine weitgehend fehlerfreie Darstellung der Projektwebseiten kann derzeit nur mit dem Browser Firefox garantiert werden. Bewusst verzichtet wurde auf die Entwicklung von spezifischen Client-Programmen, die auf den Computern der Nutzer installiert werden müssen, um auf die Datenbankinhalte zugreifen zu können. Für die Wiedergabe von Audiodateien sind jedoch geeignete Browser-Plugins (Flashplayer oder Quicktime) erforderlich. Die Version 5 des HTML-Standards (offizielle Verabschiedung für 2014 geplant) wird künftig gestatten, auf diese Plugins zu verzichten. Die standardmäßig eingesetzte Google-Technologie für die interaktive Kartographie verlangt die Aktivierung von Javascript im Browser.

3.2 Datenstrukturierung

Während die technische Seite sowohl konzeptionell wie auch im praktischen Betrieb, der durch die IT-Gruppe Geisteswissenschaften der LMU unterstützt wird, so gut wie keine

Probleme bereitet, bestehen die eigentlichen Herausforderungen in der Modellierung der zu verarbeitenden Daten. Das gilt vor allem dann, wenn Inputdaten aus mehreren, konzeptionell unterschiedlichen Quellen zusammengebracht werden sollen, wie es das geplante Projekt (v), *Verba Alpina*, vorsieht. Es stellt sich hier z.B. bei der Analyse der Heteronyme des Begriffs 'Sennhütte' das Problem, dass die zur Verfügung stehenden Sprachdaten je nach Quelle unterschiedliche Qualität aufwiesen: Während für den italienischsprachigen Alpenraum die Daten des AIS zur Verfügung standen, die – der romanistischen Tradition der Sprachwissenschaft folgend – jeweils den unmittelbaren Einzelbeleg greifbar machen, war das Datenmaterial des *Vorarlberger Sprachatlas* (VALTS) mit diesem nur bedingt vergleichbar, da dort das Sprachmaterial ausgehend von Punktsymbolkarten nur in typisierter Form, d.h. ohne Dokumentation der Einzelbelege, vorliegt. Während man bei der Analyse solch inkongruenter Daten problemlos einen Ausgleich dadurch schaffen kann, dass die Einzelbelege von der Art des AIS ihrerseits typisiert werden, so bleibt dennoch das Problem, bei der Speicherung der Daten in einer Datenbank ihre unterschiedliche empirische Qualität zu dokumentieren.

Für alle der hier vorgestellten Corpus-basierten Projekte wird grundsätzlich nach einem einheitlichen Verfahren vorgegangen. Sämtliches Sprachmaterial wird zunächst in elektronisch verarbeitbaren Text verwandelt. Dabei wird zwischen verschiedenen Graden der Digitalisierung unterschieden, was folgende Graphik illustriert:

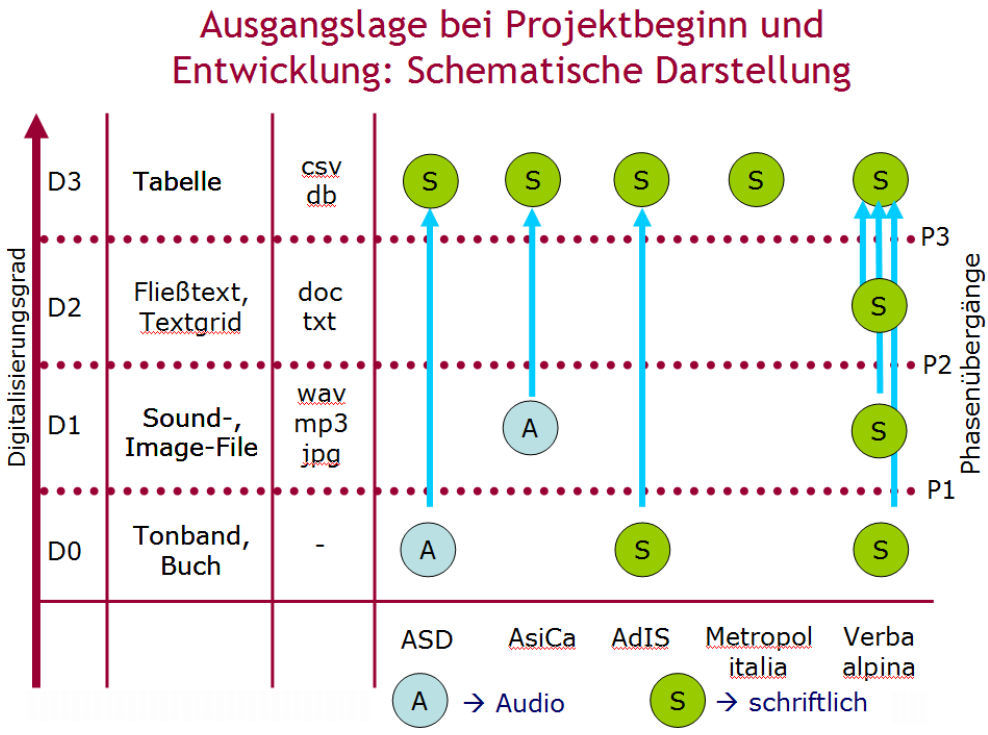

Abbildung 3: Digitalisierungsgrade und -phasen

Unter Erzeugung eines konsistenten Referenzsystems, das das spätere Wiederauffinden beliebiger Textteile im Gesamtkorpus garantiert, wird der elektronische Fließtext stufenweise in immer kleinere Einheiten zerlegt, wobei in den meisten Fällen als kleinste Einheit am Ende dieses Prozesses jeweils das 'Token' steht. In der Sprache der Informa-

tik ist jede alphanumerische Zeichenfolge zwischen Separatoren, unabhängig von ihrer Frequenz, ein 'individuelles' Token, wobei im Fall der Zerlegung eines Fließtextes als Separator in aller Regel das Spatium betrachtet wird. Als Ergebnis dieses Prozesses liegt jeweils eine Tabelle vor, deren Zeilen untereinander die Tokens des Fließtexts enthalten, wobei jedem Token in getrennten Spalten die Werte des jeweiligen Referenzsystems zugeordnet sind:[7]

interview	intervall	position	Sprecher	token
Acc1wlQ1	31	1	informante	'nannuma
Acc1wlQ1	31	2	informante	jia
Acc1wlQ1	31	3	informante	mu
Acc1wlQ1	31	4	informante	piʃka
Acc1wlQ1	31	5	informante	sempre
Acc1wlQ1	31	6	informante	i
Acc1wlQ1	31	7	informante	matina

Abbildung 4: Tokenisierung I

Das vorliegende Beispiel stammt aus dem AsiCa-Corpus und stellt die Antwort des Informanten Acc1wIQ1 auf den Stimulus F14: *Mio nonno andava a pescare sempre di mattina* dar. Das Tabellenschema bietet die Möglichkeit, jeder Zeile im Prinzip beliebig viele Spalten (auch: 'Attribute') hinzuzufügen, um auf diese Weise strukturiert weitere Daten ('Tags'), wie z.B. die Wortart jedes Tokens, anzufügen:

interview	intervall	position	Sprecher	token	Wortart
Acc1wlQ1	31	1	informante	'nannuma	Nome di parentela + poss. enclit.
Acc1wlQ1	31	2	informante	jia	verbo
Acc1wlQ1	31	3	informante	mu	Congiunzione finale
Acc1wlQ1	31	4	informante	piʃka	verbo
Acc1wlQ1	31	5	informante	sempre	Awerbio temporale
Acc1wlQ1	31	6	informante	i	Preposizione
Acc1wlQ1	31	7	informante	matina	Nome comune

Abbildung 5: Etikettierung

Die Funktionalität der Datenbank gestattet es, auf Basis des Referenzsystems jederzeit den ursprünglichen Fließtext wiederherzustellen, auch unter Integration in der Datenbank vorgenommener Ergänzungen wie z.B. im gegebenen Beispiel der Wortart:

(4) Acc1wIQ1, 31: *'nannuma* (Nome di parentela + poss. enclit.) *jia* (verbo) *mu* (Congiunzione finale) *piʃka* (verbo) *sempre* (Avverbio temporale) *i* (Preposizione) *matina* (Nome comune)

[7] Die Kombination der Werte in den Feldern `interview`, `intervall` und `position` fungiert dabei als zusammengesetzter Primärschlüssel. Datenbankintern ist überdies jedes Token mit einer eindeutigen Identifikationsnummer ('ID') versehen.

Das vorgestellte Beispiel illustriert ein Problem, das bei der Segmentierung von Fließtexten immer wieder auftritt: Das Token *'nannuma*, Ergebnis der 'Tokenisierung' nach Maßgabe des Spatiums als Separator, stellt ein Konglomerat aus einem Substantiv, konkret einer Verwandtschaftsbezeichnung, *nannu* (*nonno*) und einem als Enklitikon realisierten Possessivpronomen (*ma*) dar. Es gibt verschiedene Möglichkeiten, mit Dergleichen umzugehen; eine Lösung ist die im Exemplum vorgestellte, nämlich die Definition einer speziellen Wortart, die dann dem Token zugewiesen wird. Diese Lösung ist mit diversen Nachteilen behaftet; geschickter erscheint, bei der Tokenisierung mit einem erweiterten Set von Separatoren, nämlich solchen erster und solchen zweiter Ordnung zu arbeiten. Als Separatoren erster Ordnung gelten weiterhin die Spatien, als Separator zweiter Ordnung wird, z.B., das Gleichheitszeichen (=) eingeführt. Der Fließtext müsste also in folgender Weise vorbereitet werden:

(5) 'nannu=ma jia mu piʃka sempre i matina

Die Abbildung in einer Tabelle sähe dann folgendermaßen aus:

interview	intervall	position	subtoken	Sprecher	token	wortart
Acc1wlQ1	31	1	1	informante	'nannu	Nome di Parentela
Acc1wlQ1	31	1	2	informante	ma	poss. enclit.
Acc1wlQ1	31	2	1	informante	jia	verbo
Acc1wlQ1	31	3	1	informante	mu	Congiunzione finale
Acc1wlQ1	31	4	1	informante	piʃka	verbo
Acc1wlQ1	31	5	1	informante	sempre	Awerbio temporale
Acc1wlQ1	31	6	1	informante	i	Preposizione
Acc1wlQ1	31	7	1	informante	matina	Nome comune

Abbildung 6: Tokenisierung II (Verwendung von Separatoren unterschiedlicher Ordnung)

Die Information, dass *ma* ein Enklitikon von *'nannu* ist, wird durch die Kombination der beiden Spalten 'position' und 'subtoken' - letztere wurde eigens zu diesem Zweck zusätzlich eingefügt - kodiert: Beide Tokens besitzen denselben Wert in der Spalte ‚'position', sind jedoch durch den Wert in der Spalte 'subtoken' voneinander unterschieden, wobei der dort eingetragene Wert gleichzeitig die Abfolge der Subtokens festlegt. Auch in diesem Fall ist eine – wenn gewünscht, auch annotierte – Re-Synthetisierung des ursprünglichen Fließtextes jederzeit möglich:

(6) *'nannu*(Nome di Parentela)=*ma*(poss. enclit.) *jia*(verbo) *mu*(Congiunzione finale) *piʃka*(verbo) *sempre*(Avverbio temporale) *i*(Preposizione) *matina*(Nome comune)

Die vorgestellte Datenstrukturierung erlaubt unter anderem das problemlose Auffinden sämtlicher Enklitika, indem nach Datensätzen gesucht wird, die im Feld 'subtoken' den Wert "2" aufweisen. Ein Nachteil besteht allerdings darin, dass der Fließtext vor der Konvertierung in das Tabellenformat zunächst durch Einfügen der Separatoren zweiter Ordnung entsprechend vorbereitet werden muss. Sofern dies nicht schon bei einer allfälligen Transkription geschehen ist, ist dafür zumeist einiger Aufwand erforderlich.

Wie erwähnt, gestattet die Tabellenstruktur die Assoziation einer im Grunde beliebigen Anzahl von Attributen in Form weiterer Spalten. Ausgehend von der Basis-Entität des Tokens, ergeben sich nahezu zwingend die morphosyntaktischen Kategorien des Wortes als weitere Attribute:

interview	intervall	wort	Wortart	person	numerus	modus	tempus	genus	lemma
Acc1wlQ1	31	'nannuma	NParPoss	1	sg			m	nonno
Acc1wlQ1	31	jia	V	3	sg	ind	impf		ire
Acc1wlQ1	31	mu	C1						mu
Acc1wlQ1	31	piʃka	V	3	sg	ind	prs		pescare
Acc1wlQ1	31	sempre	AVtem						sempre
Acc1wlQ1	31	i	Prep						di
Acc1wlQ1	31	matina	N		sg			f	mattina

Abbildung 7: Erweiterte (morphosyntaktische) Etikettierung

Das Beispiel führt auch das angesprochene Segmentierungsproblem vor Augen. Im Fall des Tokens *'nannuma* passt zwar der Eintrag im Feld 'Wortart', die anderen Attribute können sich aber jeweils nur auf eines der beiden Teil-Tokens beziehen. Die Zuweisung des Tokens zum Lemma *nonno* unterschlägt das gleichzeitig gegebene Lemma *mio*, was Auswirkungen auf die Ergebnisse von Datenanalysen haben kann.

Grundsätzlich eröffnet die Anlagerung von Metadaten natürlich die gezielte Datenanalyse auf dieser Metaebene. So lassen sich z.B. problemlos Äußerungseinheiten mit jeweils mehr als einem finiten Verb finden (vgl. oben Beispiel [1]) oder auch Phänomene der Wortstellung bzw. Syntax analysieren.

3.3 Georeferenzierung

Die Verwendung der Tabellenstruktur gestattet die Verknüpfung von Daten mit einer beliebigen Anzahl von weiteren Daten. Voraussetzung ist lediglich, dass diese in einem unmittelbaren logischen Zusammenhang stehen. Auf diese Weise wird die Erzeugung eines Datennetzes ermöglicht, das z.T. höchst unterschiedliche Datenobjekte einbindet, und deren vielfältige mittelbaren Zusammenhänge abbildet. Der Georeferenzierung kommt dabei eine zentrale Rolle zu, kann sie doch die Brücke sein, über die sich solche mittelbaren Abhängigkeiten zwischen scheinbar ihrer Qualität nach unvereinbaren Daten herstellen lassen (s. unten Abbildung 17). In den oben vorgestellten Projekten aus dem Bereich der Geolinguistik spielt die Georeferenzierung in durchaus unterschiedlicher Weise eine Rolle.

Sowohl der ASD als auch der AdIS basieren auf Daten, die bei Projektbeginn bereits im analogen Sinn lokalisiert waren. Mit Lokalisierung ist dabei im Wortsinn eine Verortung gemeint, die eine mündliche oder schriftliche Äußerung einem Herkunftsort durch Nennung dessen Namens oder die Eintragung an einem bestimmten Punkt auf einer herkömmlichen 'analogen' Landkarte zuordnet:

nr	ort	Aeusserung
100	Neppendorf	æm ˈvæŋtər ˈflæjən də gəˈdrɛçt ˈblædər æn dər ˈlaft əˌræm
95	Neppendorf	o͡as ɪm ˈvɪntə ˈfli͡agŋ diː ˈtrɪkərn ˈpleːtʃn ɪ də ˈluft ʊməˌtʊm

Abbildung 8: ASD-Korpus, phonetische Transkriptionen des Wenkersatzes 1, Informanten aus Neppendorf in Siebenbürgen

Das Beispiel zeigt die Transkription des Wenkersatzes Nummer 1 ('Im Winter fliegen die getrockneten Blätter in der Luft herum') aus einer entsprechenden Audio-Aufnahme, die

in den sechziger/siebziger Jahren des vergangenen Jahrhunderts im siebenbürgischen Ort Neppendorf entstanden ist. Der Schritt von der Lokalisierung zur Georeferenzierung erfolgt durch die Ermittlung der Geokoordinaten des Ortes Neppendorf[8]. Im Internet steht zu diesem Zweck eine Vielzahl von Diensten zur Verfügung, bisweilen existieren auch schon georeferenzierte Ortslisten einzelner Regionen oder Länder. Die ermittelten Geokoordinaten werden den Sprachdaten als weitere Attribute hinzugefügt:

nr	ort	Breite	Länge	Aeusserung
100	Neppendorf	45.788	24.116	æm ˈvæŋtər ˈflæjən də gəˈdrɛçt ˈblædər æn dər ˈlaft əˌræm
95	Neppendorf	45.788	24.116	o͡as ɪm ˈvɪntə ˈfli͡əgŋ diː ˈtrɪkərn ˈpleːtʃn ɪ də ˈlʊft ʊməˌtʊm

Abbildung 9: ASD-Korpus, Georeferenzierung

Analyseergebnisse lassen sich nun auf einer georeferenzierten elektronischen Karte an den jeweils zugeordneten Koordinaten darstellen. Um beim gegebenen Beispiel zu bleiben: Man beobachtet, dass die im ASD-Korpus repräsentierten Informanten bei der Wiedergabe des Wenkersatzes 1 zwei Varianten als Entsprechung für das PPP "*getrockneten*" verwenden. Während die einen das PPP beibehalten, gebrauchen andere das Adjektiv "*trockenen*". Geeignete Datenbankabfragen ermitteln alle Belege, die zur einen bzw. zur anderen der genannten Gruppen gehören. Zusammen mit den Belegen liefern die Datenbankabfragen die Namen und Koordinaten der Herkunftsorte der Informanten (vgl. Abbildung 10 und Abbildung 11).

Die Ergebnisse dieser analytischen Datenbankabfragen können nun kontrastiv auf einer georeferenzierten elektronischen Karte dargestellt werden (vgl. Abbildung 12).

variante	ort	Breite	Länge	token	Wenkersatz
A	Neppendorf (100)	45.79	24.12	gəˈdrɛçt	æm ˈvæŋtər ˈflæjən də gəˈdrɛçt ˈblædər æn dər ˈlaft əˌræm
A	Almen (1004)	46.05	24.43	gəˈdrɛçt	æm ˈvæŋtər ˈflɛjən də gəˈdrɛçt ˈbladər æn dər ˈlɔft əˌram
A	Almen (1005a)	46.05	24.43	gəˈdrɛçt	æm ˈvæŋtər ˈflɛjən də gəˈdrɛçt ˈbladər æn dər ˈlɔft əˌram
A	Kleinschelken (1016b)	46.05	24.14	gəˈdrɛçt	æm ˈvæŋtər flæjən də gəˈdrɛçt ˈbladər æn dər ˈlɔft əˌram
A	Kleinschelken (1018b)	46.05	24.14	gəˈdrɛçt	æm ˈvæŋtər flæjən də gəˈdrɛçt ˈbladər æn dər ˈlaft əˌram
A	Schorsten (1020b)	46.03	24.06	gəˈdrɛçt	ɛm ˈvaĩŋtər flɛjən də gəˈdrɛçt ˈbladər ɛn dər ˈlɔft ˈəram
A	Donnersmarkt (1032a)	46.14	23.97	gəˈdrɛçt	æm ˈvæŋt͡ʃər flæjən də gəˈdrɛçt ˈblædər æn dər ˈlɔft əˌram

Abbildung 10: Analyseergebnis Wenkersatz 1 – Variante A: *getrockneten* (Ausschnitt)

variante	ort	Breite	Länge	token	Wenkersatz
B	Tobsdorf (1156a-04)	46.15	24.50	ˈdrɛç	æm ˈvɛŋtər flæjən də ˈdrɛç ˈblader æn dər ˈlɔft əˌram
B	Heltau (118-03)	45.72	24.15	ˈdrɛç	æm ˈvæŋtər ˈflæjən də ˈdrɛç ˈbladər æn dər lɔft əˈræm
B	Blutroth (1198a-04)	46.08	23.74	ˈdriːɛ͡ç	am ˈva͡ɲt͡çər ˈflajən də ˈdriːɛ͡ç də gəˈdriːɛ͡çt ˈbladər an dər ˈlɔft əˌram
B	Michelsberg (120-05)	45.70	24.11	ˈdɛr	æm ˈvæŋtər ˈflæjən də ˈdɛr ˈbladər fʊn də ˈbiːmən æn dər ˈlɔft əˈræm
B	Michelsberg (120-05)	45.70	24.11	ˈdɛr	æm ˈvæŋtər ˈflæjən də ˈdɛr ˈbladər æn dər ˈlɔft əˈræmər
B	Rumes (1217-05)	45.85	23.27	ˈdrɛç	æm ˈvæŋtər flæjən də ˈdrɛç ˈblædər æn dər ˈlæft əˌræm
B	Michelsberg (127)	45.70	24.11	ˈdɛr	æm ˈvæŋtər ˈflæjən də ˈdɛr ˈbladər æn dər ˈlɔft əˌræmər
B	Hermannstadt (135-02)	45.80	24.15	ˈdrɛç	æm ˈvæŋtər ˈflæjən də ˈdrɛç ˈblædər æn dər ˈlaft əˌræm

Abbildung 11: Analyseergebnis Wenkersatz 1 – Variante B: *trockenen* (Ausschnitt)

[8] Die beiden Belege in Abb. 9 zeigen übrigens sehr schön das Nebeneinander von Sachsen ('nr 100') und ursprünglich aus dem Salzburgischen stammenden Landlern ('nr 95') im Aufnahmeort Neppendorf.

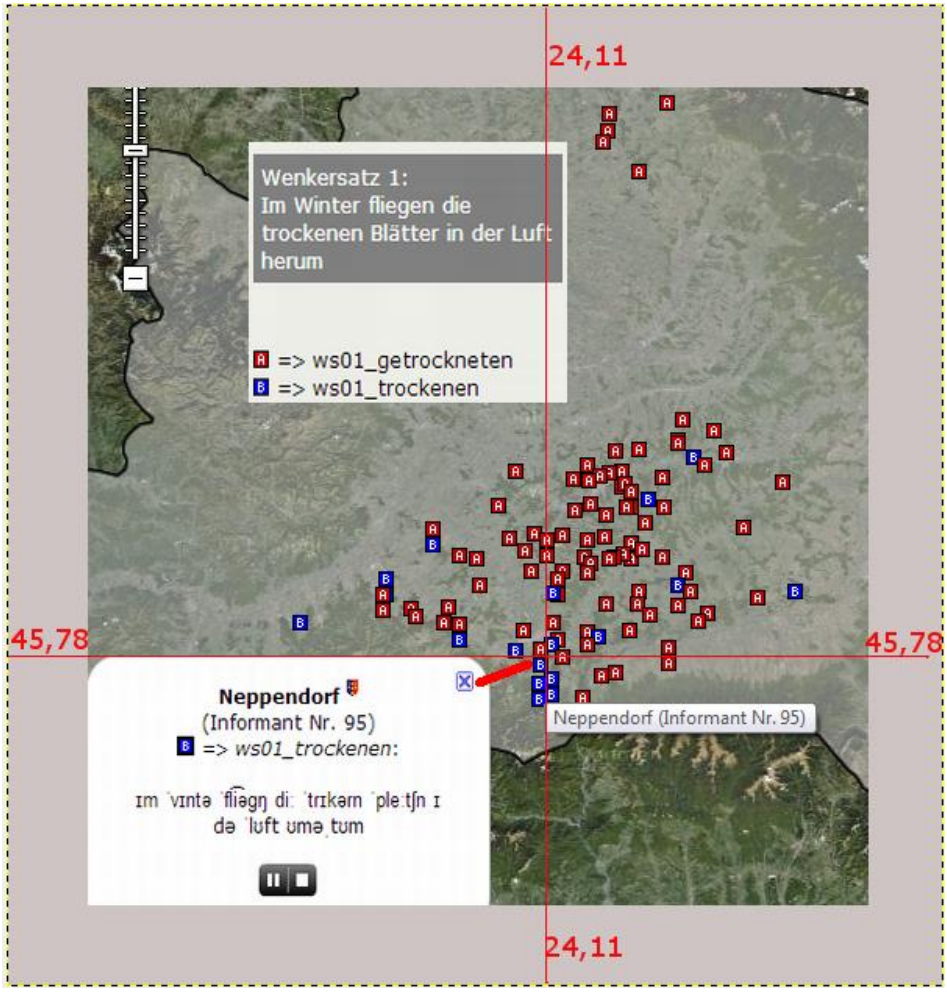

Abbildung 12: Abbildung georeferenzierter Korpusdaten auf einer Google-Karte

Sowohl der ASD wie auch AdIS und Metropolitalia verwenden derzeit den Kartendienst Google Maps für die Erzeugung der elektronischen Karten. Aufgrund sich abzeichnender Veränderungen in den Nutzungsbedingungen der Firma Google wird aktuell an einer Umstellung des Systems auf die Verwendung gemeinfreier elektronischer Karten des OpenStreetMap-Projekts gearbeitet (**www.openstreetmap.org**). Es ist zu betonen, dass es aus Projektsicht vollkommen unerheblich ist, welche Quelle die elektronische Grundkarte zur Verfügung stellt. Solche Umstellungen haben konzeptionell keinerlei Relevanz und sind mit vergleichsweise geringem Arbeitsaufwand verbunden.

Wie bereits angedeutet, besitzt die Georeferenzierung vor allem deswegen Attraktivität, weil durch ihre Vermittlung auf den ersten Blick zusammenhanglose Informationen in sinnvoller Weise miteinander verknüpft werden können. Im Sinne der heuristischen Methode können im Grunde beliebige Daten miteinander kombiniert werden. Voraussetzung ist lediglich, dass sie georeferenzierbar sind. Natürlich empfiehlt es sich, zunächst Daten miteinander zu kombinieren, von denen a priori anzunehmen ist, dass sie mit dem primären Datenbestand in irgendeiner Beziehung stehen. Illustrativ ist das folgende sprachgeschichtliche Beispiel.

Es ist seit langem bekannt, dass die Glottogenese des Italienischen in die Epoche der Spätantike bzw. des frühen Mittelalters fällt und im Zusammenhang mit den großen Migrationsbewegungen dieser Zeit gesehen werden muss. Daher lag es nahe, den Sprachdaten speziell des AdIS georeferenzierbare Daten aus eben dieser Zeit gegenüberzustellen. Bislang sind die Bemühungen in diese Richtung noch nicht über das Versuchs- bzw. Anfangsstadium hinausgekommen, jedoch können erste Ergebnisse durchaus als *proof of concept* angesehen werden.

Zur Illustration sei hier eine Karte abgebildet[9], die synoptisch die Belege für die Verwendung von italienischen Wörtern langobardischen Ursprungs, speziell des Wortes *guancia* 'Wange', zusammen mit den Fundorten langobardischer Gräberfelder sowie dem Auftreten von Ortsnamen langobardischen Ursprungs darstellt (vgl. Abbildung 13).

Die divergierende Schwerpunktbildung zwischen sprachlichen und archäologischen Belegen springt ins Auge. Der sprachwissenschaftlichen Relevanz dieses Bildes soll hier nicht weiter nachgegangen werden. Das Beispiel zeigt – und darauf kommt es hier an – die besondere Bedeutung der Kartierung strukturierter Daten: Gerade im Hinblick auf die heuristische Methode ist die Abbildung der Daten auf Karten eine unverzichtbare Ergänzung der Analysemöglichkeiten in einer Datenbank. Im Sinne der Heuristik ist geplant, den Nutzern unserer Online-Atlanten weitreichende Freiheiten bezüglich der Auswahl der synoptisch darzustellenden Daten zu gewähren.

Neben der Sammlung georeferenzierter Daten zu den Fundorten langobardischer Gräberfelder wurde bislang mit der Georeferenzierung der auf der sog. Tabula Peutingeriana fassbaren Informationen begonnen. Bei der Tabula Peutingeriana handelt es sich sehr wahrscheinlich um eine Kopie (12. Jh.) eines spätrömischen Itinerars oder gar einer Landkarte, die die Verhältnisse vermutlich des 4./5. Jahrhunderts n. Chr. widerspiegelt. Die aus der Tabula gewonnenen Daten stammen demnach ungefähr aus der frühesten Epoche der Glottogenese des Italienischen und sind daher interessante Kandidaten für eine Kontrastierung mit dem vorhandenen Sprachmaterial. Des Weiteren ist geplant, systematisch georeferenzierte und interpretationsfreie Daten zu Kommunikationswegen (Fundorte römischer Meilensteine, Passheiligtümer, antike und nach-antike Straßenverläufe, Patrozinien etc.) zu erfassen und auf diese Art und Weise, *bottom up*, Regionen bzw. Grenzen politischer und administrativer Einflusssphären (z.B. sich aus der Bistumszugehörigkeit einzelner Pfarreien abzeichnende Bistumsgrenzen) deutlich werden zu lassen. Bei all dem wird stets versucht, dem Nutzer einen oder mehrere Belege/Quellen für die Authentizität der vorgenommenen Verortung zu präsentieren. Im Fall der Sprachdaten äußert sich dieses Bestreben in der Angabe des Originalbelegs, wobei, wie im Fall des AsiCa und des ASD, auch das Anhören von Tonaufnahmen möglich ist. Die so entstehenden 'qualitativen' Karten stellen gleichsam eine Synthese aus den, traditionell in der Germanistik verbreiteten, Punktsymbolkarten (z.B. VALTS) und den, der romanistischen Tradition folgenden, Sprachkarten mit der Angabe des Einzelbelegs (z.B. AIS) dar. Die Karten liefern ein prägnantes, auf Klassifizierung beruhendes Bild durch die Verwendung unterschiedlicher Punktsymbole und machen gleichzeitig durch die jeweilige Angabe des Originalbelegs die Klassifizierung nachvollziehbar und transparent. Somit

9 Für die Erstellung der Karte danken wir Helene Eichwald und Miriam Schwemmlein.

verbinden die elektronischen Karten die Vorteile beider traditioneller Kartentypen, was ohne den Einsatz der modernen Computertechnologie nicht möglich gewesen war.

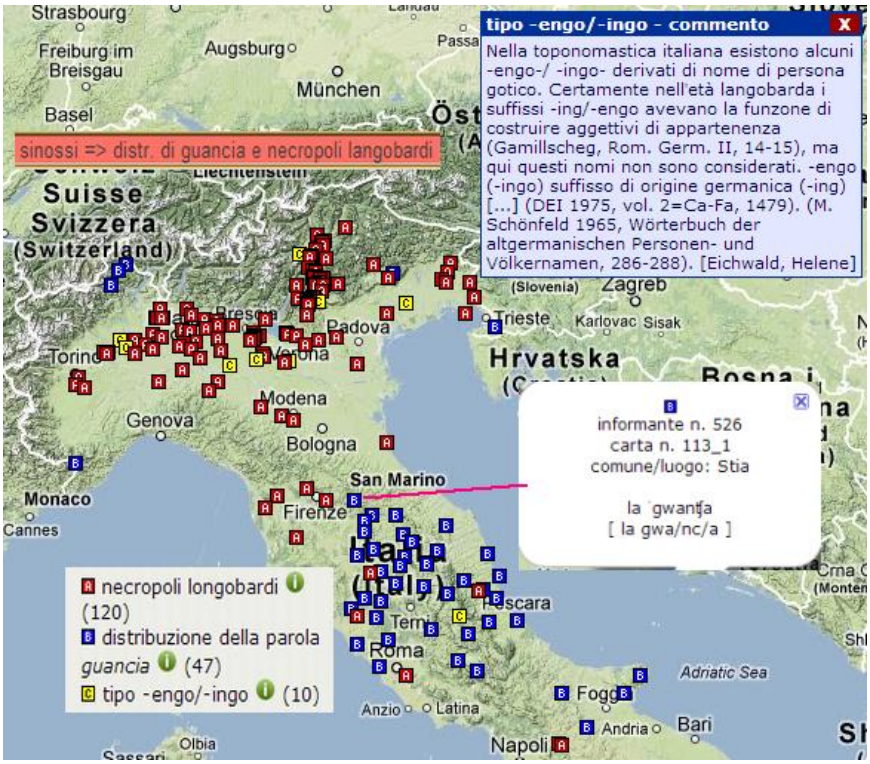

Abbildung 13: Synoptische Darstellung von Daten unterschiedlicher Art (Sprache, Archäologie)

Neben diesen qualitativen Karten, die die Summe einer Vielzahl von in der Fläche verteilten Einzelbelegen abbilden, ist auch die Erzeugung von 'quantitativen' Karten möglich, deren Symbole die Kumulation mehrerer Einzelbelege an einem Ort durch variable Größe und Farbgebung anzeigen. Die folgende Karte[10] stellt die Kumulation von sprachlichen Langobardismen im Bestand des AIS dar:

[10] Auch für die Erstellung dieser Karte danken wir Miriam Schwemmlein.

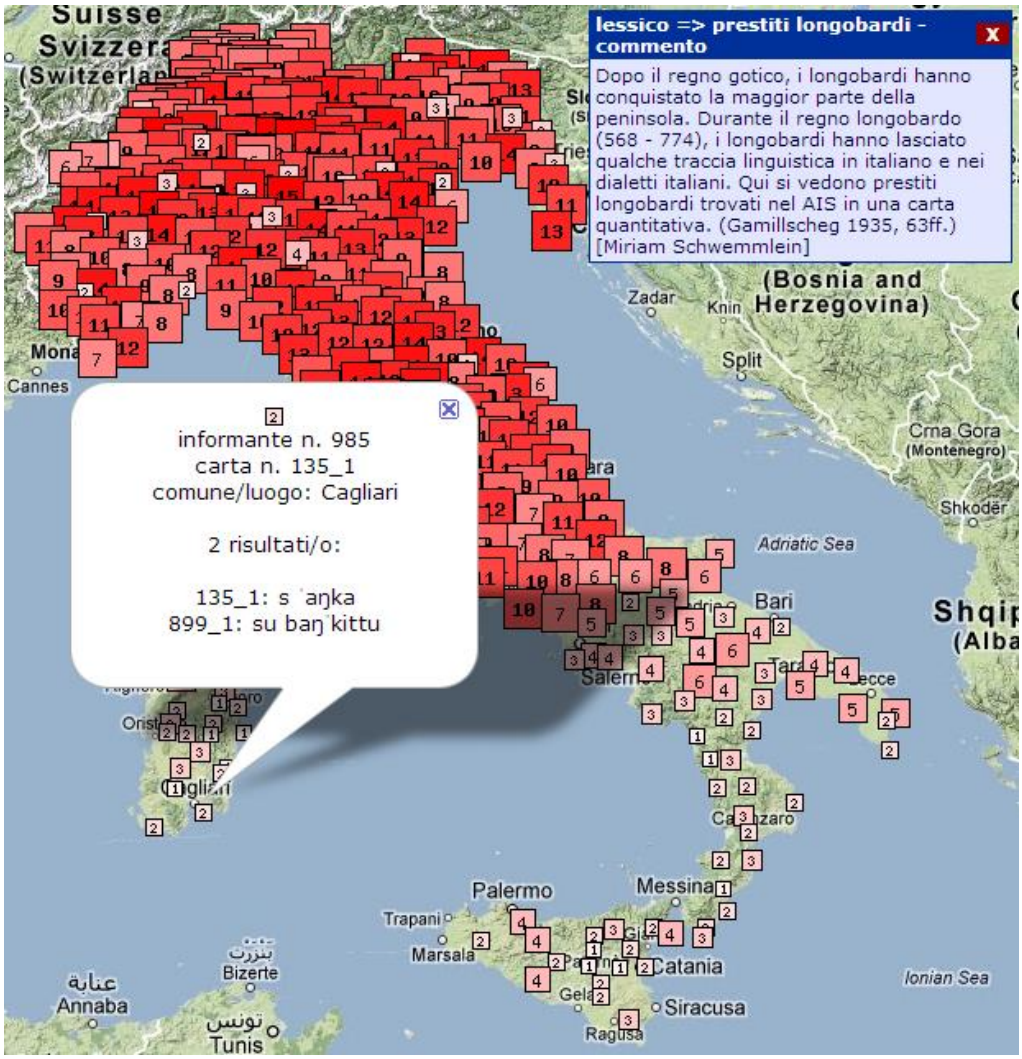

Abbildung 14: Kumulierende Darstellung von Analyseergebnissen auf 'quantitativen' Karten (Beispiel: Anzahl von Langobardismen in den Aufnahmeorten des AIS)

Interessant ist, wie sich die Massierung langobardischer Sprachreste in etwa mit dem historisch bezeugten Ausdehnungsgebiet des Langobardenreiches deckt; der punktuelle Eindruck, den die Abbildung 13 vermittelte, wird durch den sich ergebenden, konsistenten Gesamteindruck grundsätzlich relativiert.

Gezielte Datenanalyse, auch im Sinne heuristischer Ansätze, kann auch direkt in der Datenbank selbst vorgenommen werden, allerdings erfordert der Umgang mit der für relationale Datenbanken spezifischen Abfragesprache SQL (‘structured query language’) ein gewisses Maß an Einarbeitung. Das MySQL-Datenbank-Management-System verfügt über spezielle Datentypen und Funktionen, die für die Speicherung und Verarbeitung georeferenzierter Daten wie Punkte, Linien/Pfade und Flächen optimiert sind. In der aktuellen Version bestehen noch gewisse Einschränkungen, so basieren sämtliche geometrischen Berechnungen auf der planaren (euklidischen) Geometrie. Die sich daraus ergebenden Berechnungsungenauigkeiten sind bei den, bezogen auf die Gesamtgröße der Erdkugel, kleinräumigen Analysen im Rahmen unserer Projekte jedoch vernachlässigbar.

Zur Distanzmessung wurden überdies eigene Funktionen entwickelt, die die Erdkrümmung berücksichtigen und hinreichend genaue Ergebnisse liefern. Es ist außerdem zu erwarten, dass die künftigen Versionen von MySQL Berechnungen im Bereich der sphärischen Geometrie unterstützen werden.[11]

Die folgende SQL-Abfrage ermittelt alle bislang in der Datenbank erfassten Funde langobardischer Gräberfelder auf dem Gebiet der Region Trentino-Alto-Adige:

```
SELECT l.name, latitude, longitude, CONCAT_WS(', ', l.kategorie,
l.bemerkung) FROM locationsv l JOIN geopolygons p ON (ST_WITHIN
(l.geodaten,p.geodaten) = 1) WHERE p.name LIKE 'trentino%' AND
kategorie LIKE 'langobardisches Graeberfeld';
```

Die Abfrage liefert eine Liste von insgesamt 45 Belegen, deren Anfang hier abgebildet wird:

name	latitude	longitude	concat_ws(', ', l.kategorie, l.bemerkung)
Tisens	46.57496985	11.16149742	langobardisches Graeberfeld, Volker Bierbrauer, L´insediamento del periodo tardoantico e altomedievale in Trentino-Alto Adige (V-VII secolo). Fondamentali caratteristiche archeologiche e notazione per una carta sulla diffusione degli insediamenti. In: G. C. Menis (Hrsg.) L´Italia longobarda , Venedig 1991, Seite 152/153
Siebeneich	46.51154445	11.27502431	langobardisches Graeberfeld, Volker Bierbrauer, L´insediamento del periodo tardoantico e altomedievale in Trentino-Alto Adige (V-VII secolo). Fondamentali caratteristiche archeologiche e notazione per una carta sulla diffusione degli insediamenti. In: G. C. Menis (Hrsg.) L´Italia longobarda , Venedig 1991, Seite 152/153
Perdonig	46.4939193	11.23164334	langobardisches Graeberfeld, Volker Bierbrauer, L´insediamento del periodo tardoantico e altomedievale in Trentino-Alto Adige (V-VII secolo). Fondamentali caratteristiche archeologiche e notazione per una carta sulla diffusione degli insediamenti. In: G. C. Menis (Hrsg.) L´Italia longobarda , Venedig 1991, Seite 152/153
Eppan-St. Pauls	46.47003996	11.25321154	langobardisches Graeberfeld, Volker Bierbrauer, L´insediamento del periodo tardoantico e altomedievale in Trentino-Alto Adige (V-VII secolo). Fondamentali caratteristiche archeologiche e notazione per

Abbildung 15: Funde langobardischer Gräberfelder auf dem Gebiet der Region Trentino-Alto-Adige; Datenbankabfrageergebnis (Ausschnitt)

Dieses und andere Ergebnisse in Listenform lassen sich sodann ohne großen Aufwand wieder in elektronische Karten verwandeln. Abbildung 16 zeigt eine Kartendarstellung der soeben vorgestellten Ergebnismenge im Programm Google Earth .

[11] Die derzeit zuverlässigsten Geofunktionen scheint das Datenbankmanagementsystem Postgres zu besitzen. Grundsätzlich ist eine Datenmigration von MySQL nach Postgres durchaus möglich, jedoch müssen Aufwand und Gewinn sorgfältig gegeneinander abgewogen werden. Wie schon erwähnt, ist damit zu rechnen, dass künftige MySQL-Versionen entscheidende Verbesserungen auf dem Gebiet der Verarbeitung von Geodaten mitbringen werden.

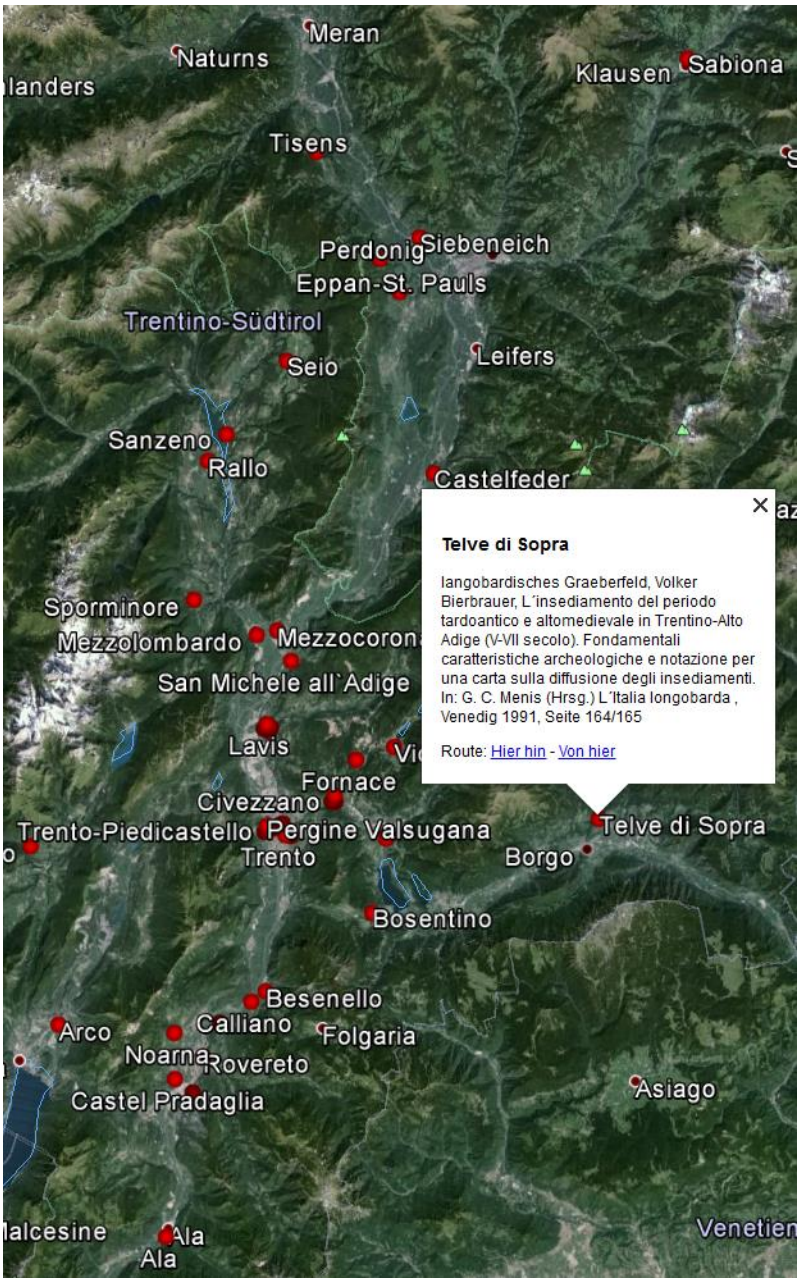

Abbildung 16: Visualisierung des Abfrageergebnisses (Abbildung 15) auf einer Google-Earth-Karte

Das Konzept der strukturierten, georeferenzierten Daten setzt der Phantasie hier buchstäblich keine Grenzen. Voraussetzung ist lediglich die Erfassung möglichst vieler bzw. nach Bedarf immer weiterer georeferenzierter Daten. Im Ergebnis entsteht ein im Grunde grenzenloses Datennetz, oder vielleicht besser: ein mehrdimensionales, grenzenloses Datengitter, zwischen dessen Knoten sich immer neue Beziehungen herstellen lassen. Die ursprüngliche Sichtweise, derzufolge primäre Sprachdaten um Metadaten anderer Art erweitert werden, löst sich damit allerdings auf – je nach gewähltem Standpunkt werden Daten zu Metadaten und umgekehrt. Entscheidend ist jedoch, dass eine induktive Konstruktion historischer Sprach- und Kulturräume ermöglicht wird, die eine Betrachtung von unterschiedlichen Standpunkten und in unterschiedlichen Perspektiven

gestattet. Auf diese Weise wird das gesammelte Datenmaterial für Forscher unterschiedlicher Disziplinen relevant. Das folgende, zweidimensionale, Schema illustriert das skizzierte System und unterstreicht die Bedeutung der Georeferenzierung als der 'Brücke', die, ihrer Qualität nach scheinbar unvereinbare, Daten zu verbinden vermag (vgl. Abbildung 17).

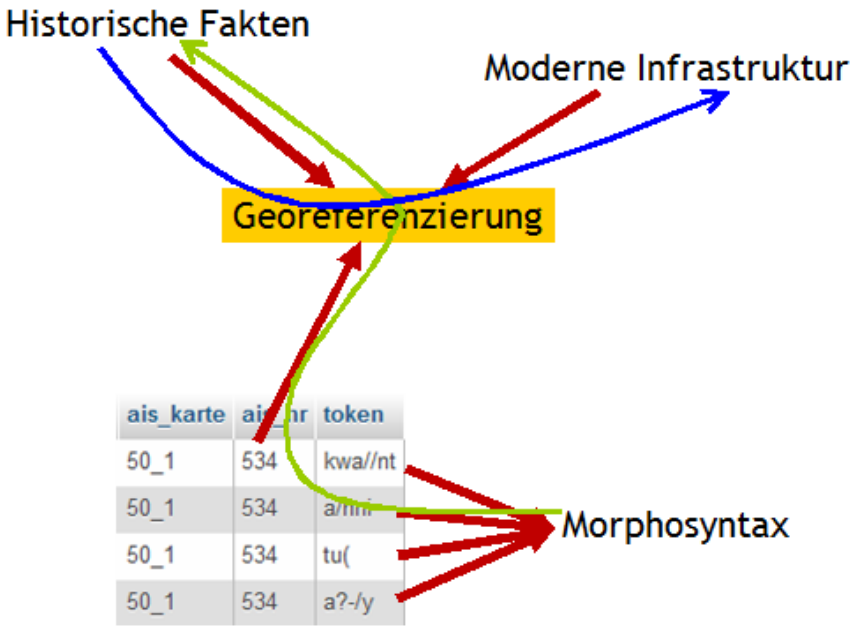

Abbildung 17: Entstehung eines Datennetzes/-gitters und die zentrale Bedeutung der Georeferenzierung

Wie eingangs erwähnt, unterscheiden sich unsere Projekte hinsichtlich der Voraussetzungen und des Umgangs mit der Frage der Georeferenzierung. Während die Projekte ASD und AdIS zumindest bislang auf zumeist eindeutig georeferenzierbaren Daten basieren, verhält sich dies bei den Projekten AsiCa, Metropolitalia und künftig auch Verba Alpina ein wenig anders. Im Fall von AsiCa besteht zwar insofern ein eindeutiger Geo-Bezug, als in diesem Projekt der Dialekt ausgewählter Ortschaften in Kalabrien untersucht wird. Ein ganz wesentlicher Aspekt bei der Datenerhebung sind jedoch auch ortsunabhängige Parameter, wie z.B. Alter, Geschlecht und Migrationserfahrung der Informanten gewesen. Dieser Umstand führte zur Wahl einer kartographischen Darstellung, die gleichsam die verschiedenen Dimensionen der Datenerhebung miteinander verbindet.

Die abgebildete Karte (Abbildung 18) visualisiert das Ergebnis der Analyse der unterschiedlichen Realisierungen des Stimulus '*comincia a piovere*' durch die Informanten, von denen jeder durch eines der kleinen Quadrate auf der Karte repräsentiert wird. Konkret wird überprüft, ob der Informant eine Infinitiv-Konstruktion verwendet hat (so, wie im hochitalienischen Stimulus) oder nicht. Je nach Analyseergebnis wird dem entsprechenden Quadrat eine bestimmte Farbe bzw. Markierung zugewiesen. Die Kodierung der immanenten nicht-geographischen Logik erfolgt durch die Anordnung der Quadrate vor

dem Hintergrund der Kalabrienkarte. Die Informantenquadrate sind in Vierergruppen gebündelt in die Nähe der jeweiligen Herkunftsorte gerückt, wobei die Vierergruppen innerhalb der Küstenlinie die ortsfesten Nicht-Migranten symbolisieren. Die Symbole der Informanten mit Migrationserfahrung sind jeweils jenseits der Küstenlinie gleichsam im Meer angeordnet. Innerhalb jeder Vierergruppe repräsentieren die beiden linken Quadrate männliche Informanten, die rechten weibliche, die oberen die Vertreter der Eltern- und die unteren die Vertreter der Kind-Generation.

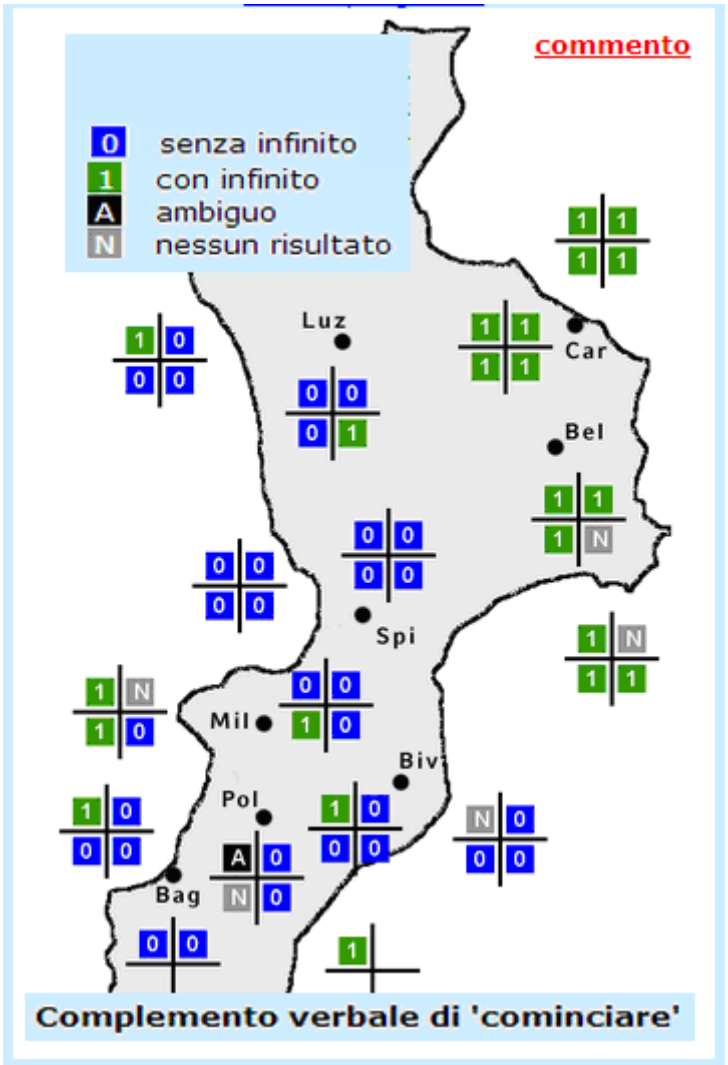

Abbildung 18: Abbildung teil-georeferenzierbarer Daten (Beispiel AsiCa)

Die beiden Projekte Metropolitalia und Verba Alpina wiederum basieren zumindest teilweise auf Daten, die zwar exakt georeferenzierbar sind, deren Authentizität jedoch zunächst nicht hundertprozentig als gesichert betrachtet werden kann. Dieser Umstand resultiert aus der speziellen Erhebungsmethode, die bei diesen beiden Projekten zum Einsatz kommt: das sog. Crowdsourcing. Dabei werden im Internet von im Grunde anonymen Informanten Sprachdaten und Einschätzungen zu deren geographischer Herkunft gesammelt. Dem Manko der mangelnden Überprüfbarkeit der Zuordnung von Sprachdaten zu Geokoordinaten durch den einzelnen Internetuser wird mit der Sammlung einer

Vielzahl von Georeferenzierungen ein und derselben Äußerung durch eine möglichst große Anzahl von Internetusern begegnet. Auf diese Weise lassen sich georeferenzierte Karten erzeugen, die die aktuelle Verbreitung bestimmter sprachlicher Phänomene abbilden und in denen eventuell fragwürdige Sondereinschätzungen nicht zum Tragen kommen oder auch algorithmisch unterdrückt werden können.

Da sowohl Metropolitalia als auch Verba Alpina sich noch in der Entwicklungs- bzw. Startphase befinden, konnten noch nicht genügend Daten gesammelt werden, um den beschriebenen Effekt überzeugend illustrieren zu können. Die folgende Karte stammt aus dem Projekt Metropolitalia und zeigt die Einschätzungen der 'crowd' bezüglich der Herkunft des Ausdrucks '*Telefonaci a tuo padre!*'. Die Datenbasis besteht zwar erst aus insgesamt vier Einschätzungen, zeigt aber bereits eine deutliche Konzentration auf Süditalien und Sizilien (vgl. Abbildung 19).

Abbildung 19: Metropolitalia, Verortung des Ausdrucks 'Telefonaci a tuo padre!' durch Internet-User

Der unterschiedliche Sättigungsgrad der Farbflächen korreliert mit der Anzahl der entsprechenden Verortungen. Das hier angewandte Konzept besitzt überdies die Eigenschaft, sich verändernde Einschätzungen der 'crowd' und somit die Dynamik der Sprachwandels abbilden zu können.

4 Bibliographie

AIS Online-Archiv: http://www.italiano.unibe.ch/content/linguistica/archivio_ais/index_ger.html [10.10.2013].

Jaberg, Karl / Jud, Jakob (1928): *Der Sprachatlas als Forschungsinstrument. Kritische Grundlegung und Einführung in den Sprach- und Sachatlas Italiens und der Südschweiz*, Halle [Saale].

Jaberg, Karl / Jud, Jakob (1928-1940): *Sprach- und Sachatlas Italiens und der Südschweiz* (AIS), Zofingen.

Jaberg, Karl / Jud, Jakob (1960): *Index zum Sprach- und Sachatlas Italiens und der Südschweiz. Ein propädeutisches etymologisches Wörterbuch der italienischen Mundarten*, Bern.

Krefeld, Thomas (2007): "Dal punto diatopico alla diatopia del punto: una prospettiva promettente", in: Raimondi, Giammario / Revelli, Luisa (Hrsg.), *La dialectologie aujourd'hui. Atti del Convegno 'Dove va la dialettologia?'*, Alessandria, 37-50.

Krefeld, Thomas (2007a): "Kalabresisch *fra pogu vegnu a ti trovu* - Fossil oder Produkt syntaktischen Wandels?", in: Stark, Elisabeth / Schmidt-Riese, Roland / Stoll, Eva (Hrsg.), *Romanische Syntax im Wandel*, Tübingen, 437-448.

Krefeld, Thomas / Lücke, Stephan (2008): “ASICA-online: Profilo di un nuovo atlante sintattico della Calabria, *Rivista di Studi Italiani* 26, 196-211.

Krefeld, Thomas / Pustka, Elissa (Hrsg.) (2010): *Perzeptive Varietätenlinguistik*, Frankfurt am Main.

Postlep, Sebastian (2019): *Zwischen Huesca und Lérida: Perzeptive Profilierung eines diatopischen Kontinuums*, Frankfurt am Main.

Purschke, Christoph (2011): *Regionalsprache und Hörerurteil. Grundzüge einer perzeptiven Variationslinguistik*, Stuttgart.

Salminger, Irmengard (2009): *Subordination und Finitheit im Kalabrischen: Eine Untersuchung im Rahmen des Atlante Sintattico della Calabria (ASiCa)*, Frankfurt am Main

Scheuermeier, Paul (1943/1956): *Bauernwerk in Italien, der italienischen und rätoromanischen Schweiz: Eine sprach- und sachkundliche Darstellung landwirtschaftlicher Arbeiten und Geräte*, 2 Bde., Zürich (ital. Übersetzung: Mailand 1980).

VALTS = Gabriel, Eugen (1985-2004): *Vorarlberger Sprachatlas mit Einschluss des Fürstentums Liechtenstein, Westtirols und des Allgäus*, vol. 1-5, Bregenz.