HUMBOLDT-UNIVERSITÄT ZU BERLIN

Fabio Tosques (Hrsg.)

20 Jahre digitale Sprachgeographie

Berlin 2014

Fabio Tosques (Hrsg.)

20 Jahre digitale Sprachgeographie

Tagungsband

Berlin 02. bis 03. November 2012

Humboldt-Universität zu Berlin

Institut für Romanistik

2014

ISBN: 978-3-00-046278-8

Inhaltsverzeichnis

Zur Tagung *20 Jahre digitale Sprachgeographie*

2. bis 3. November, Humboldt-Universität zu Berlin

Fabio Tosques, Humboldt-Universität zu Berlin

Vorwort

Als vor gut 20 Jahren mit der allgemeinen Verbreitung des Computers die elektronische Datenverarbeitung, kurz EDV, auch Einzug in die sprachgeographische Forschung hielt, hatten wir nur vage Vorstellung davon, was heute tatsächlich realisiert werden kann.

Schon früh versuchte die Forschung sich die Vorteile von Datenbanken zu Nutze zu machen, wenn es darum ging, das im Feld gesammelte Material schnell und umfassend zu verarbeiten. Die Technik sollte zu Anfang somit vor allem die internen Prozesse bei der Verarbeitung und Aufbereitung der Daten unterstützen. Mit dem Aufkommen der CD-ROM wurde der Forschung ein Instrument an die Hand gegeben, welches erstmalig ausreichende Kapazität hatte, um große Mengen von Daten zu speichern und auf der das schnelle Auffinden von Datensätzen möglich war. Damit konnte die Technik nicht nur die internen Prozesse unterstützen, sondern den Sprachatlas als solchen auch um zusätzliche Features erweitern. Zum ersten Mal konnte beispielsweise dem Nutzer das komplette Tonmaterial – sofern es vorher im Feld aufgezeichnet worden war – zur Verfügung gestellt werden. Zwar gab es hier und da vereinzelt Versuche, den Nutzern auditives Material zur Verfügung zu stellen. Dies geschah bis zur Entwicklung der CD-ROM jedoch auf Langspielplatten oder Kassetten, deren Kapazität sehr eingeschränkt war. Ein zweites Problem von analogen Datenträgern bestand im Qualitätsverlust mit jeder Nutzung. Drittens waren die Produktionskosten oft so hoch, dass nach einer Kosten-Nutzen-Abwägung in der Regel auf die Beigabe von Tonmaterial verzichtet wurde. Mit der Einführung der CD-ROM schienen diese Probleme weitgehend gelöst: Kapazität, Qualität und Kosten.

Von den neuen technischen Möglichkeiten begeistert, planten Anfang der 1990er Jahre die Sprachwissenschaftler und Dialektologen Roland Bauer (Salzburg) und Dieter Kattenbusch (damals Regensburg, heute Berlin), einen Atlas der italienischen Dialekte und Minderheitensprachen zu entwickeln, bei dem alle Antworten der Informanten nicht nur lesbar, in Form von Transkriptionen, zur Verfügung gestellt werden sollten, sondern diese auch vollständig angehört werden konnten. So begannen nach einer Planungs- und Entwicklungsphase sowie der Erstellung eines speziellen Fragebuchs im Jahr 1993 die ersten Aufnahmen auf Sizilien. Nur kurze Zeit später wurden die technisch aufbereiteten Daten der Öffentlichkeit präsentiert, der erste digitale Sprachatlas war geboren (vgl. Bauer 1995, Kattenbusch, 1995).

Während in Salzburg und in Berlin an der Verbreitung sprachgeographischer Daten auf CD-ROM gefeilt und getüftelt wurde, entwickelte ein paar hundert Kilometer weiter westlich ein Mitarbeiter des *Conseil Européen pour la Recherche Nucléaire* (CERN) in Genf ein System, welches Daten miteinander verknüpfen kann, das *Hypertext Transfer Protocol* (HTTP) und das *World Wide Web* (WWW) und revolutionierte die Mediengeschichte. Mit diesem System eröffneten sich für die Sprachgeographie ungeahnte Möglichkeiten. Gelänge es schließlich, das Tonmaterial mit den Transkriptionen und den ge-

ographischen Daten zu verbinden, so stünde der Forschung ein Tool zur Verfügung, welches die Sprachatlanten zum „Sprechen“ bringen würde.

Derart interessiert an den neuen Entwicklungen und dem Potential, das in ihnen steckte, entwickelten Carola Köhler und Marcel Müller den digitalen Sprachatlas *VIVaio Acustico delle Lingue e dei Dialetti d'Italia*[1] (VIVALDI, in der deutschen Übersetzung *Akustischer Sprachatlas der Dialekte und Minderheitensprachen Italiens*), der erstmals auch online genutzt werden konnte. Die erste Version dieses im *World Wide Web* verfügbaren Sprachatlasses konnte 1999, sechs Jahre nach dem Erscheinen der ersten CD-ROM, der Öffentlichkeit präsentiert werden.

Die Idee, das 20-jährige Bestehen des digitalen Sprachatlasses mit einer Tagung zu würdigen, wurde auf einer der vielen Reisen nach Italien im „Feld“ gdeboren. Thema der Tagung sollte sein, die aktuellen Entwicklungen und Projekte von digitalen Sprachatlanten vorzustellen. An zwei Tagen konnten so zum einen die technischen Möglichkeiten, Probleme und Lösungsvorschläge vorgestellt werden mit denen heutige Atlasprojekte realisiert werden. Zum zweiten sollten digitale Sprachatlanten vorgestellt werden, die vollständig funktionsfähig sind und von allen genutzt werden können.

Im Beitrag von Giorgio Di Nunzio (Universität Padua) und Stefan Rabanus (Universität Verona) werden Methoden vorgestellt, wie Daten der zimbrischen Dialekte in Norditalien gesammelt, digitalisiert und getaggt werden können. Ziel des vorgestellten Projekts ist die Entwicklung eines allgemein verfügbaren Tools, mit dem Tiefenstrukturen einer Sprache, hier das Zimbrische, genauer untersucht werden können.

Alois Dicklberger von der Universität Passau stellt in seinem Beitrag das von der EU geförderte Projekt *ONDa* (OrtsNamenDatenbank) vor, ein WebGIS-Tool, welches die geographische Verteilung von Namenstypen oder die urkundliche Erstbelegung von Ortsnamen auf Karten aufbereitet und exakt darstellt. Jeder Ort erhält neben der kartographischen Darstellung hilfreiche und nützliche Informationen, die aus den verschiedensten historischen, teils von schwer zugänglichen, Quellen erarbeitet wurden.

Bisher eher am Rande wurden die prosodischen Phänomene von einzelnen Dialekten und Dialektgruppen untersucht. Dabei charakterisiert doch besonders die Satzmelodie die Dialekte und lässt sich anhand dieser räumlich einordnen. Die heutzutage verfügbare Technik, sowohl bei der Aufnahme wie bei der Analyse, gibt der Forschung ein hilfreiches Werkzeug, mit dem prosodische Merkmale von Einzelsprachen sehr präzise untersucht werden können. Methodisch kann die Wissenschaft dabei auf die von Michel Contini, Grenoble, im Rahmen des *Atlas Multimédia Prosodique de l'Espace Roman* (AMPER) entwickelten Fragebuches zurückgreifen. Von aktuellen Entwicklungen und von neuen Ergebnissen berichtet Antonio Romano, der lange Zeit direkt mit Contini gearbeitet hat und inzwischen das Projekt von Turin aus betreut, in seinem Beitrag.

Auf die langjährigen Erfahrungen des Projekts AMPER setzen Anca-Diana Bibiri, Oana Panaite und Adrian Turculet in ihrem Beitrag zum *Multimedia Prosodic Atlas of Romania*. Bei der Umsetzung des Projekts greifen die Autoren auf die methodischen Grundlagen Continis zurück, erweitern diese sinnvollerweise um Besonderheiten im Rumänischen. Vom dabei verwendeten Fragebuch, den erzielten Ergebnissen und neueren Analysetechniken schreiben die drei Autoren.

[1] http://www2.hu-berlin.de/vivaldi.

Von den bedeutenden Fortschritten in der Geo- und Korpuslinguistik berichten die baskischen Sprachforscher Gotzon Aurrekoetxea und Charles Videgain. Durch den Zusammenschluss verschiedener Forscher der Université du Pays Basque und weiterer Universitäten in der EUDIA-Gruppe konnten neue Ergebnisse bezüglich der Analyse und der Behandlung von Sprachdaten gewonnen werden. Vorgestellt werden im Beitrag zwei im Web verfügbare Korpora: CORPUSLEM und TEI sowie Programme für die Erzeugung von geolinguistischen Karten und für die Dialektometrisierung von Sprachdaten.

Unter der Leitung von Hans Goebl wurde 2012 das Projekt *Atlas des Dolomitenladienischen und angrenzender Dialekte* vollendet. Welche Software dabei entwickelt wurde und wie diese eingesetzt werden kann, wird im Beitrag von Hans Goebl, Bernhard Schauer, Heidemarie Beer und Agnes Staudinger vorgestellt. Von der Erzeugung der Karten bis zum Sprechenden Sprachatlas werden die notwendigen Einzelschritte ausführlich beschrieben und erklärt.

Obschon der ALD-II erst kurz vor der Tagung veröffentlicht wurde, ist es Roland Bauer, Salzburg, gelungen, erste dialektometrische Analysen mit den Daten des ALD-II durchzuführen und zu präsentieren. Mit Hilfe des von Edgar Haimerl entwickelten Softwaretools *VisualDialectometry* (VDM) werden im Beitrag anhand eines kleineren Korpus Arbeitskarten vorgestellt, die für die Erzeugung von Ähnlichkeitsprofilen herangezogen wurden.

Seit gut zehn Jahren erheben Stefan Elspaß und Robert Möller Daten zur deutschen Alltagssprache in den deutschsprachigen Ländern Deutschland, Österreich und der Schweiz. Mit Hilfe eines Online-Fragebogens werden die Informanten indirekt nach lexikalischen, morphologischen, syntaktischen und phonetischen Besonderheiten in ihrer Region befragt. Aus den Antworten werden Karten erzeugt, die die verschiedene Verwendung von Geosynonymen und weiteren Besonderheiten in der Alltagssprache sehr anschaulich darstellen. Im vorliegenden Beitrag berichten die Autoren zudem von einer Fragerunde, in der die Informanten ihre sprachliche Situation selbst einordnen sollten, und vergleichen diese dann mit den erhobenen Daten.

Gleich vier geolinguistische Projekte werden von Thomas Krefeld und seinem langjährigem Mitarbeiter Stephan Lücke vorgestellt. Abgeschlossen und schon seit längerer Zeit online verfügbar ist der *Atlante Sintattico della Calabria* (ASiCa), der die syntaktischen Strukturen kalabresischer Dialekte von Auswanderern und Ortsansässigen verschiedenen Alters und nach Geschlecht untersucht. Das zweite online verfügbare Projekt *MetropolItalia* soll in spielerischer Weise Interessierte Vertrautes und Neues lehren. Beim *Atlante Linguistico digitale dell'Italia e della Svizzera meridionale* handelt es sich um eine digitalisierte Version des *Sprach- und Sachatlasses Italiens und der Südschweiz* (AIS), die jedoch um zusätzliche Informationen angereichert und erweitert wurde. Das Projekt befindet sich, wie auch das der *Audio-Atlas siebenbürgisch-sächsischer Dialekte* (ASD) noch in der Entwicklung. Zum Stand der Projekte und zu den Zielen der Projekte berichten die beiden Autoren in ihrem Beitrag.

Schon seit längerer Zeit online ist der Atlas des Frankoprovenzalischen von Andres Kristol und Federica Diemoz. Dabei handelt es sich um einen der ersten Atlanten, die nicht nur Tonmaterial, sondern auch Videos der Informanten online stellen. Die enorme Datenmenge erlaubt es den Autoren bisher nicht, sämtliches Material der Öffentlichkeit online zur Verfügung zu stellen. Welches Potential in den erhobenen Daten steckt und

welche Analysen damit durchführbar sind, stellen die beiden Autoren in ihrem Beitrag vor.

Wir möchten den Teilnehmern nochmals für die gelungene Tagung und die interessanten Beiträge danken und hoffen, dass sich – auch in unregelmäßigen Abständen – Nachahmer finden, die das Thema der Tagung weiter aufnehmen und Ideen und Lösungen vorstellen.

Bitte beachten Sie noch den folgenden wichtigen Hinweis: die meisten der hier gedruckten Karten entfalten ihre Wirkung erst in Farbe. Da aus Kostengründen nur in schwarz-weiß gedruckt werden konnte, möchten wir die interessierten Leser auf die online-Publikation[2] hinweisen. Dort finden Sie alle Karten in Farbe und in hoher Auflösung für den persönlichen Ausdruck.

Bibliographie

Bauer, Roland (1995): “Documentazione sonora dei dialetti siciliani”, in: Giovanni Ruffino (a cura di), *Percorsi di geografia linguistica. Idee per un atlante siciliano della cultura dialettale e dell'italiano regionale*, Palermo.

Kattenbusch, Dieter (1995): “Atlas parlant de l'Italie par régions: VIVALDI”, in: *Estudis de lingüística i filologia oferts a Antoni M. Badia i Margarit*, Barcelona.

[2] http://www2.hu-berlin.de/vivaldi/tagung.

Research on Geolinguistic Linked Data: The Test Case of Cimbrian Varieties

Giorgio Maria Di Nunzio, Department of Information Engineering, University of Padua
& Stefan Rabanus, Chair of German Linguistics, Yerevan State Linguistic University

In this paper, we present a geolinguistic linked open data approach of a multidisciplinary and collaborative project, "Cimbrian as a test case for synchronic and diachronic language variation", which provides linguists with a test bed for formal hypotheses concerning human language. Aims of the project are to collect, digitize and tag linguistic data from the German dialect varieties of Cimbrian – spoken in three areas of northern Italy: Giazza (province of Verona), Luserna (province of Trento), and Roana (province of Vicenza) – and to make available on-line a valuable and innovative linguistic resource for the in-depth study of Cimbrian.

1 Introduction

Language resources that have been publicly made available can vary in the richness of the information they contain: on one hand, a corpus typically contains at least a sequence of words, sound or tags; on the other end, a corpus may contain a large amount of information about the syntactic structure, morphology, prosody, and semantic content of every sentence, plus annotation of discourse relations or dialogue acts (cf. Bird/Klein/Loper 2009). When researchers need to perform particular linguistic analyses such as capturing fine-grained grammatical differences by comparing various dialectal translations of the same sentence, the only way to build a high accuracy language resource is by manual annotation (cf. Agosti et al. 2011, 63-64).

The heterogeneity of linguistic projects has been recognized as a key problem limiting the reusability of linguistic tools and data collections (cf. Chiarcos 2012). The rate of re-use for linguistic database technology together with related processing tools and environments is still too low. For example, the Edisyn search engine – the aim of which was to make different dialectal databases comparable – "in practice has proven to be unfeasible"[1] to date. In order to find common ground where linguistic material can be shared and re-used, the methodological and technological boundaries between different research projects have to be overcome.

The research direction we pursue in this work is to move the focus from the systems handling the linguistic data to the data themselves. We address these issues by adopting an approach based on the Linked Open Data (LOD) paradigm with the aim of enabling interoperability at a data-level by overcoming the characteristics of each collection which depend on different methodological and technological choices. For this purpose, we present a linguistic project which aims (i) to collect, digitize and tag linguistic data from the Cimbrian varieties and, (ii) to distribute data by means of an LOD. We also present a Web application which produces dynamic maps on user request that is built upon this open dataset.

[1] http://www.dialectsyntax.org/wiki/About_Edisyn. [All URLs in this paper were last accessed on January 17, 2013.]

2 Linguistic Project

In this contribution, we present the results of an ongoing multidisciplinary collaboration which is conducted in the context of the project named *Atlante Sintattico d'Italia*, Syntactic Atlas of Italy (ASIt)[2]. This project aims to implement a digital library system that provides access and enables management of curated dialect data, also by means of an advanced user interface specifically designed to update and annotate the linguistic data (cf. Agosti et al. 2012).

In this context, the Cimbrian project[3] focuses on the so-called Triveneto area in the north-eastern part of Italy, in which the Cimbrian dialects are in intense language contact with the Italian dialects belonging to the Lombard and Venetian dialect groups (cf. Pellegrini 1977). Cimbrian, spoken in the language island of Giazza (Veneto, province of Verona), Luserna (Trentino/South Tyrol, province of Trento) and – historically – Asiago/Roana (Veneto, province of Vicenza)[4], is of great interest to three important lines of research in linguistics:

- Romance dialectology: linguistic contact phenomena are visible especially at the lexical level,
- German dialectology: the language island varieties exhibit a high level of preservation of certain structural characteristics, and
- Historical linguistics: the diachronic development of a variety in isolation shows a particularly interesting mixture of preservation and innovation.

This historic language-contact situation (supplemented by the entry of spoken Regional Northern Italian in the repertoire of the speakers in the course of the 19th century) is crucial for our idea that language variation in Cimbrian depends both on its structural possibilities as a German dialect and on the multilinguism of its speakers. Hence, it is necessary to consider the Cimbrian and the Italian dialects of the area with respect to the same grammatical categories and features.

The interest for this linguistic context is witnessed by many studies on Cimbrian throughout the last decade (cf. the overviews in Bidese 2010). Furthermore, the present project, which puts its focus prominently on Cimbrian syntax, is coherent to similar projects at European level in that it creates a database of syntactic structures – which so far have been neglected in traditional dialectological work (cf. Rabanus/Alber/Tomaselli 2008). Finally, Cimbrian is an endangered language, with only few speakers of advanced age speaking Cimbrian fluently in Giazza[5]. This makes collection of linguistic data of this language all the more important.

[2] http://asit.maldura.unipd.it/.
[3] http://ims.dei.unipd.it/websites/cimbrian/.
[4] Additionally, some data from Mòcheno – another German-language island variety in Trentino which is collocated geographically and linguistically in between Cimbrian and Bavarian in South Tyrol (cf. Rabanus 2013) – have been considered. The entire area of Cimbrian and Mòcheno has been surveyed and documented in detail by Bruno Schweizer in the 1940's whose maps have been published as linguistic atlas (Schweizer 2012) only in the context of our project.
[5] The situation is much better in Luserna even though there are no children acquiring Cimbrian as mother language.

2.1 Documents

In contrast to many other German dialects Cimbrian has a tradition as written languages and a literature that goes back to the beginning of the 17th century. This makes it possible to reconstruct the language change for at least four empirically attested stages (1602, 1844, 1942, 2009/2010). The written documents that have been elaborated in order to form part of the database are "Christlike unt korze Dottrina" (1602, cf. Meid 1985), "Novena vun unzar liben Vraun" (1844, cf. Stefan 2000), "Taut6. Puox tze Lirnan Reidan un Scraiban iz Gareida on Lietzan" (1942, cf. Cappelletti/Schweizer 1942). These Cimbrian texts have been completely transcribed (faithfully to their graphic form) and segmented in sentences which have also been linked to their translations in Italian and Standard German. For contemporary Cimbrian fieldwork has been conducted in Giazza (2009 and 2010). In order to be able to compare the Cimbrian data with data from the Italian dialects and other projects on the syntax of German varieties, the questionnaire was designed as similar as possible to the ASIt questionnaires and has integrated questions elaborated by the SyHD project (*Syntax hessischer Dialekte*, Universities of Marburg/Frankfurt/Vienna)[6]. The interviews have been digitally recorded and transcribed both according to a Cimbrian orthography (developed for this purpose) and phonetically. The questionnaire so far aims to elicit syntactic and morphological data.

2.2 Tags

After segmentation of the sentences, tagging of the linguistic data is carried out. We start with tagging at the word-level, determining the parts of speech of single words. Tagging of syntactic phenomena at the sentence level and tagging of syntactic constituents will take place in a second phase of the project. The starting point for developing an adequate set of tags for Cimbrian is the tagset elaborated by the Edisyn project[7], especially for the (Dynamic) Syntactic Atlas of the Dutch dialects (DynaSAND)[8]. In collaboration with the ASIt team, we have developed a language-specific set of tags which is suitable for Cimbrian but, at the same time, allows the Cimbrian data to be linked to other databases of dialect syntax. This involves assigning the same names to same parts of speech as in the Edisyn and the ASIt databases, at most adding tags when they are needed for language-specific structures of Cimbrian, or leaving out tags which are not relevant for Cimbrian. Thus, for instance, the tag "verbal particle" has been added to identify verbal particles which can be found in German dialects (e.g. the verbal particle in the Standard German sentence, "Ich gehe weg", 'I go away'), but gender values such as "masculine" have been left out for the tag of the past participle, since past participles never inflect for gender in German varieties. We can therefore imagine the creation of a language-specific tagset as starting from a universal core, shared by all languages, and subsequently developing a language-specific periphery, which is compatible with other databases and appropriate to classify language-specific structures.

The sentences that are tagged can be searched by means of a search interface as shown in Figure 1 (see Section 3.2 for more details about this interface).

[6] http://www.syhd.info/.
[7] http://www.dialectsyntax.org/.
[8] http://www.meertens.knaw.nl/sand/.

3 Digital Geolinguistic Linked Open Data

The LOD paradigm refers to a set of best practices for publishing data on the Web[9] and it is based on a standardized data model, the Resource Description Framework (RDF).[10] RDF is designed to represent information in a minimally constraining way and it is based on the following building blocks: graph data model, URI-based vocabulary, data types, literals, and several serialization syntaxes.

3.1 Geolinguistic Ontology

The common ground defined by current European linguistic projects allows us to infer the fundamental classes and properties necessary to define an ontology for modeling and representing geolinguistic resources. Geolinguistic concepts can be organized into three major areas: geography, derivation, and tagging. The geographical area comprehends classes and properties related to physical places. The derivation area is about people speaking a certain language, their relationships, and the geographical area where they live. Furthermore, the derivation area allows for the study of the correlation between social factors, education and knowledge of the dialect, and the distinctiveness of a local dialect. Lastly, the tagging area regards language-specific classes and properties, such as documents, sentences, words, and their relationships (cf. Di Buccio/Di Nunzio/Silvello 2012, 2013a).

We present an example of how we built the ontology of a document. A document represents the composite unit of study of a dialect; it is composed by one or more sentences which are subsequently divided in words for further analysis. A document may be redacted in one language (e.g. Italian or English) and then translated into several dialects which allow for linguistic comparisons. The syntactical analyses of these parallel translations are possible thanks to the `Tag` class specialised into two main sub-classes: `Sentence Tag` and `PoS Tag`. `Sentence Tag` allows us to capture a sentence-level phenomena, whereas `PoS Tag` allows us to capture a phenomena occurring on a `Word` in a `Collocation`, i.e. a specific position, within a given `Sentence`. The `WordSentenceCollocation` class relates a tag to a word within a sentence along with the properties relating it to the `Word`, `Sentence` and `Collocation` classes. The `SentenceDocumentCollocation` class relates a sentence to a document specifying the collocation of the sentence within the document by means of the class `SentenceCollocation`.

This ontology is the starting point for modeling and describing geolinguistic resources because:

- it provides general-purpose concepts and relationships;
- it is extendable by adding more fine-grained classes;
- it permits an easy mapping from existing linguistic projects and publicly available databases.

[9] http://www.w3.org/DesignIssues/LinkedData.html.
[10] http://www.w3.org/RDF/.

This geolinguistic ontology allows us to expose the linguistic data as a Linked Open Dataset (see the details in Di Buccio/Di Nunzio/Silvello 2012a). Currently, the ASIt dataset is linked to DBpedia.[11]

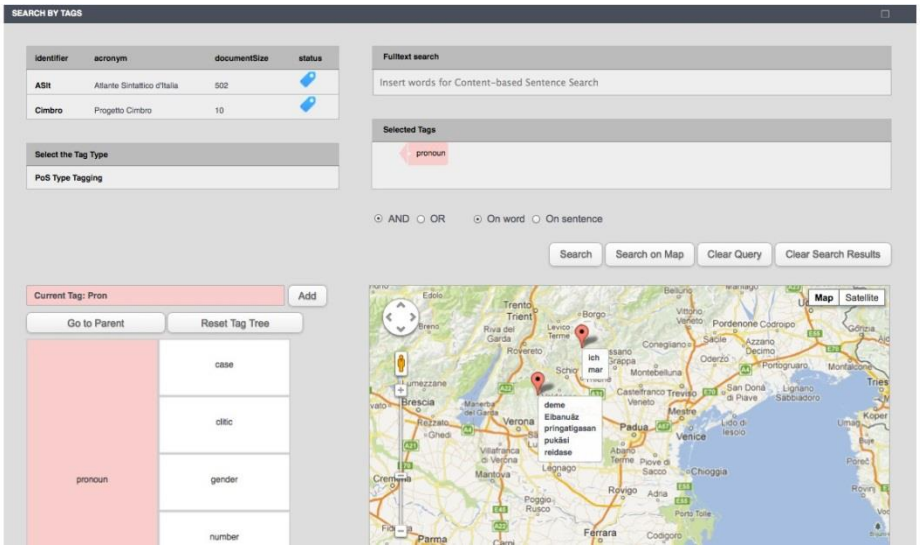


Figure 1: A screenshot of the ASIt GeoSearch interface.

3.2 Geolinguistic Web Application

The objective of the project is to provide linguists with a system for investigating variations among closely related languages. We developed a graphical user interface on top of the ASIt system that dynamically produces maps on the basis of the user request. The interface is available at the URL: http://svrims2.dei.unipd.it:8080/asit-enterprise/do/search.

A screenshot of a map produced by a tag-based search is reported in Figure 1. This type of search aims at satisfying the information need of a user searching for the geographic distribution of linguistic resources. The submitted query retrieves all the sentences that have the tags selected in the query ("pronoun" in this case). Then, for all these sentences, the system retrieves the locations and displays the words related to these tags (see details in Di Buccio/Di Nunzio/Silvello 2013).

3.3 Linguistic Analyses

The tagged corpus of Cimbrian data will be available to end users who might be linguists interested in carrying out syntactic analyses, or also informants, interested in correcting or augmenting the data. Concerning the former, it is important that the data are presented in a way which makes it usable by linguists working in different theoretical frameworks. Although it is inevitable (and, to some extent, also desirable) that the tagging of the data is influenced by theoretical considerations (in our case, the framework of genera-

[11] http://www.dbpedia.org/.

tive linguistics), it is important that the database should be of use not only to a small group of specialists.

With respect to the types of structures which can be analyzed in the tagged Cimbrian database, it will be possible to analyze syntactic structures and phenomena in great detail. It should also be possible to deduct morphological paradigms without too much effort, while it still remains a desideratum of further research projects to integrate a component which will make it possible to carry out phonological analyses on the database.

It is important that the structures in the database can be compared with structures present in other databases, since cross-linguistic comparison will be one of the major interests of an analysis of Cimbrian, which is in contact with Romance varieties (hence can be compared to the ASIt data) but has a Germanic base (hence can be compared, e.g., to the DynaSAND data). To make just one example of what an analysis in these terms could look like, consider the case of pronouns and clitics in Cimbrian. In Cimbrian documents, sentences as the following can be found (Bidese 2008, p. 134):

miar	importar-z-mar	nicht	zo	sterben
me	matter-it-me	not	to	die

'I don't mind dying'

Whereas the use of the infinitive particle *zo* and the expletive pronoun *–z* are typical of German varieties, the doubling of the object pronoun *miar, mar* could be evidence for the development of a Romance-like system of clitics in Cimbrian, differently from Standard German where clitics are not attested. The tagged database will make it possible to retrieve all sentences of the corpus containing potential clitics and will therefore create an empirical basis on which to test hypotheses as those of the development of a system of clitics in Cimbrian.

4 Conclusions

In this paper, we presented the results of an ongoing linguistic project which aims to collect, digitize and tag linguistic data from the German dialect varieties of Cimbrian. The project gave the opportunity to merge different fields of research and begin a multidisciplinary collaboration between linguists and computer scientists. Since cross-linguistic comparison will be one of the major interests of an analysis of Cimbrian, the main aim was to design and implement a digital library system that enables the management of linguistic resources of curated dialect data and provides access to grammatical data by means of a LOD approach. We imagine the use of the Geolinguistic Linked Open Dataset by third-party linguistic projects in order to enrich the data and build-up new services over them. To this purpose, we developed a graphical user interface on top of these linked data that dynamically produces maps on the basis of the user requests.

5 Acknowledgements

This work has been supported by the Project FIRB "Un'inchiesta grammaticale sui dialetti italiani: ricerca sul campo, gestione dei dati, analisi linguistica" (Bando FIRB Futuro

in ricerca 2008, cod. RBFR08KRA 003). We would like to thank Maristella Agosti, Emanuele Di Buccio, and Gianmaria Silvello of the Department of Information Engineering of the University of Padua, Paola Benincà and Diego Pescarini from the Department of Linguistic and Literary Studies of the University of Padua, Alessandra Tomaselli and Birgit Alber from the Department of Foreign Languages and Literatures of the University of Verona.

References

Agosti, M. et al. (2011): "A Digital Library of Grammatical Resources for European Dialects", in: Agosti, M. et al. (eds.): *Digital Libraries and Archives. 7th Italian Research Conference, IRCDL 2011. Pisa, Italy, January 20-21, 2011. Revised Selected Papers,* Berlin, Heidelberg, 61-74.

Agosti, M. et al. (2012): "A curated database for linguistic research: The test case of cimbrian varieties", in: Choukri, K. et al. (eds.): *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12), Istanbul, Turkey, May 23-25. European Language Resources Association (ELRA)*, 2230-2236.

Bidese, E. (2008): *Die diachronische Syntax des Zimbrischen*, Tübingen.

Bidese, E. (ed.) (2010): *Il cimbro negli studi di linguistica*, Padua.

Bird, S./Klein, E./Loper, E. (2009): *Natural Language Processing with Python*, Sebastopol.

Di Buccio, E./Di Nunzio, G./Silvello, G. (2012): "A system for exposing linguistic linked open data", in: *Research and Advanced Technology for Digital Libraries International Conference on Theory and Practice of Digital Libraries (TPDL 2012), Papho, Cyprus, September 23-27,* Berlin, Heidelberg, 172–178.

Di Buccio, E./Di Nunzio/G., Silvello, G. (2013a): "A curated and evolving linguistic linked dataset", in: *Semantic Web*, 4, 3, 265-270.

Di Buccio, E./Di Nunzio, G./Silvello, G. (2013b): "A geolinguistic web application based on linked open data", in: *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'13)*, 1101-1102.

Cappelletti, G./Schweizer, B. (1942): *Taut6. Puox tze Lirnan Reidan un Scraiban iz Gareida on Lietzan*, Bolzano.

Chiarcos, C. (2012): "Interoperability of corpora and annotations", in: Chiarcos, C./Nordhoff, S./Hellmann, S. (eds.): *Linked Data in Linguistics*, Berlin, Heidelberg, 161–179.

Meid, W. (1985): *Der erste zimbrische Katechismus.* Christlike unt korze Dottrina. *Die zimbrische Version aus dem Jahre 1602 der* Dottrina Christiana Breve *des Kardinal Bellarmin in kritischer Ausgabe. Einleitung, italienischer und zimbrischer Text, Übersetzung, Kommentar, Reproduktionen*, Innsbruck.

Pellegrini, G. (1977): *Carta dei dialetti d'Italia*, Pisa.

Rabanus, S./Alber, B./Tomaselli, A. (2008): „Erster Veroneser Workshop ‚Neue Tendenzen in der deutschen Dialektologie: Morphologie und Syntax'", in: *Vorschläge für die Ausrichtung zukünftiger Dialektsyntaxprojekte. Zeitschrift für Dialektologie und Linguistik,* 75, 72–82.

Rabanus, S. (2013): "La cartografia linguistica del mòcheno", in: Bidese, E./Cognola, F. (eds.): *Introduzione alla linguistica del mòcheno*, Turin, 129-146.

Schweizer, B. (2012): *Zimbrischer und Fersentalerischer Sprachatlas/Atlante linguistico cimbro e mòcheno*. Edited and commented by S. Rabanus, Luserna, Palù del Fersina.

Stefan, B. (2000): Novena vun unzar liben Vraun. *Die Zimbrische Mariennovene des D. Giuseppe Strazzabosco mit Übersetzung und Kommentar*, Innsbruck.

Eine WebGIS-Anwendung für die Ortsamenforschung: Das Projekt „Die ältesten Ortsnamen im bayerisch-tschechischen Grenzraum“[1]

Alois Dicklberger, Universität Passau/Universität Regensburg

Am Lehrstuhl für Deutsche Sprachwissenschaft an der Universität Passau wird seit 2008 unter der Leitung von Prof. Harnisch das grenzübergreifende Forschungsprojekt „Die ältesten Ortsnamen im bayerisch-tschechischen Grenzraum“ durchgeführt. Das auf drei Jahre angelegte Projekt wurde gefördert von der Europäischen Union im Rahmen des „Europäischen Fond für regionale Entwicklung“ und der Universität Passau. Als Partner auf tschechischer Seite konnte das Museum der Stadt Prachatice gewonnen werden. Das Untersuchungsgebiet ist der Landkreis Freyung-Grafenau auf deutscher, und der benachbarte Okres Prachatice auf tschechischer Seite.

1 Das Projekt

1.1 Zielsetzung

- Mit diesem Projekt werden die ältesten überlieferten Ortsnamen im Untersuchungsgebiet nach wissenschaftlichem Standard erhoben und dokumentiert. Dies betrifft auf tschechischer Seite alle Ortsnamen, die bis 1360 urkundlich erstmals erwähnt sind, und auf bayerischer Seite alle Ortsnamen, die bis zum 13. Jahrhundert urkundlich zu fassen sind sowie Namen mit dem Grundwort *aha* (ahd. „Fluß, Gewässer“) und *heim* und dem Suffix *-ing*, die bis 1400 belegt sind. Auf beiden Seiten werden auch alle Gemeindehauptorte unabhängig davon, wann sie erstmals belegt sind, aufgenommen.
- Jeder Ortsname wird sprachwissenschaftlich fundiert erörtert und erklärt.
- Die erhobenen Daten und deren Auswertung werden in einer Datenbank erfasst und in einer WebGIS-Anwendung der Öffentlichkeit zugänglich gemacht.

1.2 Erhebung der Daten

a) Recherchieren und Erfassen der relevanten Quellen in Archivbeständen, um die historischen Namenformen zu erfassen. Unsere Arbeit wurde dadurch erheblich erleichtert, dass inzwischen viele der für unseren Raum relevanten Archivbestände online verfügbar sind.

b) Aufnahme der Mundartlautung bei ortsfesten älteren Sprechern mit einem USB-Rekorder und phonetische Transkription der mundartlichen Lautung der Ortsnamen.

c) Festhalten der Ortsansichten und deren Koordinaten mit einer GPS-fähigen Digitalkamera.

[1] Dieser Beitrag erscheint in leicht veränderter Form im Tagungsband „7. Tagung des Arbeitskreises für bayerisch-österreichische Namenforschung“.

1.3 Auswertung

Zu jedem Ort wird ein Artikel erstellt, der eine quellenkritisch gesicherte Belegreihe, die Mundartlautung und eine Erklärung enthält, die auf Grundlage dieser Daten die Ausgangsform erschließt. Das heißt, es wird die sprachliche Form zum Zeitpunkt der Namengebung erschlossen, die morphologische Struktur des Namens beschrieben und die lautgesetzliche Entwicklung hin zur heutigen Namenform erklärt. Ist die ursprüngliche Namenform ermittelt, können auch Rückschlüsse auf die Benennungsmotivik zum Zeitpunkt der Namensgebung gezogen werden[2].

1.4 Darstellung

Anders als in bisherigen toponomastischen Publikationen soll das Ergebnis nicht in erster Linie über ein Printmedium, sondern mittels einer WebGIS-Anwendung der Öffentlichkeit zugänglich gemacht werden. Der Benutzer soll mithilfe einer kartengestützten Benutzerführung und über Datenbankabfragen einen einfachen Zugang zu den gewünschten Daten erhalten.

1.5 Forschungsstand

Leider müssen wir feststellen, dass Niederbayern bisher, mehr noch als die Oberpfalz, in der Ortsnamenforschung vernachlässigt wurde. So ist etwa für Niederbayern mit *Griesbach im Rottal, Der ehemalige Landkreis* von Josef Egginger[3] bislang nur ein Band in der Reihe *Historisches Ortsnamenbuch von Bayern* (HONB) erschienen. Auch im *Lexikon bayerischer Ortsnamen* von Wolf-Armin Frhr. v. Reitzenstein[4] werden nur einige Ortsnamen aus unserem Untersuchungsgebiet behandelt. Zu nennen ist auch die etwas oberflächliche Dissertation von Alexander Glück zu den Ortsnamen des nördlichen Altlandkreis Passau.[5]

Für die tschechische Seite liegen vor allem die Abhandlung *Die Ortsnamen der Sudetenländer als Geschichtsquelle* von Ernst Schwarz[6] und Antonín Profous', Ortsnamenbuch von Böhmen[7] vor. Wie der Titel schon sagt, beschränkt sich Schwarz nur auf Ortsnamen im deutschsprachigen Gebiet der damaligen Tschechoslowakei und berücksichtigt auch hier nicht alle. Zur Erklärung und Deutung der entsprechenden Namen wurden meist nur wenige urkundliche Belege herangezogen, überwiegend zitiert er aus Regionalstudien vor den 30iger Jahren des letzten Jahrhunderts. Die darin enthaltenen Schreibformen stammen mehrheitlich aus älteren, z. T. unzuverlässigen Quelleneditionen. Schwarz verfügte daher über eine mangelhafte Beleggrundlage, was in einigen Fällen zu ungenauen oder unzutreffenden Herleitungen führt. Ähnliches gilt für das größtenteils

[2] Windberger-Heidenkummer, Erika: Mikrotoponyme im sozialen und kommunikativen Kontext. Flurnamen im Gerichtsbezirk Neumarkt in der Steiermark. Frankfurt am Main [u. a.]: Lang 2001 (= Schriften zur deutschen Sprache in Österreich 30).

[3] Egginger, Josef: Griesbach i. Rottal. Der ehemalige Landkreis. (Historisches Ortsnamenbuch von Bayern, Niederbayern 1) München: Kommission für bayerische Landesgeschichte 2011.

[4] Freiherr von Reitzenstein, Wolf-Armin: Lexikon bayerischer Ortsnamen, Herkunft und Bedeutung, Oberbayern, Niederbayern, Oberpfalz. München: C. H. Beck 2006.

[5] Glück, Alexander : Die ältesten Ortsnamen im nördlichen Altlandkreis Passau. Dissertation. LM Univ. München 2010.

[6] Schwarz, Ernst**:** Die Ortsnamen der Sudetenländer als Geschichtsquelle. 2. durchges. teilw. umgearb. und erw. Aufl. v. 1931. München, Lerche 1961 (= Handbuch der sudetendeutschen Kulturgeschichte 1).

[7] Profous, Antonín**:** Místní jména v Čechách (Ortsnamenbuch von Böhmen). Jejich vznik, původní význam a změny. Vol. 1-5. Prag **,**Nakl. Československé Akad. Věd 1947-1960.

von Profous verfasste fünfbändige Werk *Místní jména v Čechách*. Hier werden zwar neben tschechischen oftmals auch deutsche Namen behandelt. Allerdings wird meist eine nur sehr knappe Herleitung geboten, die in vielen Fällen nicht dem heutigen Forschungsstand entspricht.

1.6 Methodik

Bei den Erhebungen der mundartlichen Namenformen wird die topographische Lage der betreffenden Siedlungen mit Digitalfotos erfasst, die ebenso wie die Tondokumente der Dialektaussprache und die zugehörigen GIS-Koordinaten zum untersuchten Ort in eine Datenbank integriert werden.

Die Recherche der historischen Namenschreibungen wurde in verschiedenen Archiven durchgeführt (Hauptstaatsarchiv München, Národní archiv Praha, Státní oblastní archiv Třeboň, Státní okresní archiv Prachatice usw.), ebenso in kleineren Stadtarchiven und Sammlungen.

Im Laufe unserer Recherchen wurden immer mehr Archivbestände online zugänglich gemacht. Etwa ein Viertel der von uns gesichteten Urkunden und Regesten sind online verfügbar und können so auch den Benutzern unserer WebGIS-Anwendung zugänglich gemacht werden. Die Archivarbeit wurde damit wesentlich erleichtert.

Auch wird mit diesem Projekt zum ersten Mal eine breite Materialgrundlage zu den ältesten Ortsnamen im Untersuchungsgebiet erarbeitet.

Über sprachhistorische Gesetzmäßigkeiten werden die Ausgangsformen von über 300 Ortsnamen erschlossen, die dadurch dem deutschen, germanischen oder slawischen Siedlungsträger zugeordnet werden können. Durch Kartierung bestimmter Ortsnamentypen lassen sich Erkenntnisse über den Ablauf der Besiedlung erzielen. Die typologische Einordnung der einzelnen Ortsnamen wird ebenfalls in der Ortsnamendatenbank abgelegt, so dass mit individuellen Abfragen in der WebGIS-Anwendung auch eine räumliche Darstellung der verschiedenen Ortsnamentypen möglich wird.

Datierungen von Erstnennungen, Belegreihen, Dialektaussprache und Etymologie werden zu den erforschten Ortsnamen im bayerischen Grenzraum abrufbar sein, ebenso wie weitere digitale Informationen. Auch für den tschechischen Grenzraum haben alle Interessierten Zugriff auf die Datenbank, die in beiden Sprachen vorliegen wird. Somit wird Kulturraum diesseits und jenseits der Grenze multimedial erfahrbar gemacht, indem Ortsnamen, vor allem auch solche, die ihren Weg vom Tschechischen ins Deutsche oder umgekehrt genommen haben, abgefragt und aktuelle Erkenntnisse bezüglich Herkunft und Bedeutung dieser Namen nachgeschlagen werden können. Schulen und Behörden wie auch Privatleute haben Zugang zu der innovativen online-Datenbank.

2 OrtsNamenDatenbank

2.1 Datenbank- und WebGIS-gestützte Publikation in der Ortsnamenforschung

Die Datenbank für Eingabe, Recherche und Verfassen der Ortsnamenartikel basiert auf MS Access, das Frontend und die Datenbankstruktur beruhen auf einer Weiterentwicklung von "DA01", einer Datenbankanwendung zur Erstellung von Ortsartikeln, die für

das *Historische Ortsnamenbuch von Bayern* (HONB) erarbeitet wurde und von einigen Autoren verwendet wird.[8] Bei der von Markus Ohlenroth an unsere Bedürfnisse angepassten Version werden die erhobenen Daten in die verschiedenen Formulare eingegeben, in den entsprechenden Tabellen abgelegt und verknüpft. Je nach Datenherkunft und Datenformat stehen entsprechende Eingabefenster zur Verfügung.

In speziellen Editoren können die Belegreihen, Kommentare und Erklärungen angelegt werden. Diese Textsorten erfordern besondere Formateigenschaften wie Lautschrift, Texthervorhebung und Verweisstrukturen, die von dieser Anwendung mit Hilfe von XML zur Verfügung gestellt werden.

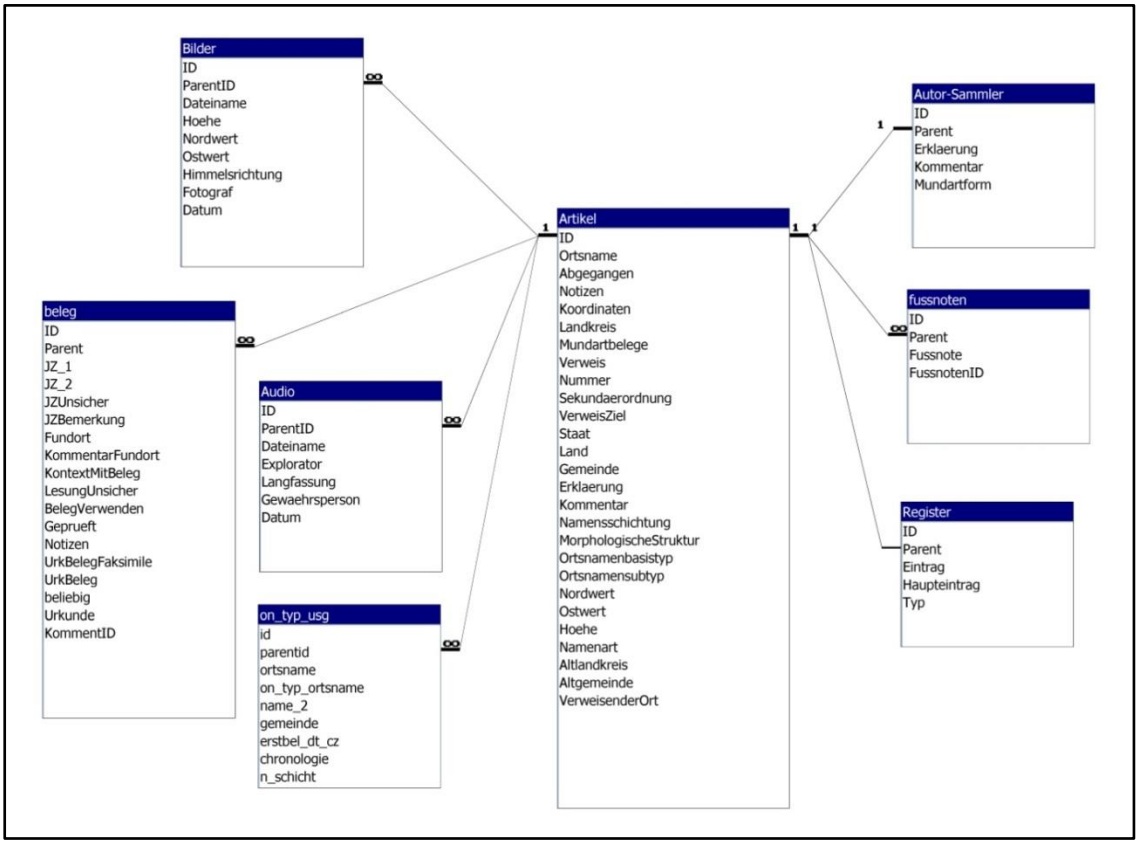


Abbildung 1: Struktur der OrtsnamenDatenbank

In speziellen Editoren können die Belegreihen, Kommentare und Erklärungen angelegt werden. Diese Textsorten erfordern besondere Formateigenschaften wie Lautschrift, Texthervorhebung und Verweisstrukturen, die von dieser Anwendung mit Hilfe von XML zur Verfügung gestellt werden.

Schon hier kann auf die strukturiert gespeicherten Daten zugegriffen werden. So kann der Bearbeiter einer Ortsnamenerklärung hier die historischen Belege sichten, ungeeignete oder nicht erforderliche von der Darstellung ausschließen und – falls vorhanden – auf das Original der Urkunde via Scan zurückgreifen. Er kann sich die Belegreihe chronologisch geordnet anzeigen lassen, die Mundartbelege einsehen oder gar die genaue Lau-

[8] Janka, Wolfgang: Konzeption und Methodik des *Historischen Ortsnamenbuchs von Bayern* (HONB). In: Arne Ziegler, Erika Windberger-Heidenkummer: Methoden der Namenforschung. Berlin, Akademie Verlag 2011. S. 169-180.

tung anhören und nicht zuletzt die Ortsansichten zur Klärung möglicher naturräumlicher Motivation der Namengebung zu Rate ziehen.

Die in der Accessdatenbank vorgehaltenen Daten können nach vielfältigen Kriterien selektiert und das Ergebnis im XML-Format ausgelesen werden. Mit einem entsprechenden XSL-Transformer können diese dann in unterschiedliche Formate übersetzt und ausgegeben werden: Zurzeit ist das Auslesen in MS-Word ab Version 2007 möglich, ein direkter Export in das Adobe-PDF Format ist geplant.

Mit der ONDa-Datenbankanwendung ist es möglich, Ortsartikel zu den einzelnen Orten im herkömmlichen Sinn zu erstellen.

Aber unser Ziel war es nicht nur, isoliert einzelne Ortsnamen zu erklären und diese Erklärungen im Netz abrufbar zu gestalten, sondern uns lag vor allem daran, das gesamte Material in einer Web 2.0-Technik zugänglich zu machen.

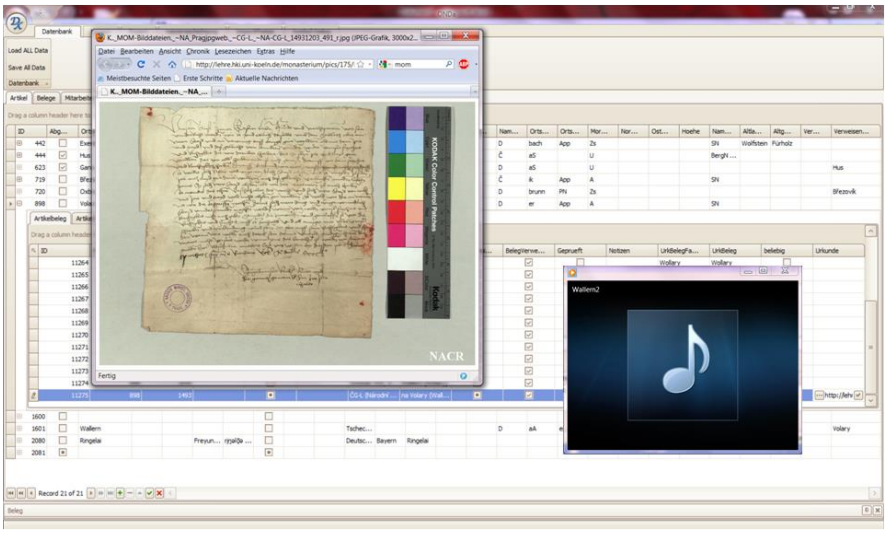


Abbildung 2: Screenshot ONDa, Urkundenansicht und Audiowiedergabe

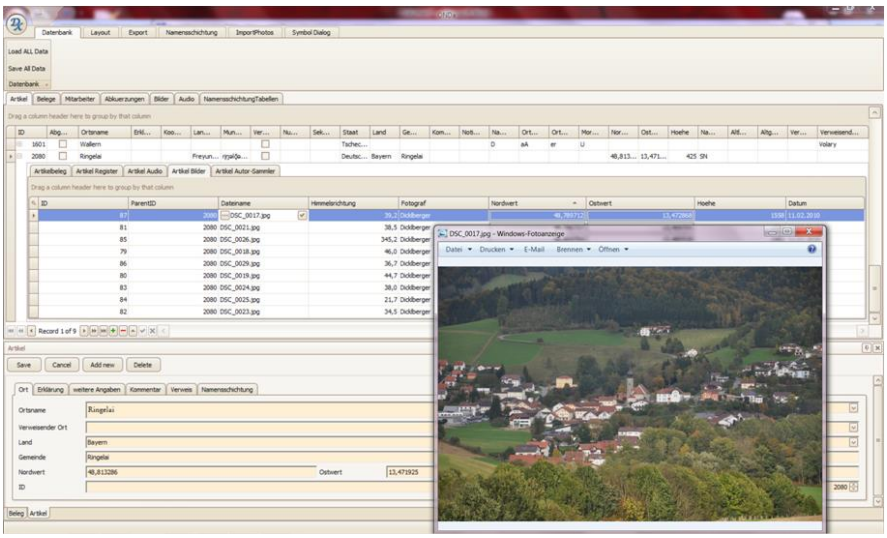


Abbildung 3: Screenshot ONDa, Fotoansicht

Dazu gehört die aktive Teilnahme des Benutzers bei der Selektion der Themen, der Realisierung eigener Abfragen in der Datenbank und der Darstellung dieser Ergebnisse in multimedialer Form.

So kann sich der Benutzer über eigene Datenbankabfragen die Verteilung unterschiedlicher Ortsnamentypen zeigen lassen. Er kann nach bestimmten Orten suchen und über diese Zugang zu den relevanten Daten erhalten.

Auch über die Kartenansicht kann er sich die von uns erhobenen Daten und die sprachwissenschaftlich fundierten Expertisen erschließen.

Vor allem aber wollen wir mit den derzeit verfügbaren technischen Möglichkeiten des internetbasierten Geografischen Informations-Systems die historische Entwicklung des Siedlungsraums an der deutsch-slawischen Sprachgrenze für den interessierten Laien, aber auch für den Fachwissenschaftler sichtbar machen.

Da MS Access keine GIS-Funktionalität bereitstellt und dessen Betrieb in heterogenen Netzwerken nicht so sicher ist, wurden die Daten in eine SQL-Datenbank portiert. Als SQL Datenbank verwenden wir PostgreSQL, da die Geodatenbank-Erweiterung PostGIS die derzeit leistungsfähigste auf dem Markt ist.

3 WebGIS-Auftritt des Projekts ONiG

Auch in der avancierteren toponomastischen Literatur ist es üblich, die geographische Verteilung von typischen Bildungsmustern oder von Elementen, die an der Bildung von Ortsnamen beteiligt sind, darzustellen. Hier ein Beispiel aus dem Ortsnamenbuch des Landes Oberösterreich.

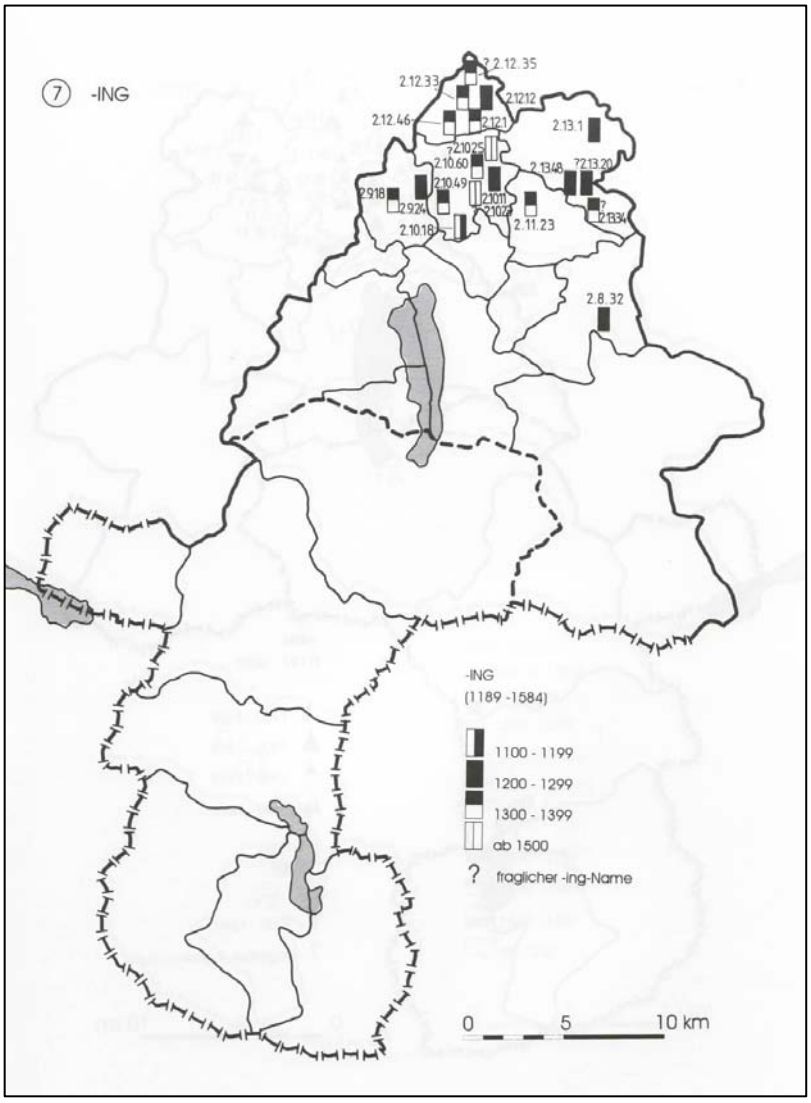


Abbildung 4: Reuter u. Wiesinger, Ortsnamenbuch des Landes Oberösterreich, Bd. 6, Karte 7

Karte 7 zur Geographie der Ortsnamen des Bezirks Gmunden[9] zeigt die Verteilung und das Jahr des Erstbelegs von Ortsnamen mit *ing*-Suffix. Diese Darstellungsform hat jedoch einige Nachteile gegenüber einer GIS-gestützten Kartographie:

- Es können nur Karten isoliert zu einem Phänomen dargestellt werden, eine beliebige Kombination von Karteninhalten ist nicht oder nur sehr schwer möglich (z.B. Beilegen transparenter Karten).
- Der Kartenbearbeiter wählt die behandelten Themen aus, der Rezipient hat auf diese Wahl keinen Einfluss.
- Karteninhalte können nicht mit externen Daten und Medien verlinkt werden.

Als Publikationsform war von Anfang an eine Web-Applikation geplant. Über die reine Darstellung von textlichen Elementen wie Belegreihen und Erklärungstext hinaus sollte diese Anwendung folgende Vorgaben verwirklichen:

[9] Reuter, Richard und Wiesinger, Peter: *Ortsnamenbuch des Landes Oberösterreich*. Bd. 6: *Die Ortsnamen des politischen Bezirks Gmunden,* Karte 7. Wien 1999.

- Die Inhalte sollten über das Internet allgemein und frei zugänglich sein.
- Das von uns erhobene und geeignete Ton- und Bildmaterial soll verfügbar gemacht werden.
- Die geografische Verteilung der bei der Ortsnamenbildung wirkenden sprachlichen Faktoren soll darstellbar sein.
- Über die Karte soll ein strukturierter Zugang auf das verfügbare Material ermöglicht werden.
- Da es sich um ein grenzüberschreitendes Projekt handelt, soll die Benutzeroberfläche zweisprachig realisiert werden, also tschechisch und deutsch.
- Alle Komponenten der Anwendung sollten gemäß Open Source Kriterien lizensiert sein.

Dieses Schema zeigt, wie diese Anforderungen in der Struktur realisiert werden können. Folgende Technologien kommen zum Einsatz:

- Webserver: Apache 2.2[10]
- Mapserver: UMN Mapserver 6.2[11]
- Scriptsprachen:
  - PHP[12]
  - Javascript[13]
  - Mapscript[14]
  - Datenbankserver: PostgreSQL 9.1.6[15]
  - CGI: ein Standard, der hier dem Datenaustausch zwischen Webserver und Mapserver dient.[16]

[10] http://httpd.apache.org/.
[11] http://mapserver.org/.
[12] http://php.net/.
[13] http://www.java.com/.
[14] http://mapserver.org/mapscript/index.html.
[15] http://www.postgresql.org/.
[16] http://www.w3.org/CGI/.

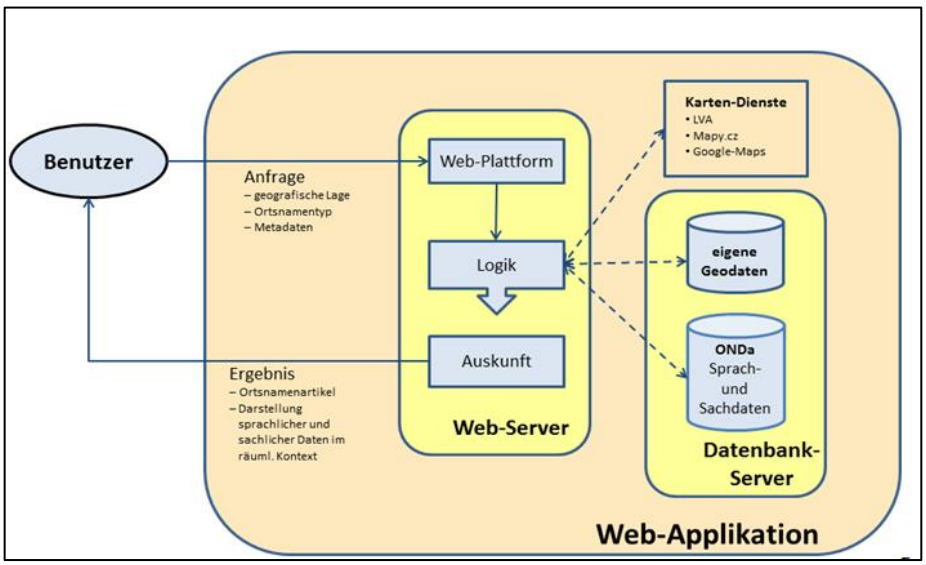


Abbildung 5: Schematische Darstellung der verwendeten Techniken.

4 Ortsnamenkarten heute

Hier einige Beispiele, die die Funktionalität dieser Anwendung zeigen (siehe Abbildung 6).

In der Bedienleiste im linken Rahmen sind folgende Auswahlmöglichkeiten gegeben:

- **Sprache**: Tschechisch oder Deutsch
- **Zoom/Pan:** Zoommodus und -faktor oder Panmodus
- Kartenmodus:
  - "Browse": Bewegen und Zoomen auf der Karte
  - "Abfrage": Inhalt zu einer Ebene abfragen, z.B. Ortsname, Gemeinde etc.
  - "n-Abfrage": Inhalt zu mehreren Ebenen abfragen, z.B. Ortsname und Gemeinde etc.
- **Abfragbare Layer:** Diese Layer oder Ebenen werden auf unserem eigenen Server vorgehalten, da nur so sinnvoll Abfragen zu räumlichen Daten durchgeführt werden können.
  - Gemeinden (Gemeindegrenzen, Gemeindeschlüssel und Name, in PostGIS abgelegt, © LVG, © cenia)
  - Ortsnamen (ebenfalls in PostGIS gespeichert, enthält die umfangreichste Datensammlung)
  - Namensschicht
  - Morphologische Struktur
  - Namenbasistyp
  - Namensubtyp

Abbildung 6: Kartenausschnitt mit Ortsnamen und Orthophoto des LVG Bayern im Hintergrund

- **Verschiedene Hintergrundkarten:** Hier handelt es sich in erster Linie um Ebenen, die von externen Kartenservern als sogenannte WebMapServices (WMS) bezogen werden. Ihre Qualität ist abhängig vom Anbieter, der sie zur Verfügung stellt (Vergl. Abbildungen 7 bis 11).
  - Uraufnahmeblätter[17] (© LVG Bayern)
  - Franziszeische Landesaufnahme[18] (© cenia[19])
  - topografische Karten (© LVG Bayern)
  - Orthophotos (© cenia, © LVG Bayern)
  - geologische Formation (© cenia, © LVG Bayern)
  - Bodengüte (© LVG Bayern)

Einige Beispiele zeigen, welche Hintergrundinformationen abrufbar sind.

[17] Die Uraufnahmeblätter sind in Bayern zwischen 1808 bis 1864 im Maßstab 1:5000 und kleiner erstellt worden.
[18] Die Franziszeischen Landesaufnahmen sind das habsburgische Pendant zu den Uraufnahmeblättern und wurden von 1806 bis 1869 erarbeitet.
[19] Česká informační agentura životního prostředí (CENIA), Vršovická 1442/65, Praha 10, 100 10.

Abbildung 7: Ortsnamen und Franziszeische Landesaufnahme im Hintergrund

4.1 Abfragemöglichkeiten

Über diese Kartenhintergründe kann der Benutzer die von ihm gewünschten thematischen Karten legen. Hier z.B. die Namenschichtung: Abbildung 12 zeigt, dass in unserem Untersuchungsgebiet fast nur deutsche und tschechische Ortsnamen belegt sind. Die Typologisierung ist noch nicht abgeschlossen, so dass noch kein endgültiges Ergebnis vorliegt. Tschechisierte deutsche Ortsnamen und eingedeutschte tschechische Ortsnamen sind sicher zu belegen. Eher unwahrscheinlich ist das Auftreten von Ortsnamen germanischer, keltischer, indogermanisch-voreinzelsprachlicher Schicht.

Abbildung 13 zeigt z.B., ob das Bestimmungswort oder die Ableitungsbasis ein Appellativum oder ein Personenname ist. An der Symbolverteilung ist zu erkennen, dass im später besiedelten grenznahen und geographisch höher gelegenen Bereich an die Stelle von Personennamen Appellativa treten.

Abbildung 8: Ortsnamen und topographische Karte im Hintergrund

Abbildung 9: Ortsnamen und Orthophoto als Hintergrund

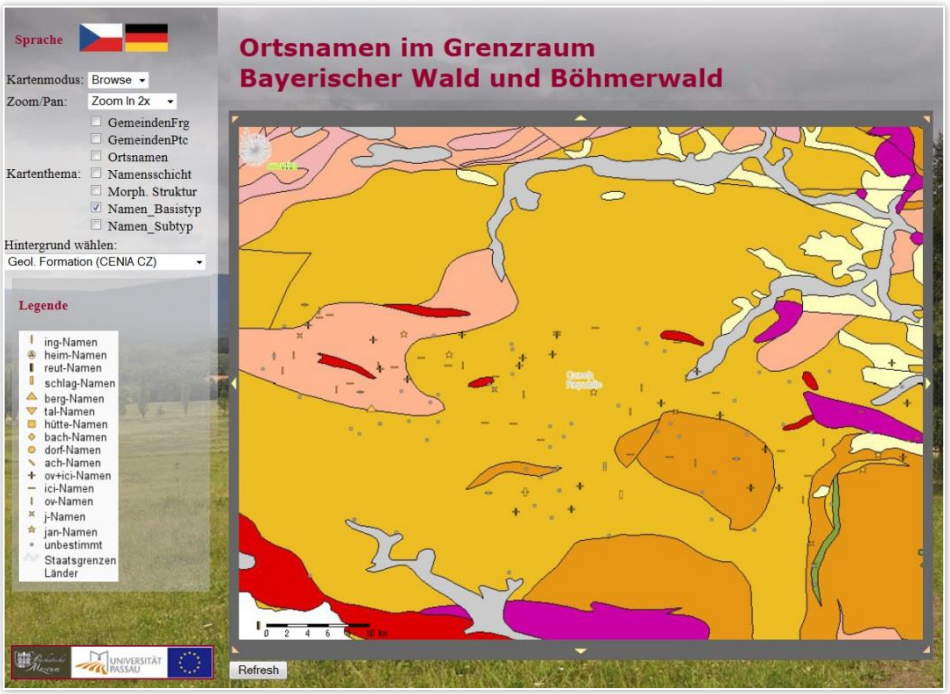


Abbildung 10: Ortsnamentyplogie mit geologischer Formation als Hintergrund

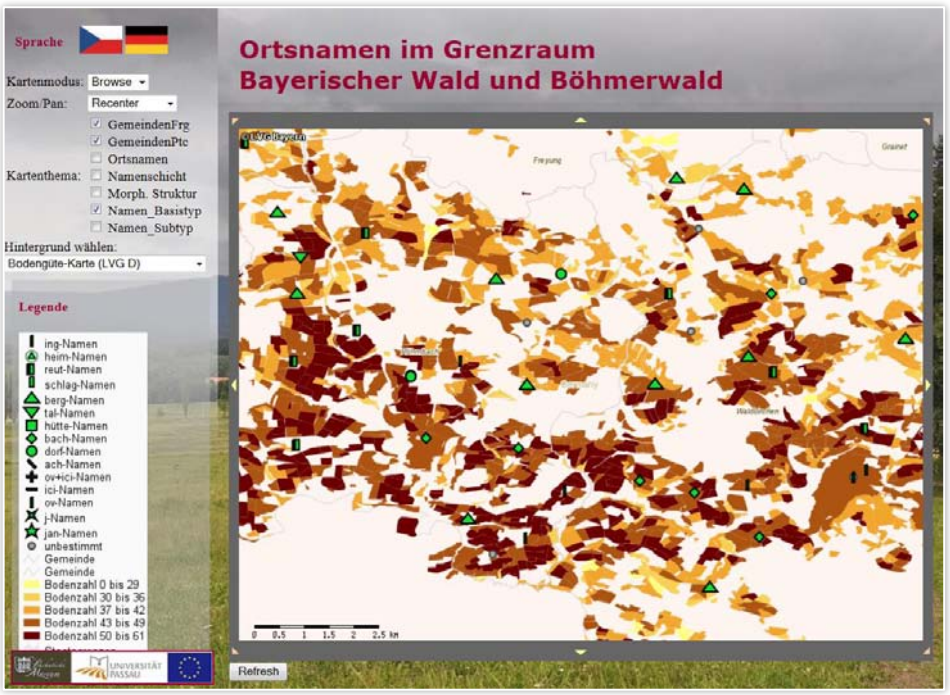


Abbildung 11: Ortsnamensubtyp mit Bodengüte als Hintergrund

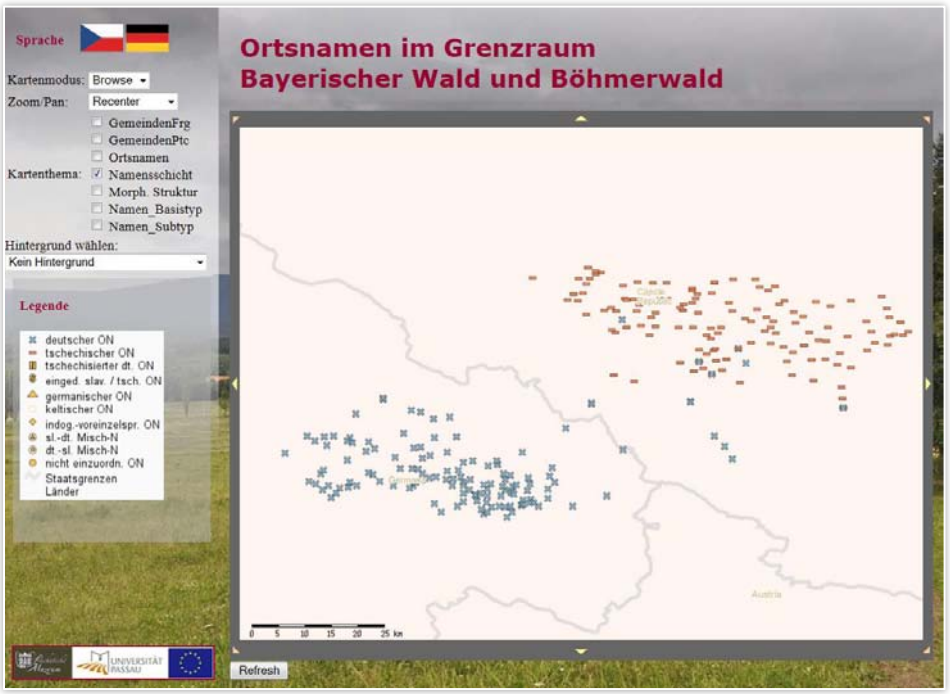


Abbildung 12: Darstellung der Namenschichtung

Später werden auch noch historische Karten (z. B. kirchliche und weltliche Herrschaftsgebiete) hinzukommen.

Weitere Funktionalitäten sind in Vorbereitung. So soll in einem weiteren Schritt das Jahr des Erstbelegs aller Ortsnamen symbolisch dargestellt werden. Damit könnte man die Siedlungsbewegungen veranschaulichen und zugleich gewisse Wortbildungsmuster optisch mit dem Zeitraum ihrer größten Verbreitung in Beziehung setzen.

Bisher können einzelne Wortbildungsmuster und -typen abgebildet werden. Für eine tiefergehende sprachgeographische Analyse des Materials wäre aber auch eine vom Benutzer selbst definierbare Kombination aller an der Wortschöpfung beteiligten Faktoren nützlich.

4.2 Fokussieren kleinräumiger Phänomene

Des Weiteren können mit unserer Anwendung Fragen des Terrains beantwortet werden, die für die Erklärung von Ortsnamen aber auch von Mikrotoponymen von Bedeutung sind. Über Uraufnahmen können Flussverläufe und Wege aus der ersten Hälfte des 19. Jahrhunderts erschlossen werden: Mittels Höhenlinien werden Eigenheiten des Geländes deutlich (vgl. Abbildung 15).

4.3 Informationen zu einzelnen Ortsnamen abrufen

Nach Umschalten auf den Abfragemodus können durch Anklicken des gewünschten Ortes oder durch eine Datenbankabfrage weitere Informationen zu den Ortsnamen abgerufen werden.

Über dieses Fenster, das sich nach der Auswahl eines Ortes öffnet, können weitere Angaben zum Ort eingesehen werden, und von einigen Feldern aus ist es möglich, weitere Dienste aufzurufen.

Abbildung 13: Darstellung des Ortsnamensubtyps

Abbildung 14: Detailansicht Uraufnahmeblätter

Abbildung 15: Detailansicht Topographie

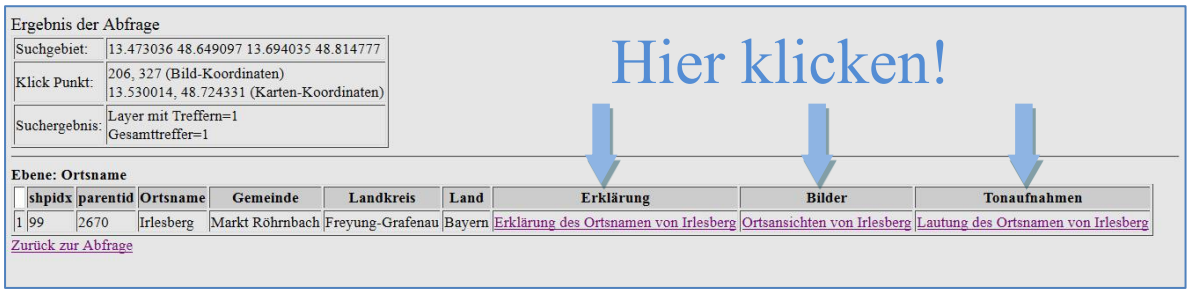


Abbildung 16: Abfrageergebnis

So kann durch einen Klick auf das Feld "Erklärung" der Artikel zu dem entsprechenden Ort, hier Irlesberg, im pdf-Format geöffnet werden, die Ortsansichten können über das Feld "Bilder" betrachtet und die Lautung des Ortsnamen über "Tonaufnahmen" angehört werden.

In einer weiteren Ausbaustufe soll neben den oben angeführten Erweiterungen der Funktionalität auch die Benutzerführung einfacher und intuitiver gestaltet werden.

Damit aber das Instrument zur Analyse und räumlichen Darstellung von Ortsnamentypologien aussagekräftige Ergebnisse liefern kann, muss das Untersuchungsgebiet beträchtlich erweitert werden. Erst mit der Einbeziehung des gesamten bayrisch-tschechischen Grenzraumes in die Ortsnamentypologie könnte man solidere Aussagen über den Verlauf der Besiedlung und über Verbreitung und Wandel von Bildungsmustern und -motiven bei Ortsnamen treffen.

Bibliographie

Egginger, Josef (2011): Griesbach i. Rottal. Der ehemalige Landkreis. (Historisches Ortsnamenbuch von Bayern, Niederbayern 1) München: Kommission für bayerische Landesgeschichte.

Glück, Alexander (2010): *Die ältesten Ortsnamen im nördlichen Altlandkreis Passau*. Dissertation. LM Univ. München.

Janka, Wolfgang (2011): Konzeption und Methodik des *Historischen Ortsnamenbuchs von Bayern* (HONB). In: Arne Ziegler, Erika Windberger-Heidenkummer: *Methoden der Namenforschung*. Berlin.

Mitchell, Tayler (2008): *Web-Mapping mit Open-Source-GIS-Tools*. Deutsche Ausgabe überarbeitet von Astrid Emde und Arnulf Christl.

Profous, Antonín (1947-60): *Místní jména v Čechách* (Ortsnamenbuch von Böhmen). *Jejich vznik, původní význam a změny*. Vol. 1-5. Prag.

Reitzenstein, Wolf-Armin Freiherr von (2006): Lexikon bayerischer Ortsnamen, Herkunft und Bedeutung, Oberbayern, Niederbayern, Oberpfalz. München.

Schwarz, Ernst (1961): Die Ortsnamen der Sudetenländer als Geschichtsquelle. 2. durchges. teilw. umgearb. und erw. Aufl. (Handbuch der sudetendeutschen Kulturgeschichte 1), München.

Windberger-Heidenkummer, Erika (2001): *Mikrotoponyme im sozialen und kommunikativen Kontext. Flurnamen im Gerichtsbezirk Neumarkt in der Steiermark* (= Schriften zur deutschen Sprache in Österreich 30), Frankfurt am Main [u. a.].

Reuter, Richard/Wiesinger, Peter (Hrsg.) (1999): Die Ortsnamen des Politischen Bezirks Gmünden (Südwestliches Traunviertel). Ortsnamenbuch des Landes Oberösterreich, Bd. 6, Wien.

L'Atlas Multimédia Prosodique de l'Espace Roman: uno strumento per lo studio della variazione geoprosodica[1]

A. Romano, Università degli Studi di Torino, Dip. Lingue e L. S. e C. M. e LFSAG – Laboratorio di Fonetica Sperimentale "Arturo Genre", Italia & M. Contini & J.P. Lai, Université Stendhal Grenoble III – GIPSA-Lab, Francia

1 Introduzione

Il progetto dell'*Atlas Multimédia Prosodique de l'Espace Roman* (*AMPER*), nato da un'idea di Michel Contini (presentata a Bilbao nel 1991; cfr. Contini 1992), è un 'cantiere aperto'. La sua struttura è talmente ramificata al punto che i nuovi partner, che si associano in continuazione (interessati a uno studio fonetico e dialettologico delle variabili prosodiche), spesso non si conoscono tra loro e non hanno contatti immediati con la coordinazione della rete. Fondandosi su fonti di finanziamento locali e individuali, non è neanche vincolato da scadenze amministrative e, perciò, ha la possibilità di crescere e allargarsi con tempi e modalità molto "permissivi" (i cui vantaggi e svantaggi sono elencati, almeno relativamente alla rete *AMPER-ITA*, limitata al solo dominio italo-romanzo, in Romano, in c. di p.)[2].

L'obiettivo generale del progetto è la descrizione della variazione della prosodia nelle parlate romanze di tutto il pianeta mediante un confronto, almeno nelle prime fasi esclusivamente fonetico, tra i dati di una Base di Dati sonora (*BD-AMPER*), preventivamente raccolti e analizzati con una metodologia comune.

Una delle motivazioni originarie del progetto (anticipata in Contini 1992) è legata alla considerazione che le ricerche in prosodia condotte tradizionalmente nello spazio romanzo (così come anche in altri spazi linguistici) si erano concentrate fino a quel momento allo studio di fenomeni diversi dipendenti da caratteristiche specifiche del singolo dominio dialettale (alcuni Paesi erano interessati maggiormente allo studio di accentazione e, ad es., relazioni tra intonazione e punteggiatura, altri Paesi si erano rivelati come i pionieri nello studio della prosodia metrica o del ritmo del parlato oppure ancora domini interessati – per via dell'indipendenza linguistica di alcune loro aree – ad approfondire gli aspetti della variazione diatopica dell'intonazione; cfr. Romano 2001a&b). Inoltre, anche in funzione di un diverso investimento nelle distinte aree nello sviluppo di metodi strumentali, a ciascuna tradizione specifica corrispondevano anche modelli teorici e/o sperimentali diversi che conducevano a risultati quasi sempre incomparabili.

In generale, fino a tutti gli anni '80, si constatava (e, in parte, ciò accade ancora oggi) una grande diffusione di studi basati sul parlato letto e/o di laboratorio e su una visione letteraria delle lingue (laddove molti dialetti non hanno una tradizione scritta)[3].

---

[1] Sebbene impostato nelle sue grandi linee congiuntamente dai tre autori, l'articolo è stato scritto da AR.

[2] Mentre la coordinazione generale del progetto si divide fra Grenoble (M. Contini e segreteria) e Torino (A. Romano), la coordinazione delle ricerche condotte per *AMPER-ITA* è affidata al *LFSAG* di Torino.

[3] Per lo spazio italo-romanzo, un'eccezione era rappresentata da alcuni lavori pionieristici che, comunque, non erano andati al di là di una prima esplorazione sommaria (cfr. Panconcelli-Calzia 1939). Più esaustivamente si erano affermati in quegli anni gli studi originali proposti da L. Canepari (v. Canepari 1985) basati tuttavia su dati e risultati non sempre verificabili. Dal lancio del progetto *AMPER* molte cose sono cambiate: anche piuttosto recentemente altri progetti hanno intrapreso la stessa via e, seppur con metodi, strumenti e finalità diversi, stanno perseguendo la realizzazione di basi di dati simili alla nostra. Orientati quasi esclusivamente all'allestimento di dati relativi alle varietà regionali delle lingue nazionali (tra l'altro rappresentate anche in diverse sezioni del progetto *AMPER*) questi progetti si situano nell'ambito di mo-

Oltre alla trascuratezza generalmente riservata in certi ambiti di studio alle numerose fonti di variazioni linguistica della prosodia[4], un ultimo punto d'insoddisfazione degli studi prosodici condotti in quegli anni era nella difficoltà d'illustrare in modo convincente gli snodi prosodici rilevanti ai fini della caratterizzazione (o, in alcuni casi più avanzati, del 'funzionamento') della parlata e del confronto tra il suo 'sistema' e quello delle parlate simili o con cui, comunque, si poneva in contrasto[5]. Benché le modalità di raffigurazione delle curve (o delle sequenze) di valori assunti dalle variabili prosodiche siano ancora oggi spesso insoddisfacenti e/o disuniformi (in termini di unità di misura, scala, qualità dell'analisi), si sono diffusi strumenti analitici (*software*) che rendono molto agevole la creazione e la manipolazione di grafici che superino questa difficoltà[6].

Anche la cartografazione dei dati o l'allestimento di sistemi di consultazione dei dati sonori hanno permesso di risolvere molti degli originari difetti di questi studi. Ovviamente nell'ambito di *AMPER* la finalità atlantistica era presente sin dal suo lancio, con gli obiettivi di rappresentazione cartografica che nell'ambito della prosodia erano suggeriti da lavori pionieristici come quelli di E. Gårding & G. Bruce (v. tra gli altri Bruce & Gårding 1978). Il riferimento alla realizzazione dei cosiddetti "Atlanti parlanti" (alla cui idea originaria ha contribuito tra l'altro anche il Centre de Dialectologie de Grenoble, *CDG*, suggerendo linee di sviluppo e applicazione microareali) era ovviamente tra gli obiettivi iniziali e ha in parte trovato una sua prima realizzazione, parziale e dimostrativa, nel DVD di *AMPER* (2011)[7].

delli interpretativi fonologici di largo consenso e si avvalgono di tecniche di elicitazione d'ispirazione dialettologica simili a quelle proposte in *AMPER* sin dalla fine degli anni '90.

[4] Sebbene in misura variabile da dominio a dominio, l'intonazione ad es. è marcata sul piano diatopico, ma anche diastratico, diafasico e diamesico (v. §3), e risente notoriamente di numerosi fattori extralinguistici. Quanto ai condizionamenti subdoli che l'intonazione di frase può subire per effetto di queste dimensioni di variazione (oltre che per le difficoltà dell'elicitazione) si veda Lai *et alii* (1997).

[5] Ciò ha condotto talvolta allo sviluppo di *savoir-faire* dipendenti 'localmente' dalle caratteristiche specifiche delle varietà studiate (e sta contribuendo a ritardare ancora, ad es., l'affermazione di una teoria generale dell'intonazione; la diffusione di modelli semplic(istic)i e, spesso, autoreferenziali si associa infatti alla tendenza di alcuni autori a sviluppare una sorta di diffidenza per gli altri approcci; v. sotto).

[6] Alcuni dei motivi che rendevano difficili l'integrazione tra le informazioni fornite dai vari studi erano stati illustrati in Contini (1992) e sono ora attualizzati in Contini (2008) dove sono ricordati anche altri difetti della ricerca in questo settore che portano all'incomparabilità dei dati e all'impossibilità di allestire un quadro d'insieme a partire da questi (alcune riflessioni sul tema sono anche in Romano *et alii* 2011). Tenendo conto delle osservazioni di diversi specialisti, registriamo inoltre, negli ultimi anni, l'affermarsi di nuove sorgenti di dispersione. L'impossibilità del confronto è oggi legata anche a: la diversità nei materiali raccolti e nelle metodologie di raccolta; la presenza nei dati di fattori di condizionamento (in molti casi controllabili, ma in genere diversamente controllati); la mancanza di metodi d'analisi obiettivi (rivolti dapprima alla comprensione della natura fonetica dei dati raccolti cui si aggiunge la confusione imperante tra i livelli di analisi segnalata tra gli altri da Hirst *et alii* 2000); le difficoltà nell'allestire uno stato dell'arte completo ed esaustivo (alcuni ricercatori riscoprono individualmente, dopo un certo investimento in termini di tempo e di energia, la presenza di variabili fino a quel momento trascurate nei loro studi e che, magari, erano state al centro delle attenzioni di altri gruppi di ricerca con altro orientamento) e la mancanza di una formazione multidisciplinare appropriata per i ricercatori che manipolano i dati prosodici (alcuni autori hanno rilevato come lavori recenti si siano talvolta basati su analisi strumentali grezze e su dati inaffidabili – si veda ad es. quello segnalato da Romano & Interlandi 2002 –, oppure sull'abuso di strumenti statistici applicati alla descrizione di fatti poco rilevanti suggeriti da ipotesi di partenza viziate; cfr. Martin 2003, Romano 2003, 2005, Martin 2012).

[7] Anche in questo caso, oltre alle realizzazioni multimediali di I. Marquet e J. Médélice del *CDG*, l'idea s'ispira *mutatis mutandis* ai progetti portati avanti per lo studio della variazione dialettale in Svezia (come ad es. la sezione d'interesse geoprosodico nel progetto *SweDia2000*, *Database tools for a prosodic analysis of the Swedish dialects*, a cura del gruppo di lavoro coordinato da G. Bruce *et alii*). Nel nostro caso, oltre alla particolare natura della sezione di dati che siamo riusciti finora ad allestire (ristretta a una selezione di dati relativi al corpus 'fisso', v. §2), la realizzazione del DVD ha risentito del problema della

2 Obiettivi del progetto *AMPER*, procedure analitiche e modellizzazione

Nel corso delle tappe fondamentali della sua genesi ventennale (1991, 2001 e 2011)[8], il progetto *AMPER* ha cercato di coinvolgere specialisti nello studio dialettologico dei fatti prosodici puntando in più occasioni sull'interesse per alcuni obiettivi di portata molto ampia.

In effetti, pur provenendo da ambiti di ricerca piuttosto differenziati, i suoi partner sono in genere interessati a perseguire un approccio descrittivo e variazionale, orientato a uno studio geoprosodico condotto sui dialetti romanzi storici e sulle varietà regionali delle lingue nazionali dei Paesi di questo spazio dialettale.

Oltre a permettere allo specialista di condurre le sue analisi ai diversi livelli (distinguendo interessi e ambiti di ricerca 'intonetici' e 'intonologici'), l'allestimento dei dati è finalizzato alla diffusione di dati comparabili raccolti su un ambito territoriale decisamente vasto ed è condotto anche in vista di finalità formative e divulgative.

Le modalità di raggiungimento di questi obiettivi sono approfondite nel §2: anticipiamo qui soltanto la necessità essenziale di basarsi su corpora di parlato progettati e realizzati partendo da assunzioni di partenza identiche e seguendo procedure conformi il più possibile a un protocollo condiviso applicato in condizioni almeno simili.

Per fare ciò, pur preservando la libertà di approfondire materiali specifici definiti di volta in volta in base alle peculiarità della varietà studiata (sulla base di un corpus 'libero'), occorre accettare la necessità di riferirsi a un corpus comparabile basato su un questionario con caratteristiche comuni (corpus 'fisso / sperimentale', come concordato coi primi partner del progetto e come ribadito in Contini 2008).

Le verifiche possono poi essere condotte su corpora spontanei o (semi-)spontanei (procedendo nelle modalità illustrate da alcuni partecipanti che hanno cominciato a seguire questa strada; v. contributi in Turculeţ 2008 e la *BD-VIAP* di *AMPER-ITA*)[9].

In sintesi, quindi, il trattamento dei dati dalla raccolta alla valutazione finale si esegue attraverso un certo numero di tappe analitiche che prevedono:

1. la definizione di una versione locale del 'Questionario Comune di Base' (*QCB*);
2. lo svolgimento d'inchieste sul campo e il ricorso a metodi di elicitazione dialettologici;
3. il rispetto di specifiche sulla qualità delle registrazioni e sul formato finale dei dati sonori (*.wav*);
4. l'esecuzione di un'analisi strumentale multiparametrica (frequenza fondamentale - f_0, durata - D e intensità relativa / energia - E) finalizzata all'ottenimento di file di dati numerici comparabili (stilizzazione → file di dati di tipo *.txt*);

riservatezza dei dati sonori che ha indotto alcuni partner a riservarsi l'uso esclusivo delle registrazioni originali (ragion per cui, anche l'accesso alla *BD on-line* è tuttora ristretto).

[8] In seguito al primo annuncio dell'idea (proposto, come anticipato nell'Introduzione, nell'intervento di M. Contini al « Nazioarteko Dialektologia Biltzarra Agiriak » di Bilbao, 1991; cfr. Contini 1992), il lancio effettivo del progetto può essere considerato coincidente col 1º *workshop* di *AMPER* organizzato dieci anni dopo da M. Contini, J.P. Lai e A. Romano a Grenoble nel 2001 alla presenza di un primo nucleo di colleghi in rappresentanza di tutti i domini romanzi (v. *Grenoble* 2001). Ad altri dieci anni di distanza da questo, segue la pubblicazione di una prima *tranche* di dati (v. *AMPER* 2011; cfr. §4).

[9] Le prime proposte sono in Romano (2001a) e Lai (2002).

5. la disamina di sequenze di valori di f_o, D e E di ripetizioni diverse attraverso l'osservazione di andamenti medi e la considerazione degli scarti/deviazioni (1ª prototipizzazione → file di dati numerici *o.txt*);
6. la verifica percettiva mediante test d'ascolto e/o di percezione di versioni sintetiche dei prototipi (file sonori di tipo *.ton* o *_ton.wav*);
7. una modellizzazione linguistica (2ª prototipizzazione) derivante dal confronto tra le strutture presenti nel *QCB* e tra queste e i dati presenti nei corpora 'liberi'.

Partendo dai risultati raggiunti in 5) o 6) è possibile procedere al caricamento dei dati nella *BD* e un'analisi trasversale dei dati, la quale si può svolgere in termini impressionistici (eventualmente tipologici e/o cartografici, come proposto in Romano 2004) o sulla base di valutazioni fonologiche (cioè in base al trattamento automatico di misure di correlazione e/o distanze dialettometriche oppure in base a procedure di clusterizzazione di modelli definiti in termini di tratti tipologici, v. §5).

Una prima valutazione del grado di accettazione dei vincoli del progetto da parte dei partner che finora hanno aderito al piano di lavoro è offerta in Contini *et alii* (2009). Secondo questa, è possibile stabilire l'affidabilità dei dati forniti / conferiti alla *BD*, in base alla considerazione del numero di tappe analitiche raggiunte per i dati di una data inchiesta.

Si ha quindi un grado '0' nel caso in cui siano stati rispettati soltanto i vincoli su tipi e strutture del corpus e sulle condizioni di registrazione. Un grado '1' si ottiene ricorrendo alle procedure d'analisi sviluppate nell'ambito del progetto (*AMPER-fox*, *AMPER-2006*, *AMPER-pour-PRAAT* etc. v. §5). Il ricorso a metodi di valutazione preliminare basati su rappresentazioni grafiche convenzionali consente di raggiungere un grado '2'[10].

I gradi successivi si valutano tenendo conto dei lavori di approfondimento sui dati pubblicati dagli autori delle inchieste e/o delle analisi.

Un grado '3' corrisponde all'uso di una tecnica di lettura dei dati (e dei tracciati) obiettiva e rigorosa in termini fonetici (non condizionata o filtrata da una classificazione precoce degli eventi osservati in base a modelli teorici di tipo fonologico). Un grado '4' corrisponde infine all'adozione di una procedura analitica contrastiva di confronto intermodale e interdialettale confermata da test d'ascolto e/o di percezione.

3 Il metodo *AMPER*

Come si deduce dalle modalità di conseguimento degli obiettivi anticipate nel §1, il metodo di valutazione messo a punto nell'ambito di questo progetto si ripropone l'osservazione (e quindi la misurazione) di diverse variabili che riflettono l'organizzazione prosodica degli enunciati (e cioè almeno f_o, D e E). Le strutture selezionate per il *QCB*, come anticipato in Lai *et alii* (1997), prevedono – nei limiti del possibile – il ricorso a sequenze segmentali di tipo CVCV...CV, con C = consonanti sorde, vincolo che ha imbarazzato alcuni dei partecipanti (e ne ha scoraggiati altri potenziali) perché – come noto – questo

[10] Se gli autori dell'inchiesta non hanno svolto un'osservazione preliminare delle sequenze di valori relative alle diverse ripetizioni, modalità, strutture etc. si ottengono insiemi di dati non omomorfici al livello segmentale, tali cioè da non consentire l'allineamento degli eventi presenti in posizioni strutturali corrispondenti. La mancata esecuzione di un confronto grafico preliminare produce inoltre in certi casi, profili variabili da struttura a struttura per la stessa modalità o per sequenze identiche di posizioni accentuali.

impedisce d'individuare l'esatto allineamento di quelli che in certi approcci fonologici sono considerati bersagli tonali.

Nel nostro caso, la perdita di queste informazioni è ampiamente compensata dalla maggiore facilità di segmentazione (che in futuro potrà essere affidata anche a procedure semi-automatiche, permettendo così di allargare in modo molto rapido e quantitativamente più soddisfacente l'insieme di dati su cui si svolgono le valutazioni)[11].

In funzione del tipo di strutture segmentali prescelto e una volta definite, dunque, le vocali come sedi privilegiate (anche se, ovviamente, non esclusiva) della codifica delle informazioni prosodiche, sugli enunciati segmentati si esegue, secondo un protocollo prestabilito, la misurazione di valori delle tre variabili f_o, D e E al cui sviluppo temporale è associata la codifica prosodica essenziale dell'enunciato (come prova il riascolto della versione risintetizzata a partire da questi)[12].

Queste sequenze di valori rappresentano quindi il risultato di una prima stilizzazione di ciascuna ripetizione della stessa struttura del *QCB*. A partire dalle versioni stilizzate di enunciati coerenti corrispondenti alla stessa frase si ottengono poi gli andamenti medi sui quali si eseguono le operazioni successive di prototipizzazione (e verifica percettiva) e modellizzazione (v. esempi al §3.2)[13].

3.1 Gli strumenti e le procedure d'analisi

Alla definizione progressiva degli strumenti d'analisi diffusi nell'uso delle diverse *équipe* che aderiscono al progetto ho dedicato una rassegna in Romano (2008).

Una prima versione delle *routine* predisposte per agevolare l'applicazione del metodo descritto al §2 (poi nota localmente *ante litteram* come *CDG-AMPER*) risale agli anni 1997-1999 (v. Romano 2001a) ed era sviluppata nell'ambito di Matlab™.

Alcune revisioni tecniche e l'aggiunta di nuove funzioni in grado di permettere di gestire fenomeni, come la cancellazione vocalica, la cui presenza è stata ritenuta prosodicamente rilevante sul piano ritmico-intonativo, hanno condotto a una seconda versione delle *routine* originarie, sviluppate con l'aiuto di (e, in parte, presso) il Departamento de Línguas e Culturas dell'Università di Aveiro (Portogallo)[14]. Le *routine* sono state in segui-

[11] Come osservato in Romano (2001a), queste strutture hanno lo stesso contenuto prosodico di quelle con sole consonanti sonore. Inoltre, come facilmente verificabile osservando il parlato naturale, l'efficacia prosodica di enunciati di questo tipo è prosodicamente equipotenziale a quella di enunciati con consonanti sonore in grado di far percepire l'allineamento di questi presunti bersagli (il cui raggiungimento non è – evidentemente – condizione necessaria ai fini funzionali).

[12] Per ogni nucleo sillabico si rileva un valore di durata e d'energia, ma tre distinti valori di f_o (all'inizio, alla fine e nel punto di cambiamento di pendenza della curva che si sviluppa in corrispondenza di questo segmento (nelle misurazioni eseguite con *AMPER-pour-PRAAT* la posizione della misurazione centrale è ridefinita, per semplicità, nel punto medio, v. dopo).

13 Sebbene quest'ultima operazione sia stata proposta in un'ottica sovrapposizionale e sia stata apparentemente rifiutata da alcuni coordinatori d'area del progetto, la natura stessa del corpus allestito ha indotto alcuni partner a descrivere i fenomeni osservati in termini di organizzazione locale e globale accettando implicitamente parte della modellizzazione proposta dalla coordinazione generale del progetto. Secondo questa visione, lo stesso insieme di parametri acustici è ritenuto responsabile (1) della messa in rilievo locale di unità complesse nella catena segmentale (accentuazione) e (2) dei fenomeni relativi alla strutturazione intonativa (modalità, gerarchizzazione tra costituenti, focus, organizzazione informativa).

[14] In particolare questa revisione (richiesta dalla collega L. de Castro Moutinho) ha introdotto la possibilità di segnalare la posizione in cui, pur non essendo possibile misurare valori di f_o, è presente una vocale 'cancellata' ancora parzialmente percepibile in molti casi per via del rilascio della consonante precedente. Alcune operazioni aggiuntive autorizzano ripensamenti nella segmentazione e danno la possibilità di cancellare i demarcatori permettendo l'ascolto dei segmenti da questi provvisoriamente delimitati. Altre novità di questa revisione (che dobbiamo ad A. Teixeira, Aveiro) sono legate all'aggiunta di uno spettrogramma nella visualizzazione (per migliorare le condizioni di segmentazione) e alla definizione di nuove

to totalmente revisionate e la procedura d'analisi e stilizzazione è stata ridefinita come successione di quattro *script* (*vox*, *fox*, *dox* e *box*) raggruppati nella 'sezione d'analisi' *Amper-fox* presentata ai *partner* del progetto nel 2003 (in occasione del *workshop* di Santiago d. C., Spagna – Inst. Ramon Piñeiro). Tra i file prodotti da *Amper-fox* si trovano quelli di tipo *.txt* che contengono, secondo una tabulazione prestabilita, i valori finali delle curve stilizzate, e informazioni sul posizionamento temporale dei segmenti definiti[15].

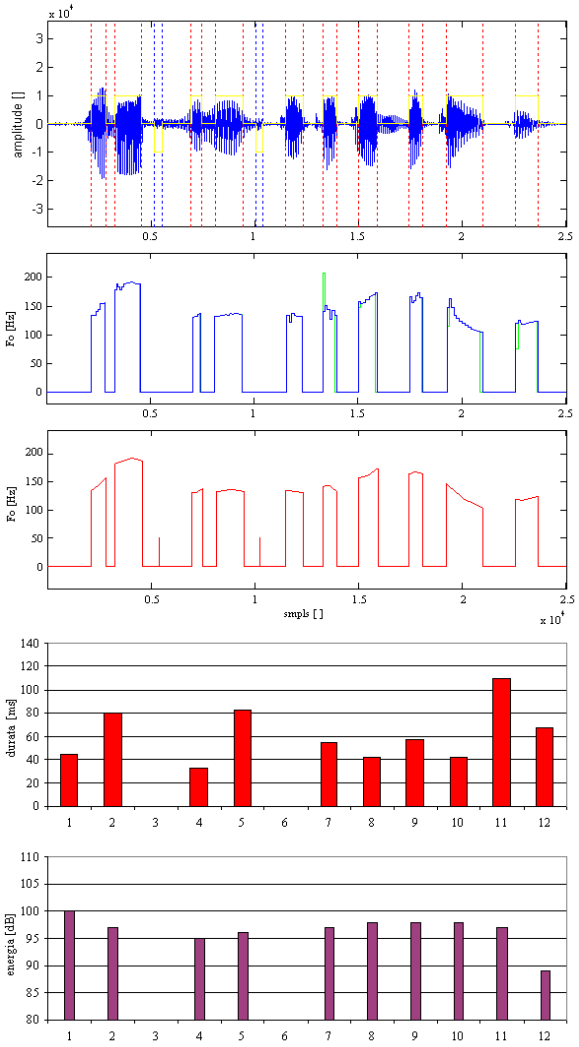


Figura 1: Esempio di analisi completa di un enunciato. Ripetizione numero 5 (codificata come *744twji5*) della frase interrogativa polare [*dU'nat@ s@ 'viv@t U ka'fE mbU'kat@?*] SAMPA 'Donato (si) beve il caffè riscaldato?' pronunciata da un locutore di Aliano (Matera, Basilicata, Italia mer., codice *744* poi *07f2*). Dall'alto in basso: grafico della forma d'onda usato per la segmentazione (con segnalazione delle posizioni in cui si è avuta cancellazione vocalica); curva di f_0 e, sovrapposta, curva risultante da correzione automatica; curva di f_0 stilizzata con annotazione convenzionale della posizione in cui è avvenuta la cancellazione; istogramma delle durate delle 12 vocali attese; istogramma dell'energia locale di ciascuna vocale [questi dati sono discussi nel complesso in Avolio & Romano 2010].

routine per la normalizzazione e per il calcolo della f_0 media (del parlante) a partire da tutti gli enunciati prodotti. Alcune di queste funzioni sono disponibili anche in *Amper-2006*, uno *script* più maneggevole sviluppato successivamente dall'*équipe* di Oviedo (Carmen Muñiz Cachón et coll.).

15 Tutte le *routine* sono disponibili all'indirizzo: http://www.lfsag.unito.it/amper/amperfox_amperdat.zip (alla pagina http://www.lfsag.unito.it/amper/fox.html; v. dopo).

Un esempio della successione di grafici prodotta da questi *script* è in Figura 1 insieme agli istogrammi delle sequenze di valori di durata ed energia locale desumibili dal contenuto del file *.txt* riportato in *tabella I*[16].

```
744twji5.txt      size: 25027
27-Oct-2007

            duration [ms]          energy [dB]  fo1   fo2   fo3 [Hz]

 1                  45                100          134   144   157
 2                  80                 97          183   193   188
 3                   0                  0           50    50    50
 4                  33                 95          131   132   138
 5                  83                 96          133   137   133
 6                   0                  0           50    50    50
 7                  54                 97          136   133   131
 8                  42                 98          142   144   133
 9                  57                 98          157   162   174
10                  42                 98          163   168   165
11                 110                 97          147   119   104
12                  68                 89          119   118   124

values at: 2117 2474 2831 3259 4085 4544 5353 5353 5353 6923 7185 7446 8112 8839
9445 10230 10230 10230 11491 11919 12347 13299 13632 13965 15011 15463 15915
17438 17771 18104 19246 20126 21007 22577 22987 23671
```

Tabella I

Insieme ad *AMPER-fox* è disponibile una 'sezione di trattamento e visualizzazione' *Amper-dat* costituita da versioni aggiornate di diversi *script* finalizzati al post-trattamento dei risultati delle analisi (e distribuiti in una nuova veste a tutti i *partner* in occasione del II incontro di Grenoble 2004, v. *AMPER* 2005).

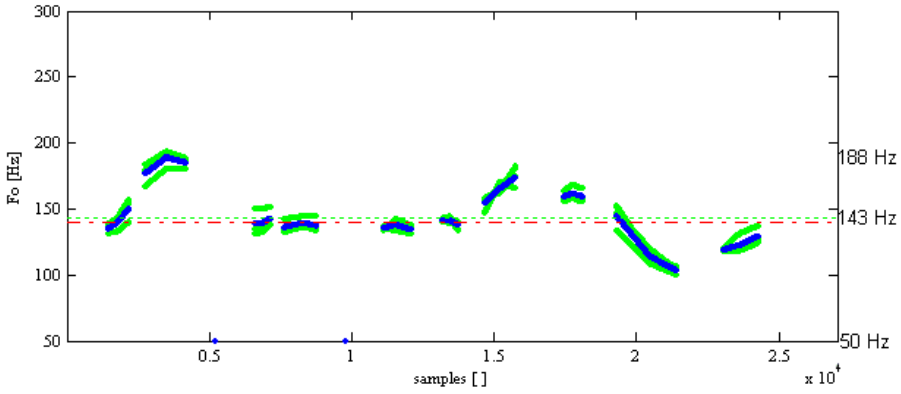


Figura 2: Esempio di curva stilizzata media (prototipo, in blu/gr. scuro) derivante dal confronto tra le diverse ripetizioni di una stessa frase (in verde/gr. chiaro, in questo caso dalle ripetizioni 3, 5 e 6 della struttura *twji*) [*dU'nat@ s@ 'viv@t U ka'fE mbU'kat@?*]$_{SAMPA}$ 'Donato (si) beve il caffè riscaldato?' (locutore di Aliano, Matera, v. Figura 1).

Tra gli *script* di *Amper-dat* si trovano quelli della serie nota come *avrgx,* che permette la determinazione di un modello 'medio' a partire da più enunciati dello stesso tipo (1ª

[16] Sin dalle prime analisi multiparametriche disposte in Romano (2001a&b), gli istogrammi si ottengono riportando i dati dei file *.txt* in un foglio elettronico. Dal 2004, i dati delle diverse frasi si possono 'incollare' in una serie di fogli (*Template*) di un file *.xls* predisposto dall'*équipe* di Aveiro in modo tale da generare automaticamente tutti i grafici necessari per le analisi quantitative/statistiche.

prototipizzazione, v. Figura 2), lo *script gentxt*, utile ai fini della generazione delle versioni sintetiche delle singole ripetizioni o del risultato di questa prototipizzazione (in vista di test d'ascolto o di percezione), e quelli del tipo *hzgr*, *stgr*, *hzcomp* etc. che consentono l'ispezione visiva dei dati contenuti nel file *.txt* o la sovrapposizione di più profili allineati (v. Figura 3).

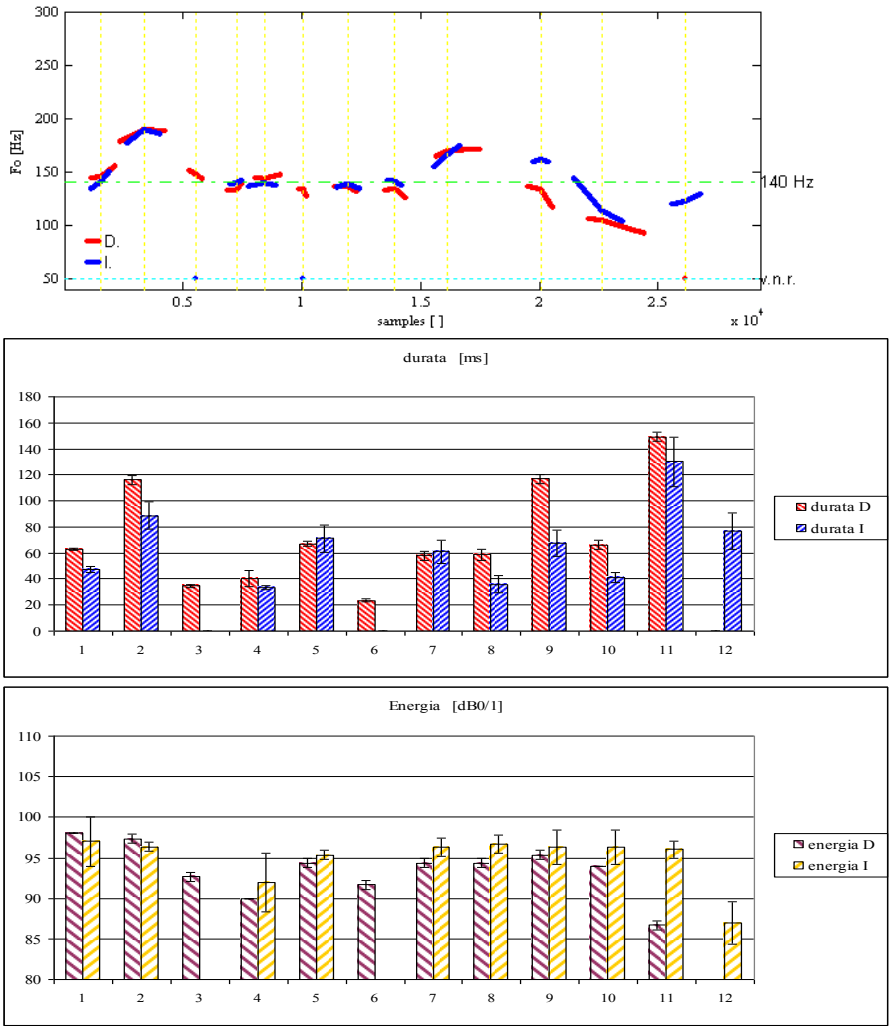


Figura 3: Esempio di confronto tra le sequenze di valori rilevati per le tre variabili f_0, durata ed energia in corrispondenza delle 12 vocali di una stessa frase in due modalità [*dU'nat@ s@ 'viv@t U ka'fE mbU'kat@./?*]$_{SAMPA}$ 'Donato (si) beve il caffè riscaldato./?' (locutore di Aliano, Matera, v. Figura 1). In alto: curve stilizzate medie (prototipiche) di f_0 (dichiarativa *twja*, in rosso/gr. chiaro, e interrogativa polare *twji*, in blu/gr. scuro; cfr. Avolio & Romano 2010). Al centro e in basso: istogrammi di durata ed energia (medie e deviazioni standard, *D* = dichiarativa e *I* = interrogativa polare; cfr. Avolio & Romano 2009)[17].

[17] L'utilità dei grafici risiede nelle indicazioni che danno riguardo alle sedi in cui le due modalità differiscono maggiormente. In questo caso, il confronto delle curve stilizzate di f_0, oltre a segnalare le diverse posizioni in cui avvengono le cancellazioni (ben visibili anche negli altri grafici), indica una sostanziale coincidenza dello sviluppo melodico delle due curve con variazioni più significative a partire dalla vocale nº 9 (notevole ad es. la maggiore altezza dell'ultima preaccentuale alla modalità interrogativa e il diverso andamento in corrispondenza degli ultimi tre tratti delle curve a confronto). Altre indicazioni vengono dagli istogrammi che permettono di segnalare le posizioni in cui le due modalità inducono distintamente ad allungamenti o accorciamenti vocalici (talvolta anche quantitativamente significativi) e/o a una maggiore o minore intensificazione locale.

Come illustrato dall'esemplificazione appena esposta, è implicita nelle applicazioni di questo metodo analitico la possibilità di riferirsi a strumenti alternativi esterni all'ambiente di programmazione ed esecuzione offerto da Matlab™ visto che quest'ultimo non è, in genere, disponibile tra le risorse di molti laboratori di fonetica.

Anche per questo motivo, alcuni *partner* hanno sollecitato a più riprese il trasferimento dell'intera procedura o, almeno, di parti di essa in ambienti di sviluppo e di fruizione più comuni. Sin dal 2005, alcuni *script* per *PRAAT*[18], parzialmente compatibili con le esigenze del progetto, sono stati implementati da P. Barbosa (Campinas, Brasile). Uno di questi, rinominato *AMPER_pour_PRAAT*, è stato successivamente adattato da A. Rilliard (Parigi) e – grazie anche allo sviluppo di un'interfaccia di trattamento dei dati – è stato diffuso, dal 2008, in una versione in grado di convertire una segmentazione-etichettatura eseguita con *PRAAT* in un file *.txt* formattato come quelli in uso nel progetto (v. *AMPER_PRAAT_Textgrid2Txt.psc*)[19].

3.2 Confronto intra-varietà e modellizzazione

Osservando l'insieme dei grafici (e ascoltando i prototipi risultanti dalle prime fasi della modellizzazione) è possibile individuare gli schemi con cui si manifestano alcuni vincoli strutturali locali (in genere riconducibili a esigenze di realizzazione degli schemi accentuali) e con cui si definisce l'intonazione globale di questi enunciati.

Confrontando ad es. la selezione di grafici della colonna di sinistra in Figura 4 (modalità dichiarativa) è possibile notare che lo sviluppo delle curve nel caso senza espansioni è complessivamente ascendente, fin quasi a tutta l'estensione della prima vocale accentata, per poi convertirsi in discendente (con una certa permanenza attorno alla frequenza media del locutore) fino all'ultima vocale accentata (nella sillaba *-fè*) in corrispondenza della quale si ha una discesa decisa (anche se non ripida) su valori bassi. In presenza di espansioni nel complemento, l'ultima vocale accentata del nome (della sillaba *-fè* vista sopra) si trova invece su valori piuttosto alti (e andamento ascendente) tanto più se nella parola seguente si presenta una prima sillaba non accentata: l'andamento basso-discendente osservato nel primo caso si localizza ora sulla vocale accentata (negoziando una prima parte di raccordo con l'ultima vocale accentata precedente in assenza di sillabe interaccentuali), mentre le postaccentuali finali si presentano cancellate o ridotte. Nella colonna di destra (modalità interrogativa) si osserva invece, in assenza di espansioni un profilo della curva con picco iniziale sulla prima vocale accentata e andamento appiattito (o lievemente ascendente) attorno alla frequenza media fino all'ultima accentuale (*-fè* che in questo caso particolare si trova in finale assoluta) e presenta un profilo nettamente discendente. In presenza di espansioni, la stessa sillaba presenta una vocale con valori piuttosto alti (a seconda della presenza di vocali interaccentuali), su un tratto ascendente della curva, che preludono alla discesa netta che si svolge in corrispondenza dell'ultima vocale accentata, la quale è ancora seguita da un ultimo tratto ascendente finale. Tra le conclusioni che si potrebbero trarre vi è quella legata all'assenza di quest'ultimo tratto finale nel caso di ossitono finale che fa pensare a una parlata in cui il contorno finale di

[18] *Public domain software* (Boersma & Weenink, Università di Amsterdam, versione *5.0*, 2007).

[19] L'interfaccia di manipolazione dei dati finalizzata alla produzione dei grafici (analitici e comparativi) e dei prototipi, tanto in formato testuale quanto in formato *_ton.wav*, è ora distribuita, in versione Beta7, come eseguibile *Interface_AMPER* e può essere usata su qualsiasi PC previa installazione del *Matlab Compiler Runtime* (*MCRInstaller.exe* che non necessita l'acquisto di nessuna licenza). Tutti gli *script* e gli eseguibili sono disponibili all'indirizzo: http://www.limsi.fr/Individu/rilliard/InterfaceAMPER.html.

modalità non viene compresso in caso di ossitono finale ma viene troncato (Avolio & Romano 2009).

L'insieme delle caratteristiche di questa parlata, limitatamente alle strutture presenti nel corpus, può essere infine racchiuso in rappresentazioni schematiche (con rapporti di altezza espressi in semitoni) come quelle proposte in Figura 5 (a & b)[20]. In queste, la possibilità che configurazioni locali, gravitanti attorno alle posizioni individuate dai tratti ispessiti, si spostino in funzione delle sedi accentuali è indicata dalle frecce (che in altri casi possono indicare lo scivolamento di queste in configurazioni globali più rigide), mentre le linee tratteggiate segnalano quei tratti di curva soggetti a troncamento.

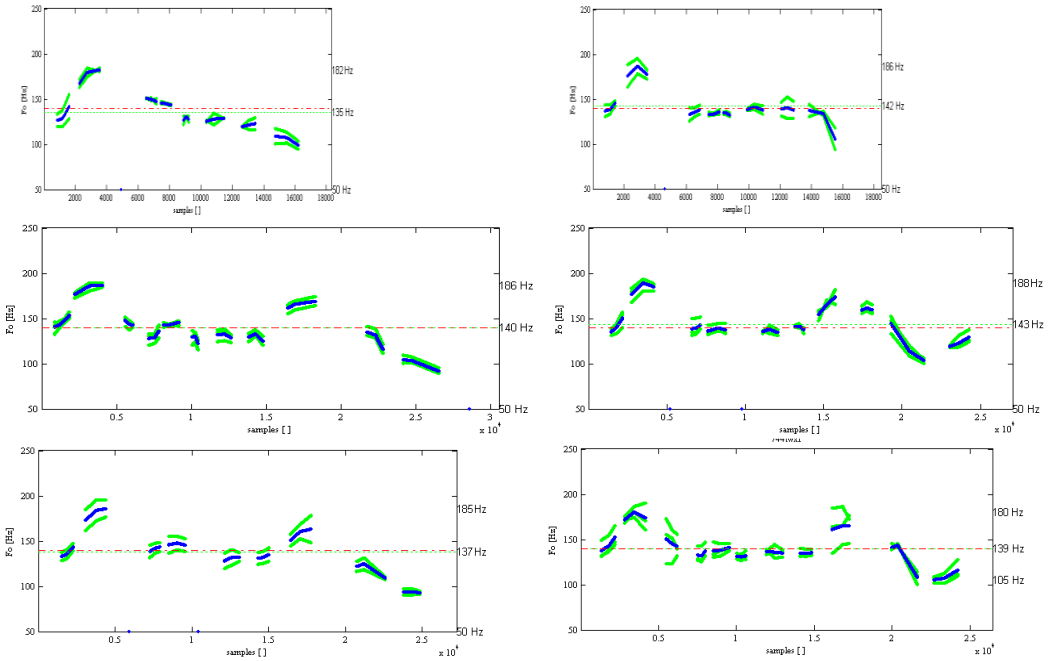


Figura 4. Esempio di confronto tra curve prototipiche per frasi con espansioni aggettivali nei sintagmi, nella modalità dichiarativa (colonna di sinistra) o interrogativa (colonna di destra). Dall'alto in basso: frase senza espansioni (*twk*, 'Donato (si) beve il caffè./?'); frase con espansione parossitona (trisillabica, *twj*, 'Donato (si) beve il caffè scaldato./?'), frase con espansione proparossitona (con riduzione da trisillabica a bisillabica, *twx*, 'Donato (si) beve il caffè torbido./?' (locutore di Aliano, Matera, v. Figura 1).

Ovviamente, prima di procedere a una modellizzazione di questo tipo è opportuno aver manipolato sperimentalmente tutti i dati presenti negli enunciati analizzati e rappresentati in termini oggettivi (fonetici). Pur presentandosi piuttosto soggettiva, possiede dunque adeguate dimensioni di astrattezza in grado di far emergere proprietà strutturali di alto livello partendo da valutazioni quantitative su materiali costruiti per annullare (a turno) alcune sorgenti di variazione. Essa non è l'unica proposta nell'ambito del progetto che poggia su esperienze condotte sin dagli anni '70-'80 sulla base di altri modelli (v., tra gli altri, Contini & Profili 1989): è attualmente allo studio da parte di M. Contini & A. Rilliard una modellizzazione ottenuta mediante classificazione automatica di tratti locali e globali e attribuzione di ponderazioni (motivate da variabili linguistiche) a varie porzioni delle sequenze di valori estratte.

[20] Omettiamo, anche per ragioni di spazio, la discussione dei dati di *D* ed *E* (cfr. Avolio & Romano 2010).

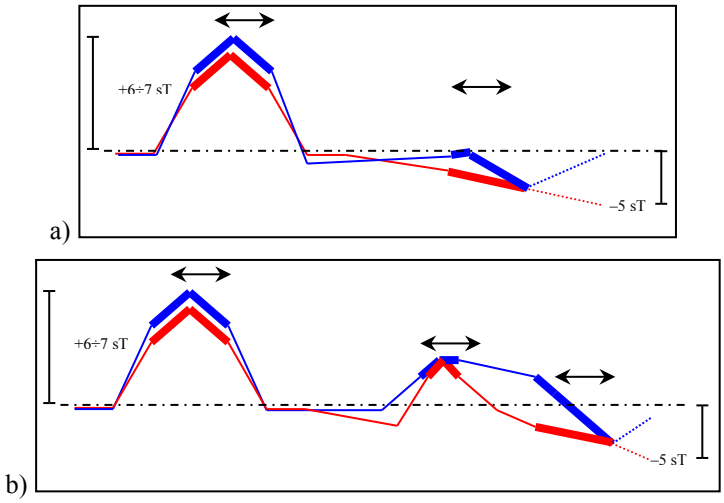


Figura 5. Confronto tra gli schemi intonativi prototipici delle due modalità (*D* - dichiarativa, rosso/gr. più chiaro, e *I* - interrogativa, blu/gr. più scuro) nelle produzioni del locutore di Aliano (v. Figura 1): a) nelle frasi senza espansione; b) nelle frasi con un'espansione nel sintagma verbale (v. testo).

3.3 Confronto inter-varietà

L'insieme delle caratteristiche così messo in evidenza può anche essere sfruttato per confronti esterni (inter-dialettali), a condizioni di 'normalizzare' le altezze medie e l'estensione degli intervalli di variazione dei prototipi. Per darne un'esemplificazione di massima confrontiamo in Figura 6 gli schemi prototipici di Aliano (di Figura 5) con quelli ottenuti su dati del corpus raccolto e analizzato per la vicina località di Alianello (cfr. Avolio & Romano 2010). Se le differenze che si localizzano in corrispondenza della prima sede accentuale possono essere ritenute secondarie e gli schemi essere giudicati simili nel complesso, appaiono evidenti due o tre fonti di dissimilarità: innanzitutto il diverso profilo locale sulla penultima vocale accentata e la flessione presente tra questa e la preaccentuale seguente nel dialetto di Alianello (difficile stabilire se questo tratto sia saliente in termini percettivi, in assenza di test mirati); secondariamente potrebbe essere rilevante la diversa pendenza/forma dell'andamento discendente sull'ultima vocale accentata; soprattutto però, infine, sembra caratteristico il mancato troncamento finale delle interrogative presenti nei dati di Alianello, il quale mostra – al contrario – una tendenza a innescare un'ulteriore discesa dopo l'andamento ascendente sulla postaccentuale, quindi a dilatarsi su un elemento sonoro di tipo paragogico.

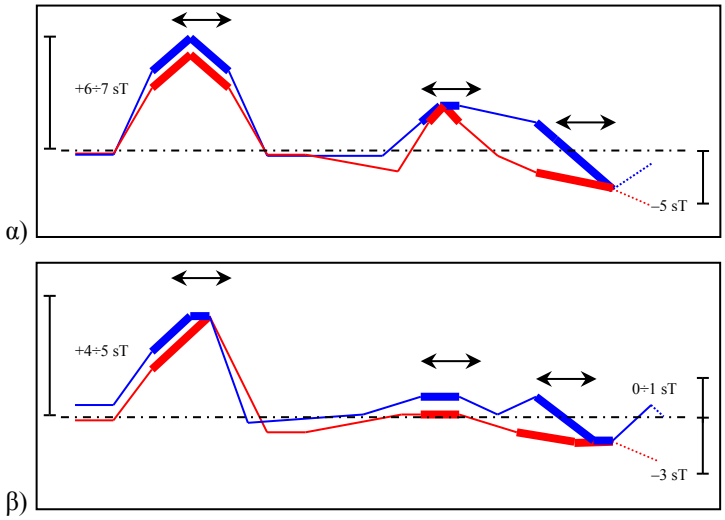


Figura 6. Schemi intonativi prototipici delle modalità *D* (dichiarativa, rosso/gr. più chiaro) e *I* (interrogativa, blu/gr. più scuro) nelle frasi con un'espansione nel sintagma verbale nel dialetto di Aliano (α) a confronto con schemi equivalenti nel dialetto di Alianello (β) (cfr. Figg. precc.; v. testo).

4 Applicazioni allo studio della variazione prosodica

Come già segnalato al §1, il lavoro che i diversi partner del progetto conducono su dati di questo tipo presenta in genere finalità dipendenti dall'orientamento degli interessi scientifici dei collaboratori che compongono l'équipe di ricerca. Oltre a una frequente inclinazione a studiare aspetti dialettologici e sociolinguistici di determinate micro-aree (o a contribuire all'allestimento di dati relativi ad aree più estese) vi sono ricercatori interessati a discutere le modalità di connotazione che i dati raccolti assumerebbero su altri piani. Sebbene si tratti, infatti, di dati elicitati, sembra rilevante valutarne la caratterizzazione pragmalinguistica o gli aspetti di organizzazione informativa che presuppongono o che inducono nell'ascoltatore chiamato a manifestare un giudizio sul valore e sull'effetto che enunciati con quelle determinate caratteristiche intonative assumono nella parlata in questione[21].

Nel caso più generale, gli andamenti dei singoli enunciati possono presentarsi variabili su questi piani e i prototipi intonativi individuati necessitare valutazioni percettive preliminari che, oltre a contribuire alla comprensione delle regole di strutturazione prosodica della parlata studiata, permettano di classificare i dati raccolti in termini di variabilità individuale, stilistica, situazionale etc.[22]

Nonostante i dati parziali per certi domini permettano già di procedere a valutazioni tipologiche e/o dialettometriche potenzialmente convincenti[23], l'applicazione del metodo

[21] Varietà simili all'interno di una stessa area o comunità possono differire per la frequenza con cui alcuni parlanti ricorrono a forme di tematizzazione/topicalizzazione/dislocazione e/o in base alle modalità specifiche con cui realizzano queste possibilità (cfr. Romano & Mattana 2008).

[22] Le premesse sul grado di esplicitazione raggiunto nella comunità dei parlanti dal livello di codificazione linguistica della prosodia della singola parlata o anche solo di alcuni *cliché* intonativi di cui si abbia un controllo o, per lo meno, una consapevolezza metalinguistica, necessitano di un riferimento a un quadro variazionale quanto più possibile completo, come quello disponibile oggi grazie al contributo di diversi autori (sin da Coşeriu 1958).

[23] Si veda l'intervento sintetico di Eugenio Martínez Celdrán sulle varietà intonative dello spagnolo in Spagna presentato alle IV giornate internazionali del progetto *AMPER* (Siviglia, Spagna, 18-19/02/2010; v. *Sevilla* 2010). Sebbene limitato a uno studio pilota, mi permetto di segnalare anche il mio primo contributo esplorativo sull'intero spazio romanzo europeo in Romano (2004).

AMPER si "limita" oggi a studi di variazione geoprosodica applicati in genere localmente su scala micro-regionale (o regionale) oppure a studi sui confronti tra dialetti e varietà regionali o ancora sulla variazione sociale e stilistica della lingua nazionale[24].

Come esempio di contributo allo studio della variazione micro-regionale, mi permetto di rimandare ai miei lavori sul Salento, nei quali ho analizzato la diffusione di due modelli prosodici differenti, discutendone diversi aspetti strutturali e mettendo in evidenza, in particolar modo, la diversa realizzazione della modalità interrogativa (Romano 1997, 2001a&b, 2003).

4.1 La variazione micro-regionale

Partendo dall'osservazione di diverse centinaia di enunciati (e dalle rappresentazioni grafiche raccolte in un intero volume; v. dati in AMPER 2011) è stato possibile definire i due modelli prototipici di due dialetti (il primo, 061 = Parabita, appartenente a un'area solitamente ritenuta più conservativa, il secondo, 062 = Sannicola – a poco più di 7 km dal precedente –, orientato verso i modelli intonativi dei centri urbani più popolati del Salento centrale, v. Figura 8). Come discusso in Romano (2001b), i due modelli valgono, nel caso di parlanti particolarmente spontanei, anche per l'italiano regionale (IR) tipico dei due centri (anche questi documentati in AMPER 2011)[25].

Oltre alle differenze nei contorni terminali di modalità tra le due varietà intonative, gli schemi permettono di osservare i rapporti tra dichiarativa e interrogativa, anche in quelle porzioni più soggette a variazioni di tonalità (pragmalinguistica e/o affettiva) o più in generale alla presenza di tratti parafonici (cfr. Canepari 2004: 245-249).

Notare infine che questi modelli prototipici (come in quelli di Figura 6) possono adattarsi alle forme reali che si presentano al variare della posizione delle sedi accentuali (cui sono associati i tratti grafici di maggiore spessore): il senso delle frecce indica lo spostamento che subiscono questi tratti di curva e lasciano immaginare la riorganizzazione locale che avviene in conseguenza di ciò[26]. La significatività statistica di differenze localiz-

[24] Tra i primi a presentare un allestimento di dati atlantistici nell'ambito di *AMPER*, sono gli studi curati da Josefa Dorta Luís e coll. per *AMPER-CAN* (Universidad de la Laguna, Tenerife - Isole Canarie, Spagna), da Ana Maria Fernandez Planas & Eugenio Martínez Celdrán per *AMPER-CAT* (Universitat de Barcelona, Spagna) e da Elisa Fernandez Rei e coll. per *AMPER-GAL* (Universidad de Santiago de Compostela, Spagna). Una menzione a parte merita l'équipe di Carmen Muñiz Cachón (Universidad de Oviedo, Spagna) per le ricerche condotte nell'ambito di *AMPER-AST* in una molteplice dimensione di studio, dalla variazione geoprosodica dell'asturiano (e delle sue modalità di differenziazione in modelli urbani e rurali), a quella del castigliano delle Asturie. Veri e propri pionieri in diverse fasi del progetto sono stati Lurdes de Castro Moutinho di *AMPER-POR* (Universitade de Aveiro, Portogallo) e Adrian Turculeţ *AMPER-ROM* (Universităţea "Al. Ioan Cuza", Iaşi, Romania), la prima per la creazione di una rete uniforme in grado di perseguire la descrizione della variabilità del portoghese in due dei suoi spazi dialettali più consistenti (con contributi all'analisi strumentale, alla rappresentazione dei dati e all'analisi tipologica) e il secondo per la problematizzazione delle dimensioni di variazioni del romeno anche al di fuori dello spazio dacoromanzo.

[25] Per questa caratteristica si definisce la possibilità di considerare almeno due grandi aree intonative salentine (parzialmente sovrapposte e – date le attuali condizioni di contaminazione diasistemica e d'interferenza tra fasce di popolazione diverse – confuse nell'imprevedibilità delle scelte e nella complessità delle esperienze individuali).

[26] Diversamente da ciò che accade nel caso di questi dialetti, i profili si presentano talvolta poco variabili al variare della posizione degli accenti, come se i movimenti locali indotti dalla presenza di una prominenza si adattassero a una curva globalmente più 'rigida'. Questo può produrre profili melodici diversi all'approssimarsi (o all'allontanarsi) della sede accentuale a posizioni prosodicamente più forti o a confini intonativi: 'scivolamenti' dei profili melodici locali sono registrati ad es. per il primo accento in protonia interrogativa nei dati di friulano in Romano & Miotti (2008: 241) oppure nel caso dell'accento nucleare, nei dati che descrivono l'intonazione interrogativa dello spagnolo di Málaga, o di quello immediatamente

zate in posizioni interaccentuali (come quella dell'avvallamento tra i primi due accenti in protonia in alto a sinistra in Figura 8) e associate alla percezione della differenza tra una modalità e l'altra è discussa in Romano (2005) a conferma di dubbi sollevati in precedenza sulla validità di certe trascrizioni (cfr. Romano 2003).

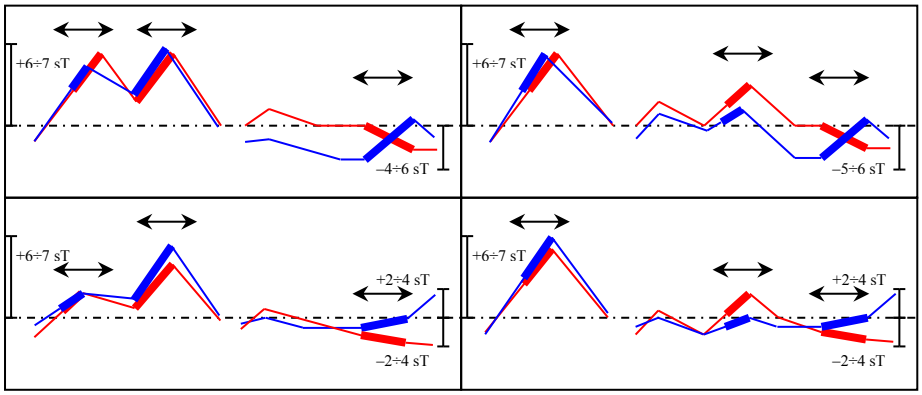


Figura 8. Confronto tra schemi intonativi prototipici delle due modalità *D* (dichiarativa, rosso/gr. più chiaro) e *I* (interrogativa, blu/gr. più scuro) in due varietà di salentino (in alto 061 = Parabita e in basso 062 = Sannicola). Modelli relativi a frasi con un'espansione: nel primo sintagma nominale (a sinistra) oppure nel sintagma verbale (a destra).

4.2 La variazione regionale

Per esemplificare la variazione intonativa dell'*IR*, in Romano (2005), ad esempio, sono discusse – seppure con dati provvisori e incompleti – alcune proprietà tipiche delle varietà di Roma, Torino e Nuoro sulla base della resa dei profili di frasi corrispondenti[27]. La presenza di elementi caratterizzanti delle tre varietà considerate nelle diverse sedi è illustrata dai grafici di Figura 7. Si possono notare in particolare le differenze presenti nella tipica intonazione che contraddistingue i contorni terminali della modalità interrogativa: mentre nell'*IR* romanesco si presenta una preaccentuale con profilo ascendente medio-basso seguita da un picco sulla vocale accentata e un andamento postaccentuale discendente medio, i dati dell'*IR* torinese (per un'approfondita discussione dei quali si rimanda a Interlandi 2004) presentano invece un andamento discendente medio-alto sulla preaccentuale, minimo spezzato in corrispondenza della sede accentuale e un profilo nettamente discendente, da medio a basso, sulla finale. In modo percettivamente ben differenziabile, il profilo tipico dell'*IR* sardo di Nuoro, presenta una preaccentuale alta, una discesa su valori medio bassi sulla vocale accentata e una risalita finale sulla postaccentuale con valori in un intervallo più che altro medio (cfr. Lai 2002).

prenucleare alla modalità dichiarativa nella varietà veneta di Motta di Livenza (cfr. Romano & Miotti 2009: 67 e 70).

27 Anche questi lavori – così come quelli pionieristici di M. Contini sull'*IR* (v. tra gli altri, Contini & Profili 1989), che pure hanno avuto una certa diffusione internazionale e hanno riscosso un certo successo per alcune proposte originali che contengono – sono curiosamente ignorati da molti sedicenti specialisti italiani d'intonazione (compresi alcuni di quelli che hanno proposto in anni recenti una descrizione della variazione intonativa dell'*IR* - una significativa eccezione è offerta da Sorianello 2006).

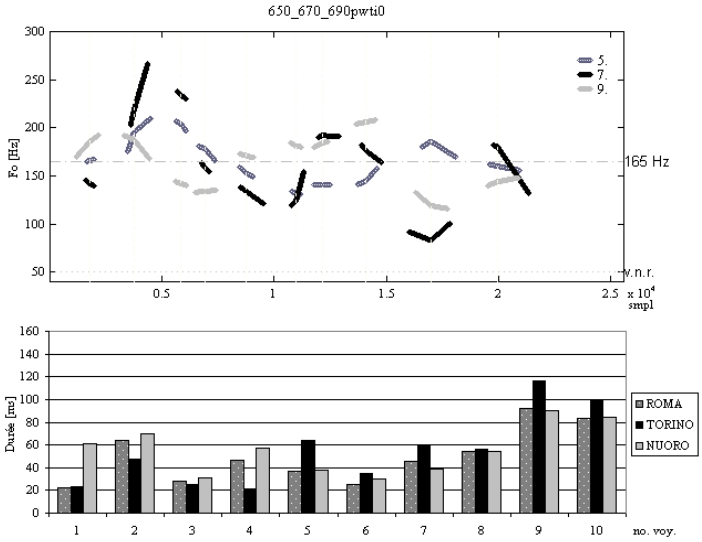


Figura 7. Confronto tra profili melodici e istogrammi di durata prototipizzati per una stessa struttura (*pwti* corrispondente a "La papera mangia la patata") nelle varietà d'italiano regionale di Roma (5.= (0)650), Torino (7.= (0)670) e Nuoro (9.= (0)690).

4.3 La variazione stilistica

Sebbene sia piuttosto intuitivo quanto possa essere rilevante la connotazione stilistica delle produzioni osservate ai fini della caratterizzazione prosodica delle lingue (per diversi spazi dialettali e contesti sociolinguistici), nell'ambito di *AMPER* questo tipo di variazione è stato indagato in particolare da A. Turculeţ e colleghi (cfr., tra gli altri, Turculeţ *et alii* 2003) che hanno messo in evidenza la coesistenza di soluzioni intonative diverse nella stessa varietà e per gli stessi parlanti.

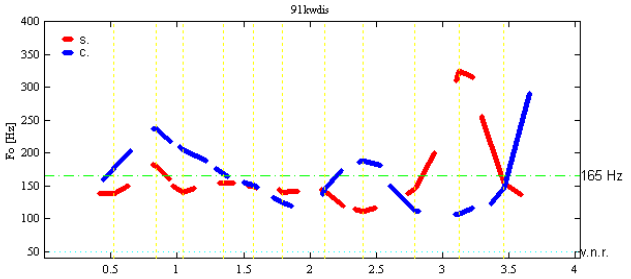


Figura 8. Confronto tra profili melodici per una stessa struttura (*Pasărea vede fantoma galbenă?* 'L'uccello vede il fantasma giallo') nelle produzioni di uno stesso parlante di Iaşi in uno stile controllato (c. = pronuncia 'letteraria' del romeno) e uno stile più spontaneo (s. = pronuncia con intonazione regionale).

In Figura 8 (tratta da Turculeţ *et alii* 2005) sono proposti ad es., sovrapposti, i due profili caratteristici relativi a una stessa frase pronunciata da una parlante di Iaşi ricorrendo alle soluzioni prosodiche che è in grado di controllare. I due tipi di enunciazione si presentano nettamente differenziati in più punti. Oltre a una tonalità più alta nella parte protonica si distinguono per alcuni movimenti che si svolgono in modo contrario a partire dall'ultimo accento pretonico: mentre nell'intonazione tipica di una pronuncia più formale del romeno (qui definita controllata) l'andamento melodico sulla vocale nucleare si presenta discendente (con massimo allineato sulla sillaba precedente e minimo sulla

sillaba seguente) per inerpicarsi su un'estesa risalita finale (su quasi due ottave!)[28], nell'intonazione tipica dell'enunciazione regionalmente marcata (spontanea) di quest'area (come anche in quella di regioni vicine; v. Turculeţ 2008 e bibliografia citata; cfr. Romano 2006) prevede un profilo discendente nella resa dell'accento pretonico, una ripida risalita su quello tonico/nucleare (con minimo sulla sillaba precedente e massimo sulla seguente) e contorno finale discendente.

Oltre a suggerire, quindi, di definire con esattezza il codice che si sta elicitando, nella fase di raccolta dei dati, questi esempi basterebbero per far riflettere sulle dimensioni di variazione da esplorare (e sulle modalità di descrizione da adottare).

4.4 La verifica percettiva tra dialettologia e sociolinguistica

Come già anticipato in §1, nell'ambito della percezione della prosodia il progetto *AMPER* ha sviluppato una metodologia originale che non ha mancato di attirare l'attenzione di numerosi specialisti[29].

Un apporto sostanzioso è venuto dalla proposta *ante litteram* di modalità di verifica che, per quanto inizialmente piuttosto 'ingenue' (cfr. Romano 1997), sono venute poi raffinandosi nell'ambito di studi successivi (v., tra gli altri, Interlandi & Romano 2004), basati su modalità di somministrazione collaudate e su tecniche di manipolazione degli stimoli sintetici sempre più sofisticate[30].

Tra le applicazioni che hanno riscosso un notevole interesse anche in contesti internazionali, vi sono quelle relative alle variabili socioprosodiche osservate nei dati di Parma[31].

5 Verso una cartografia delle variabili prosodiche: bilanci e previsioni

Gli obiettivi di rappresentazione cartografica interattiva di *AMPER* sono posti sin dal suo lancio, ma si trovano ancora allo stato attuale di fronte al problema di una copertura geografica incompleta e disuniforme (anche se localmente sono stati progressivamente alle-

[28] Inutile sottolineare che, indipendentemente dal diverso impegno articolatorio che questo tipo di parlato presuppone (senza essere necessarimente più scandito), a queste caratteristiche se ne associano generalmente altre, temporali ed energetiche, che qui non menzioniamo (v. Turculeţ *et alii* 2005).

[29] Anche se sviluppato in tempi e modi diversi, il metodo poggia su basi simili a quelle proposte da 't Hart *et alii* (1990) e condivide molti punti programmatici con i lavori di Ch. Gooskens (1997), di P. Mertens (2004) e, in parte, col sistema *Momel* (v. Hirst *et alii* 2004) i quali si prefiggono tuttavia obiettivi di etichettatura e di modellizzazione automatica. La nostra proposta di una verifica percettiva delle possibilità di discriminazione di varietà dialettali ha fornito un modello allo studio di Peters *et alii* (2003) sulla caratterizzazione prosodica di varietà del tedesco di Germania. Grazie alla mediazione dei suoi partner, le procedure *AMPER* sono oggi osservate con interesse da gruppi di ricerca anche esterni al progetto che, soprattutto in Francia, Spagna e Brasile, perseguono una modellizzazione funzionale della prosodia e una valutazione percettiva delle variabili, degli indici discriminatori più salienti e delle possibilità di variazione più comuni nello spazio romanzo. Anche riguardo allo studio dell'organizzazione temporale, la notevole versatilità dei dati formattati secondo le specifiche *AMPER* ha fornito lo spunto a diversi autori (v. ad es. alcuni contributi apparsi in *RILI* 2011) per condurre alcune misurazioni in base alle metriche ritmiche più diffuse (come suggerito sin da Mairano & Romano 2008). L'uso degli stimoli sintetici di tipo *.ton* è servito inoltre in alcuni studi sulla caratterizzazione ritmica di lingue e dialetti grazie alla facilità con cui è possibile manipolare le variabili interessate (v., tra gli altri, Romano/ Mairano 2010; Mairano/Romano 2012).

[30] Sebbene molti partner abbiano spinto le loro ricerche in questa direzione ciascuno a suo modo (e spesso con modalità ancora più ingenue), la metodologia più affidabile risiede in un'attenta valutazione preliminare dei file *.ton* usati per questi stimoli (e nella loro manipolazione, ora facilitata dallo *script Stim_Creator_V2.m* realizzato da D. Avesani).

[31] In particolare negli studi sociofonetici di Felloni & Avesani (2010) e Felloni (2011).

stiti alcuni siti per la consultazione *on-line* dei dati – pensiamo allo spazio balcano-romanzo, alle Canarie, alla Catalogna o al Portogallo, v. §3, n. 23).

Trattandosi della rappresentazione di dati ben diversi da quelli lessicali della dialettologia tradizionale, l'idea di fornire solo dati sonori o grafici per i punti esplorati conduce al modello degli "Atlanti parlanti" (per i quali il Centre de Dialectologie de Grenoble ha fornito in un recente passato interessanti applicazione microareali) e sembra ancora troppo semplicistica e valida tutt'al più in una prospettiva divulgativa (v. Romano 2004). Benché alcuni lavori pionieristici nell'ambito della prosodia (v. Bruce & Gårding 1978 e fonti citate) abbiano suggerito alcune linee esemplari per una rappresentazione cartografica di alcune di queste variabili (principalmente configurazioni accentuali), manca ancora una riflessione concordata sui criteri di selezione, classificazione e visualizzazione degli eventi caratterizzanti della prosodia di frase. Gli sforzi di diversi partner sono attualmente orientati verso una tipologizzazione convincente degli schemi relativi alle due modalità maggiormente presenti nella *BD* sulla base delle interessanti valutazioni offerte da Vaissière & Boula de Mareüil (2004)[32].

In queste condizioni d'incertezza si è preferito offrire una prima rappresentazione, parziale e dimostrativa, dello spazio di variabilità finora indagato basandosi sui soli dati sonori prototipici, non filtrati da alcuno schema classificatorio, ma affiancati interattivamente dalla visualizzazione delle sequenze stilizzate dei valori assunti dalle principali variabili prosodiche (quelle indicate sin dall'inizio: f_o, D e E) che ne descrivono l'andamento.

Il 'colpo d'occhio' sulle possibilità offerte da questo modello di consultazione dei dati è offerto dal DVD di *AMPER* (2011) realizzato da P. Mairano (v. Figura 9) nel quale si trovano i risultati delle ricerche condotte da diversi partner[33].

Per ogni punto considerato è possibile ad es. interrogare la *BD* interna secondo le frasi presenti nel corpus di dati raccolto in un dato punto (v. Figura 10)[34].

[32] Al di là delle distinzioni tra le forme dei profili più tipici descritti per ciascuna varietà osservata in determinate posizioni e per le due distinte modalità (come ad es. il contorno terminale, che i colleghi spagnoli hanno da tempo cominciato a tipologizzare come "tonema"), sembrerebbe rilevante poter includere tratti descrittivi relativi all'organizzazione temporale (varietà con riduzione e/o cancellazione vocalica vs. varietà con maggior controllo delle durate vocaliche) e ai fenomeni di allungamento finale e adattamento dei contorni alle variazioni di posizione dell'ultima sede accentuale (varietà con espansione vs. compressione, paragoge vs. troncamento). Il problema delle scelte di rappresentazione simbolica (e visualizzazione) è, infine, reso più delicato dal rischio di cartografare soluzioni insufficienti a discriminare persino indici di differenziazione tra i più stereotipati (come è accaduto in alcuni tentativi recenti condotti per l'*IR* sulla base di trascrizioni di tipo *ToBI*).

[33] La diffusione dei metodi (e degli strumenti) proposti nell'ambito di *AMPER* ha coinvolto fino a oggi più di 40 équipe di una decina di Paesi diversi. In vista di questa prima pubblicazione, è stata completata la revisione dei dati relativi a 108 locutori di 62 punti d'inchiesta (notizie dettagliate sul contenuto del DVD *AMPER* 2011 sono disponibili all'indirizzo http://w3.u-grenoble3.fr/dialecto/AMPER/DVD). Il DVD propone appena un terzo dei più di 30.000 file sonori 'grezzi' (da sottoporre ancora a verifiche di formato e a valutazioni qualitative generali) che sono attualmenti presenti nella *BD-AMPER* (v. dopo). I partner che hanno elaborato i dati del DVD sono indicati punto per punto (la lista delle inchieste si trova all'indirizzo http://w3.u-grenoble3.fr/dialecto/AMPER/DVD/consultation/liste_enquetes.html). L'elenco completo dei partner (storici) che hanno conferito dati alla *BD*-AMPER è invece consultabile all'indirizzo http://amper.limsi.fr

[34] Per le altre possibilità si v. http://w3.u-grenoble3.fr/dialecto/AMPER/DVD/consultation/guide.html.

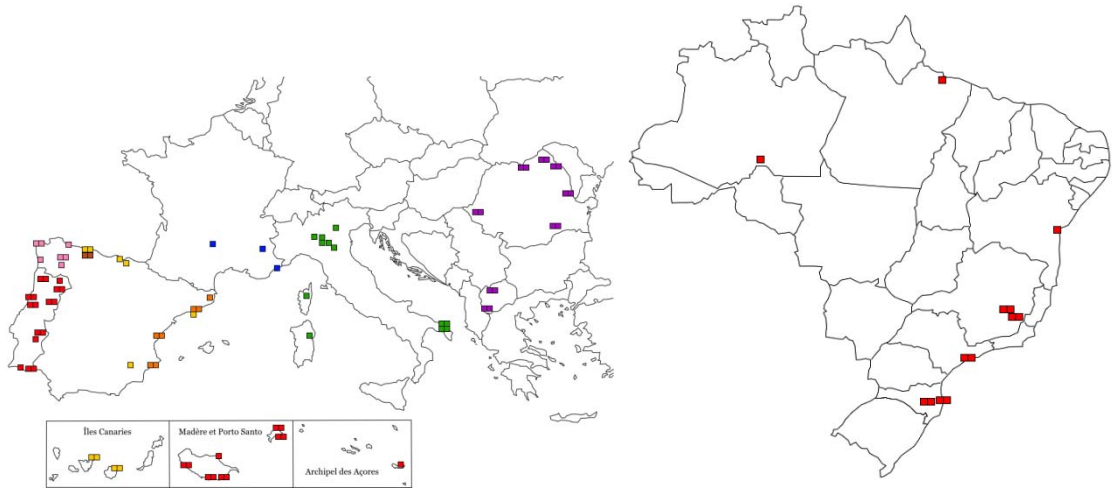


Figura 9. Copertura geografica offerta dai dati presenti nel DVD *AMPER* (2011) nei due spazi geografici europeo e sud-americano.

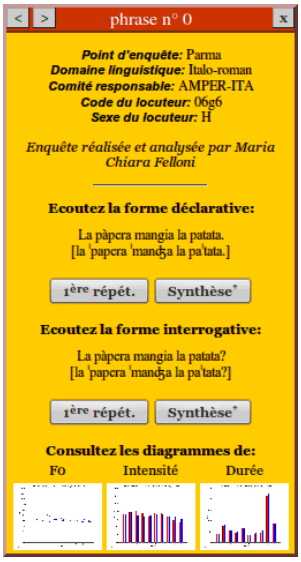


Figura 10. Esempio di finestra grafica attivata nella consultazione dei dati di un punto di rilevamento relativi a una struttura preselezionata del *QCB* (immagine modificata da Contini et al., in c. di p.).

Il DVD rappresenta la sezione della Base di Dati complessiva (*BD-AMPER*) che è stato finora possibile verificare e uniformare in vista di una consultazione che assicuri anche la comparabilità tra le rese delle stesse strutture in due punti diversi (con limitazioni che restano ancora da segnalare adeguatamente). Si tratta di un traguardo provvisorio che illustra le caratteristiche di un 'cantiere aperto' il quale non ha mancato tuttavia di conseguire già un certo numero di risultati significativi[35].

[35] Lo stato di avanzamento dei lavori è stato al centro di quattro convegni internazionali e diverse sessioni speciali organizzate nel corso di diversi congressi di fonetica. Oltre che in numerose tesi di Dottorato (e di diploma), l'applicazione del metodo *AMPER* caratterizza centinaia di articoli in riviste e atti di congressi internazionali (un elenco aggiornato al 2012 è al link http://w3.u-grenoble3.fr/dialecto/AMPER). Oltre al già citato n. 4 HS di *Géolinguistique* (*AMPER* 2011), tra le pubblicazioni più rilevanti risaltano inoltre in particolare i seguenti volumi monografici: il n. 3 HS di *Géolinguistique* (*AMPER* 2005, v. anche *Grenoble* 2004); il vol. XIV di *Estudios de Fonética Experimental* (dedicato ad *AMPER-en-España*, v. *EFE XIV*

6 Misure oggettive di somiglianza o differenziazione prosodica

Una possibilità di sviluppo originale attualmente ancora poco frequentata dai diversi partner è nella definizione di misure di distanza obiettive tra le caratteristiche prosodiche presenti nei formati convenzionali dei dati raccolti per *AMPER*.

Allo stesso tempo in cui apparivano le pubblicazioni di D. Hermes su questo tema (v., tra gli altri, Hermes 1998), l'interesse di questi studi era stato segnalato nel corso della preparazione del Dottorato di Ricerca di A. Romano[36] e presentato a più riprese in vari contributi (v., tra gli altri, Lai & Rilliard 2008).

L'idea è, innanzitutto, quella di sottoporre i dati delle singole inchieste a verifiche di *coerenza* intra-locutore e a misure di *congruenza* inter-locutore, in modo da evidenziare caratteristiche generali di affidabilità dei dati in vista di confronti inter-varietà, basati su misure di *correlazione* (Romano 2001a) o di *distanza* (Romano & Miotti 2008; Romano *et alii* 2011)[37]. Queste indagini, affidate per ora a strumenti provvisori e a procedure in via di definizione e di test (v. Moutinho *et alii* e Fernández Planas *et alii* in *EFE XX*, 2011), presuppongono valutazioni di soglie di riferimento come quelle introdotte per i confronti discussi in Romano (2001a)[38].

In Figura 11 riportiamo le stime di coerenza e congruenza per una selezione d'inchieste di *AMPER-ITA* relative a dati salentini (0616-0621-0625) ed emiliani (06g5-06g6-06h7). A titolo d'esempio discutiamo solo dei primi (già oggetto di esemplificazione al §3.1). Le stime si basano su un numero diverso di enunciati (con una media di 5 ripetizioni per ciascuna di 16 frasi alle due modalità), rispettivamente 169, 162 e 161. La coerenza tra i dati dei tre sub-corpora (basata sulla dispersione dei valori di correlazione dell'insieme di enunciati relativi alla stessa struttura) varia in media tra 0,93 e 0,94 (v. grafico a sinistra di Figura 11) presentandosi quindi decisamente buona (rispetto ad es. a quella nei dati del locutore 06h7 attestata attorno a 0,81, con una varianza piuttosto elevata). Anche la congruenza tra i dati dello stesso punto, ad es. 062 (ottenuta dalla correlazione incrociata tra 0621 e 0625), si presenta molto alta (> 0,94)[39].

L'esempio di valutazione dialettometrica condotto su simili dati è invece offerto dalla misura delle distanze tra i dati del sub-corpus della *BD* di *AMPER-POR* (Romano *et alii*, in c. di p.). Premesso che nei dati portoghesi sono stati da tempo segnalati almeno due modelli diversi di realizzare le domande totali (Moutinho *et alii*, 2004), i dati si presentano piuttosto variati geograficamente, tanto sul piano degli andamenti di f_o, quanto su quelli della durata, legata a sequenze di valori fortemente divergenti per via dei fenomeni di cancellazione vocalica. Oltre a offrire una proposta di raggruppamento per i modelli intonativi del portoghese arbitrariamente scelti per questo studio-pilota e rappresentati

2005); *La prosodia en el ámbito lingüístico románico* (a cura di J. Dorta, 2007); gli *Atti di AMPER-POR* 2007 (a cura di L. de Castro Moutinho & R.L. Coimbra, v. *Aveiro* 2007); *La variation diatopique de l'intonation dans le domaine roumain et roman* (a cura di A. Turculeţ, 2008); il n. 1 (17) del vol. IX della *Revista Internacional de Lingüística Iberoamericana* (v. *RILI* 2011).

36 In Romano (2001a&b) le proposte di un confronto quantitativo tra dati prosodicamente differenziati beneficiano di una prima formalizzazione e di alcuni trattamenti preliminari.

37 Il procedimento ricorda in parte alcuni di quelli usati in dialettometria (Goebl 1983, 1996; Saramago 1986, Saramago & Bettencourt 2003). Per queste valutazioni si può ricorrere a formule per il calcolo della correlazione tra due sequenze X e Y di n valori relativi allo sviluppo temporale di alcune variabili ritenute prosodicamente rilevanti (e, in queste applicazioni, coincidenti con quelle definite nel §2).

38 La possibilità di riferirsi tanto a enunciati dialettali quanto a produzioni in *IR* basate su strutture comparabili dà ad es. la possibilità di studiare la correlazione tra le soluzioni intonative usate nei due codici (cfr. Romano 2001b).

39 Ciò contribuisce a classificare i dati raccolti in questo punto come poco variati tra loro e, quindi, rappresentativi di un modello di pronuncia sicuro e costante.

dai dati di sei punti d'inchiesta, la Figura 12 dà una dimostrazione dei benefici che si potrebbero trarre dall'applicazione di una simile tecnica ai dati della *BD* con interessanti indicazioni di lettura[40].

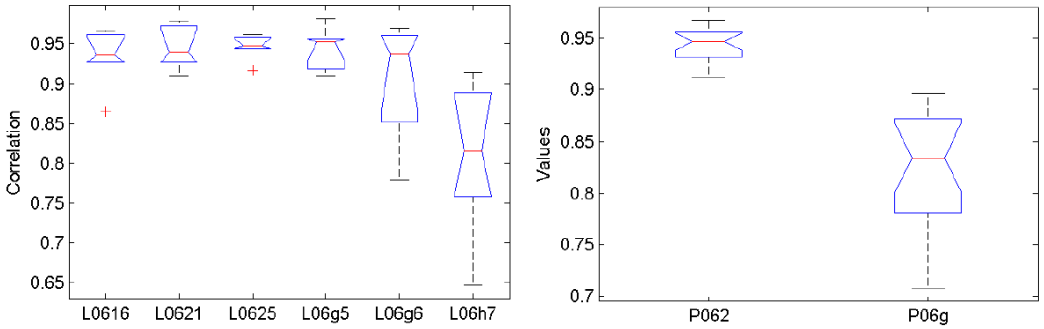


Figura 11: Misure di coerenza (intra-locutore, a sinistra) e di congruenza (inter-locutore, a destra) (dati di *AMPER-ITA* in *AMPER* 2011, trattamenti proposti in Romano *et alii*, in c. di p.).

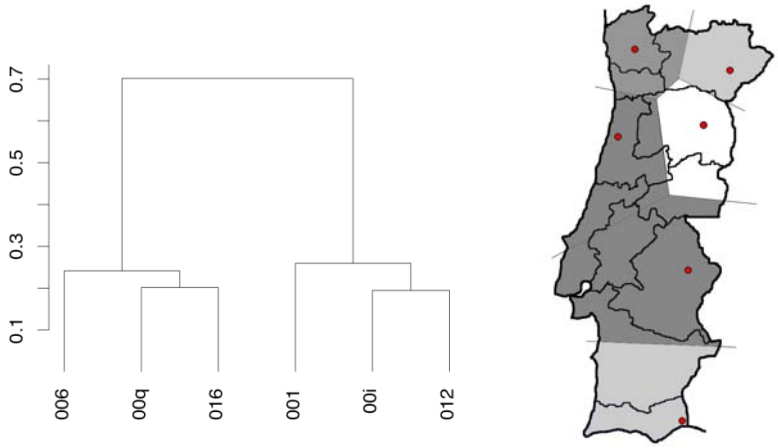


Figura 12: Dendrogramma con *clustering* gerarchizzato (a sinistra) e mappa dialettometrica (a destra) della distanza prosodica media tra dati di diverse regioni del Portogallo (dati di *AMPER-POR*) in riferimento al punto d'inchiesta 016 (Trinta, Beira Alta, in bianco nella mappa; cfr. Moutinho *et alii* 2011). Il grado di annerimento è una funzione lineare della distanza tra le modalità di resa prosodica delle stesse frasi nei diversi punti (da Nord a Sud: 001 = Prado - Braga, Minho, 006 = Alfândega da Fé, Trás-os-Montes, 012 = Aradas, Beira Litoral, 00i = Monforte, Alto Alentejo, 00q = Monte Gordo, Algarve).

Ovviamente le direzioni di sviluppo più interessanti sono quelle offerte dalle possibilità di ponderazione di queste distanze in base all'inclusione delle condizioni di resa di variabili linguistiche (attese o percepite), come le posizioni degli accenti lessicali. I tentativi più promettenti sono quelli che tengono conto dei valori presenti nelle posizioni di maggiore salienza (in termini multiparametrici) e che sono già stati in parte integrati nelle misure di distanza sfruttate per le analisi multi-dimensionali svolte finora e illustrate sopra.

Un altro aspetto interessante deriva dall'applicazione di misure d'intercorrelazione all'insieme degli enunciati di un dato locutore o di un dato punto per studiare gli effetti della realizzazione di accenti in diverse sedi sulla caratterizzazione del profilo melodico

[40] In questo caso il modello presente nei dati di riferimento sembrerebbe ritrovarsi in aree periferiche, in contrasto evidente con la diffusione di un modello verso il quale convergerebbero le aree costiere centro-settentrionali. L'indicazione è naturalmente da confermare nei dati linguistici delle singole inchieste.

complessivo (una stima delle diverse misure di sensibilità alle variazione di posizione degli accenti è in Romano & Miotti 2008)[41].

7 Conclusioni

Il cantiere del progetto *AMPER* conta su una storia ventennale (1991-2011) durante la quale ha coinvolto dialettologi e altri linguisti interessati allo studio della variazione prosodica. Queste finalità vi sono perseguite con un approccio descrittivo, rivolto al confronto di diversi indici fonetici della strutturazione prosodica degli enunciati nello spazio dialettale dei dialetti romanzi storici e delle varietà regionali delle lingue romanze.

Tra gli obiettivi del progetto vi è la costituzione di un corpus di parlato che, oltre a essere allestito secondo modalità convenzionali di confronto tali da garantire una certa omogeneità (e uniformità) tra i dati raccolti, consenta l'applicazione di un protocollo d'analisi perseguito attraverso un certo numero di tappe analitiche variamente condivise dai ricercatori che vi prendono parte.

I partner che vi aderiscono possono infatti decidere di condividere solo alcune di queste fasi procedurali e conferire i dati alla *BD* in base a garanzie di qualità soggette ad auto-valutazione (secondo uno schema che abbiamo qui riprodotto).

Al di là di alcuni vincoli piuttosto stringenti sulle modalità di raccolta dei dati e sul formato delle registrazioni sonore e dei documenti associati, per la sua natura aperta e permeabile, *AMPER* è in costante crescita e si sta arricchendo progressivamente di dati che l'hanno portato a una prima pubblicazione (*AMPER* 2011).

Oltre a un progressivo miglioramento delle modalità di astrazione linguistica dei fenomeni osservati, metodi originali di trattamento automatico dei dati presenti nella *BD* sono attualmente studiati in vista di una modellizzazione funzionale e la definizione di criteri oggettivi di confronto e rappresentazione (in termini dialettometrici e geoprosodici).

L'allargamento della rete di collaborazioni (o il maggiore coinvolgimento di ricercatori finora solo marginalmente interessati) dovrebbe portare a migliorare una copertura geografica attualmente ancora insoddisfacente, così come alla determinazione di elementi condivisi per condurre uno studio tipologico e giungere a un modello teorico convincente per un'analisi dell'organizzazione ritmico-intonativa in termini variazionali.

Bibliografia

AMPER-AST = http://www.unioviedo.es/labofone/ (Carmen Muñiz Cachón, Universidad de Oviedo – Spagna).

AMPER-CAN = http://webpages.ull.es/users/labfon/proampercan/ (Josefa Dorta Luís, Universidad de la Laguna, Tenerife – Isole Canarie, Spagna).

AMPER-CAT = http://www.ub.edu/labfon/amper/index_ampercat_cat.html (Ana Maria Fernández Planas & Eugenio Martínez Celdrán, Universitat de Barcelona – Spagna).

[41] Limitatamente al tipo di strutture qui studiate, se in una lingua determinata la correlazione tra sequenze di valori di f_0 di enunciati con posizioni accentuali non allineate è molto più alta di quella di un'altra lingua, si deduce che l'organizzazione intonativa della prima riguardo a questa variabile non risente molto delle modalità di realizzazione degli accenti: queste possono dunque essere considerate meno vincolanti che nella seconda. Lo studio suggerisce una maggiore invariabilità degli enunciati nelle varietà di spagnolo osservate vs. quelle d'italiano, almeno per quello che riguarda gli accenti di tonìa (nucleari).

AMPER-GAL = http://ilg.usc.es/amper/ (Elisa Fernández Rei, Universidad de Santiago de Compostela – Spagna).

AMPER-ITA = http://www.lfsag.unito.it/amper/ (Antonio Romano, Università di Torino - Italia).

AMPER-POR = http://pfonetica.web.ua.pt/AMPER-POR.htm (Lurdes de Castro Moutinho, Universitade de Aveiro – Portogallo).

AMPER-ROM = http://amprom.uaic.ro/ (Adrian Turculeţ, Universitătea "Al. Ioan Cuza", Iaşi – Romania).

AMPER (2005) – « Projet *AMPER* - Atlas Multimédia Prosodique de l'Espace Roman », in: *Géolinguistique*, hors-série 3 (a cura di J.P. Lai).

AMPER (2011) – "Intonations Romanes", in: *Géolinguistique*, hors-série 4 (a cura di P. Mairano).

Aveiro (2007) – L. de Castro Moutinho & R.L. Coimbra (a cura di), *Actas das I Jornadas Científicas AMPER-POR* (Aveiro, Portogallo, 29-30 ottobre 2007).

Avolio F. & Romano A. (2009): "Nuovi dati fonetici e dialettologici ai margini dell'area Lausberg: le varietà di Aliano e Alianello", in: L. Romito, V. Galatà & R. Lio (a cura di), *La fonetica sperimentale: metodi e applicazioni* (*Atti del IV Convegno Nazionale AISV - Associazione Italiana di Scienze della Voce*, Cosenza, Italia, 3-5 dicembre 2007), Torriana (RN), 372-404.

Avolio F. & Romano A. (2010): "Ai margini dell'area Lausberg: le varietà di Aliano e Alianello nei risultati di un'indagine dialettologica e fonetica", in: M. Iliescu, H. Siller-Runggaldier & P. Danler (a cura di), *Atti del XXV Congrès International de Linguistique et de Philologie Romanes* (Innsbruck, Austria, 3-8 settembre 2007), Berlin, New York, vol. 4, 25-36.

Bruce G. & Gårding E. (1978): "A prosodic typology for Swedish dialects",in: *Nordic prosody, Trav. de l'Inst. de Ling. de Lund*, 13, 219-228.

Canepari L. (1985): *L'intonazione. Linguistica e paralinguistica*, Napoli.

Canepari L. (2004): *Manuale di fonetica*, Monaco.

Contini M. (1992): "Vers une géoprosodie", *Atti del « Nazioarteko Dialektologia Biltzarra Agiriak »* (Bilbao, 1991), Bilbao: Publ. Real Academia de la Lengua Vasca, 83-109.

Contini M. (2008): « Le projet *AMPER* : passé, présent et avenir », in: *Aveiro* (2007), 9-19.

Contini M./Profili O. (1989): « L'intonation de l'italien régional. Un modèle de description par traits », in: A. Bothorel *et alii* (a cura di), *Mélanges de phonétique expérimentale offerts à P. Simon*, Strasburgo, 854-870.

Contini M./Romano A. (2011): "Au départ, un projet de dialectologues", in: *AMPER* (2011), 3-11.

Contini M. et alii (2009): "L'avancement des recherches en géoprosodie et le projet *AMPER*", *EFE XVIII* (2008), 109-122.

Contini M. et alii (in c. di p.): « Présentation du DVD "Intonations Romanes" », *Atti del V Congreso Internacional de Fonética Experimental* (Cáceres, Spagna, 25-28 ottobre 2011), in c. di p.

Coşeriu E. (1958): Sincronía, diacronía e historia: el problema del cambio lingüístico, Madrid: Gredos (ed. it. Sincronia, diacronia e storia, Torino, 1979).

Dorta J. (a cura di) (2007): *La prosodia en el ámbito lingüístico románico* (*Atti delle "III jornadas científicas del proyecto AMPER"*, La Laguna – Tenerife, Isole Canarie, 24-25 ottobre 2006), Santa Cruz de Tenerife.

EFE XIV (2005) – Estudios de Fonética Experimental, XIV (volume monografico dedicato ad AMPER-en-España).

EFE XVIII (2008) – Estudios de Fonética Experimental, XVIII (ed. speciale per il Simposio Internacional - 30è aniversari del laboratori de fonètica de la UB, Barcellona, Spagna, 2-6 dicembre 2008).

EFE XX (2011) – *Estudios de Fonética Experimental*, XX (volume contenente studi-pilota nel campo della valutazione di distanze prosodiche oggettive tra le varietà di diversi domini linguistici).

Felloni M.C. (2011): *Prosodia sociofonetica: l'italiano parlato e percepito a Parma*, Milano.

Felloni M.C. & Avesani D. (2010): "La percezione della interrogativa globale nell'italiano regionale di Parma". In: F. Cutugno, P. Maturi, R. Savy, G. Abete & I. Alfano (a cura di), *Parlare con le macchine, parlare con le persone* (*Atti del VI convegno AISV*, Napoli, Italia, 3-5 febbraio 2010), Torriana (RN): EDK, 139-171.

Fernández Planas A.M., Roseano P., Martínez Celdrán E. & Romera Barrios L. (2011): "Aproximación al análisis dialectométrico de la entonación en algunos puntos del dominio lingüístico catalán", *EFE XX*, 33-55.

Goebl H. (1983): « Eléments d'analyse dialectométrique (avec application à l'AIS) », *Revue de Linguistique Romane*, 45, 349-420.

Goebl H. (1996): « La convergence entre les fragmentations géo-linguistiques et géo-génétiques de l'Italie du Nord », *Revue de linguistique romane*, 60, 25-49.

Gooskens Ch. (1997): On the Role of Prosodic and verbal information in the perception of Dutch and English language varieties, Katholieke Universiteit Nijmegen, Doctoral dissertation.

Granada (2008) - A. Pamies, M.C. Amorós & J.M. Pazos (a cura di), *Experimental Prosody* (Actas del IV Congreso Int. de Fonética Experimental, Granada, Spagna, 23-25 febbraio 2008), *Language Design*, special issue 2.

Grenoble (2001) - *I rencontre internationale AMPER* (Grenoble, Francia, giugno 2001).

Grenoble (2004) - *II rencontre internationale AMPER* (Grenoble, Francia, giugno 2004), v. *AMPER* (2005).

't Hart J., Collier R. & Cohen A. (1990): *A perceptual study of intonation*, Cambridge: Cambridge University Press.

Hermes D.J. (1998): « Measuring the Perceptual Similarity of Pitch Contours », *J. Speech, Language and Hearing Research*, 41, 73-82.

Hirst D., Di Cristo A. & Espesser R. (2000): "Levels of representation and levels of analysis for the description of intonation systems", In: M. Home (ed.), *Prosody: Theory and Experiment*, Dordrecht: Kluwer, 51-87.

Interlandi G.M. (2004): *L'intonazione delle interrogative polari nell'italiano parlato a Torino: tra varietà regionale e nuova* koiné, Tesi di Dottorato in Linguistica (inedita), Università degli Studi di Pavia.

Lai J-P. (2002): *L'intonation dans le parler de Nuoro*, Thèse de Doctorat en Sciences du Langage, Université Stendhal (resp. Michel Contini) – Grenoble III.

Lai J.P. & Rilliard A. (2008): "Outils pour le calcul et la comparaison prosodique dans le cadre du projet AMPER- l'exemple des variétés Occitane et Sarde", In: A. Turculeţ (2008), 217-229.

Lai J.P., Romano A. & Roullet S. (1997): "Analisi dei sistemi prosodici di alcune varietà parlate in Italia: problemi metodologici e teorici", *Bollettino dell'Atlante Linguistico Italiano*, 21, 23-70.

Mairano P. & Romano A. (2008). "Distances rythmiques entre variétés romanes". In: A. Turculeţ (2008), 251-272.

Mairano P. & Romano A. (2012): "Testing the perception of speech rhythm on natural and artificial stimuli", *Proceedings of the 6th International Conference Speech Proso-*

dy 2012 (Shanghai, Cina, 22-25 maggio 2012), vol. II, Shanghai: Tongji University Press, 482-485.

Martin Ph. (2003): « *ToBI* : l'illusion scientifique ? », In: V. Aubergé, A. Lacheret-Dujour & H. Lœvenbruck (eds.), *Actes des Journées Prosodie 2001* (Grenoble, Francia, 10-11 ottobre 2001), 109-113.

Martin Ph. (2012): The Autosegmental-Metrical Prosodic Structure: not fit for French?, *Proceedings of the 6th International Conference Speech Prosody 2012* (Shanghai, Cina, 22-25 maggio 2012), vol. I, Shanghai: Tongji University Press, 131-134.

Mertens P. (2004): "The Prosogram : Semi-Automatic Transcription of Prosody based on a Tonal Perception Model", *Proceedings of Speech Prosody 2004* (Nara, Giappone, 23-26 marzo 2004), 549-552.

MIDL (2004): Actes du colloque "Identification des langues et des variétés dialectales par les humains et par les machines" (Paris, 29-30 nov. 2004), Paris: École Nationale Supérieure des Télécommunications.

Moutinho L. de Castro, Coimbra R.L., Rilliard A. & Romano A. (2011): "Mesure de la variation prosodique diatopique en portugais européen", *EFE XX*, 33-55.

Moutinho L. de Castro, Coimbra R.L., Pereira Bendiha U., Romano A., Contini M. (2004): "Estudo comparativo da variação prosódica em duas línguas românicas: o Português e o Italiano", *Atti dell'Incontro Annuale dell'APL – Associação de Linguística Portuguesa* (Lisbona, Portogallo, 1-3 ottobre 2003), Lisbona: APL, 719-723.

Panconcelli-Calzia G. (1939): « Über die " Frageton " im Italienischen », *Vox Romanica*, 4/1, 35-47.

Peters J., Gilles P., Auer P. & Selting M. (2003): "Identifying regional varieties by pitch information: A comparison of two approaches", *Proceedings of the 15th International Congress of Phonetic Sciences* (Barcellona, Spagna, 3-9 agosto 2003), 1065-1068.

PRAAT - Boersma P. & Weenink D., *Praat: doing phonetics by computer (Version 5.0)*, Università di Amsterdam, 2007 (*Public domain software*: http://www.praat.org).

RILI (2011) - *Revista Internacional de Lingüística Iberoamericana*, Vol. IX, No. 1 (17) (a cura di Y. Congosto Martín).

Romano A. (1997): "Persistence of prosodic features between dialectal and standard Italian utterances in six sub-varieties of a region of Southern Italy (Salento): first assessments of the results of a recognition test and an instrumental analysis". *Atti di EuroSpeech'97* (*5th European Conference on Speech Comm. and Technology*, Rodi, Grecia, 22-25 settembre 1997), 175-178.

Romano A. (2001a): Analyse des structures prosodiques des dialectes et de l'italien régional parlés dans le Salento: approche linguistique et instrumentale, Lille: Presses Univ. du Septentrion (10 vol. della Thèse de Doctorat de l'Université Stendhal de Grenoble discussa nel dic. 1999, resp. M. Contini).

Romano A. (2001b): "Variabilità degli schemi intonativi dialettali e persistenza di tratti prosodici nell'italiano regionale: considerazioni sulle varietà salentine", In: A. Zamboni, P. Del Puente & M.T. Vigolo (a cura di), *La dialettologia oggi fra tradizione e nuove metodologie* (*Atti del Conv. Internazionale*, Pisa, Italia, 10-12 febbraio 2000), Pisa: ETS, 73-91.

Romano A. (2003): "Applicabilité des systèmes de transcription et d'analyse de l'intonation aux cas de variabilité dialectale présentés par la situation géoprosodique italienne", In: V. Aubergé, A. Lacheret-Dujour & H. Lœvenbruck (eds.), *Actes des Journées Prosodie 2001* (Grenoble, Francia, 10-11 ottobre 2001), 115-118.

Romano A. (2004): "Indices acoustiques suprasegmentaux dans la caractérisation des langues romanes: identification de variétés linguistiques et description des traits prototypiques", In: *MIDL* (2004), 91-92.

Romano A. (2005): "Utilisation des données *AMPER* pour une description de la variation linguistique : tests de perception et contrôles statistiques", in: *AMPER* (2005), 39-64.

Romano A. (2006): "Sulla variazione intonativa di frasi dichiarative e interrogative romene", in: *Quaderni di Studi Italiani e Romeni*, 2, 121-133.

Romano A. (2008): "Éléments théoriques et pratiques de l'analyse multiparamétrique de la prosodie dans le cadre d'*AMPER*", in: *Aveiro* (2007), 115-126.

Romano A., Contini M., Lai J.P. & Rilliard A. (2011): "Distancias prosódicas entre variedades románicas en el marco del proyecto *AMPER*", in: *RILI* (2011), 17-26.

Romano A. & Interlandi G. (2002): "Quale intonazione per il torinese?", In: A. Regnicoli (a cura di), *La fonetica acustica come strumento di analisi della variazione linguistica in Italia* (*Atti delle XII Giornate di Studio del GFS*, Macerata, Italia, 13-15 dicembre 2001), Roma: Il Calamo, 117-122.

Romano A. & Maairano P. (2010): "Speech rhythm measuring and modelling: pointing out multi-layer and multi-parameter assessments", in: Michela Russo (a cura di), *Prosodic Universals: comparative studies in rhythmic modeling and rhythm typology*, Roma: Aracne, 79-116.

Romano A. & Mattana P. (2008): "Comparaison des corpus d'AMPER-ITA: l'incidence diatopique de la variable focus dans les données salentines et de l'aire centrale", in: *Granada* (2008), 293-301.

Romano A. & Miotti R. (2008): "Distancias prosódicas entre variedades románicas", in: A. Turculeţ (2008), 231-249.

Romano A. & Miotti R. (2009): "Un contributo per il confronto tra l'intonazione veneta e quella andalusa", In: L. Romito *et alii* (a cura di), *La fonetica sperimentale: metodi e applicazioni* (*Atti del IV Convegno Nazionale AISV - Associazione Italiana di Scienze della Voce*, Cosenza, Italia, 3-5 dicembre 2007), Torriana (RN): EDK, 62-76.

Romano A., de Castro Moutinho L., Coimbra R.L. & Rilliard A. (in c. di p.): "Medidas da variação prosódica diatópica no espaço românico", in: *Atti del VII giornate del GSCP della Società di Linguistica Italiana* (Belo Horizonte, Brasile, 29 febbraio – 3 marzo 2012), in c. di p.

Saramago J. (1986): « Différenciation lexicale (un essai dialectométrique appliqué aux matériaux portugais de l'ALE) », in: *Géolinguistique*, 2, 1-31.

Saramago J. & Bettencourt Gonçalves J. (2003): « Diferenciação lexical interpontual nos Açores (estudo dialectométrico aplicado em materiais do ALEAç) », in: R. Caprini (a cura di), *Parole romanze. Scritti per Michel Contini*, Alessandria: Dell'Orso, 421-440.

Sevilla (2010) - *IV incontro internazionale AMPER* (Sevilla, Spagna, 18-19 febbraio 2010), v. *RILI* (2011).

Sorianello P. (2006): *Prosodia: modelli e ricerca empirica*, Roma: Carocci.

Turculeţ A. (a cura di) (2008), *La variation diatopique de l'intonation dans le domaine roumain et roman*, Iaşi: Editura Universităţii Al. I. Cuza (Atti del convegno di Iaşi, Romania, 21-23 ottobre 2008).

Turculeţ A., Botoşineanu L., Minuţ A.M. & Romano A. (2004): "L'intonation du roumain au sein du projet AMPER", in: *Bollettino dell'Atlante Linguistico Italiano*, 27, Torino (2003), 269-274.

Turculeţ A., Botoşineanu L., Minuţ A.M. & Romano A. (2005): "Recherches acoustiques sur quelques aspects régionaux de l'intonation du roumain littéraire", in: AMPER (2005), 281-310.

Vaissière J. & Boula de Mareüil Ph. (2004): « Divers aspects de l'identification d'une langue ou d'un *accent*: du segmental à la prosodie », in: *MIDL*, 1-5.

Outils pour la géolinguistique automatisée

Gotzon Aurrekoetxea (UPV/EHU) & Charles Videgain (UPPA-Iker)

La géolinguistique a parcouru un vaste itinéraire durant ces dernières décennies, surtout dans l'utilisation de ressources informatiques comme aide au traitement et à l'analyse de données. Il existe actuellement un vaste éventail d'équipes de recherche qui utilisent quotidiennement de tels outils informatisés. L'équipe de recherche EUDIA (*EUskal DIAlektologia*) compte sur différents outils pour le traitement de l'information dialectale, outils qui fonctionnent *online*: l'outil "CorpusLem" est un élément qui transforme les textes en données. Il permet de travailler en *online* comme en local, en déchargeant un fichier *xls*, lequel peut ensuite être chargé sur le réseau. La base de données EDAK est un outil de gestion des données, qui supporte en lui-même diverses bases de données et qui, à partir des outils de gestion des données, est consultable sans avoir besoin de s'enregistrer. La base héberge actuellement près de 100.000 données sur toutes les variétés de l'euskara ou langue basque. Ces données peuvent servir à des analyses diatopiques ou diastratiques. La base de données héberge des données audiovisuelles, de telle forme que l'usager puisse en même temps lire et écouter la réponse intégrée dans la base de données. Enfin, l'outil cartographique et dialectométrique "DiaTech" possède divers modules: module de la base de données elle-même, module statistico-quantitatif et module cartographique. Il s'agit aussi d'un outil *online*. Le module cartographique a deux options: l'une de cartographie thématique avec possibilité d'un atlas parlant et une autre de cartographie thématique. L'outil permet l'importation de bases de données provenant d'autres projets géolinguistiques, la réalisation d'analyses statistiques et une cartographie quantitative.

1 Géolinguistique automatisée: état de la question

Dans le champ de la dialectologie on tient généralement pour assuré le fait que la quantification des données a commencé avec les travaux quantitatifs de Jean Séguy et son souci de trouver l'espace qui réunisse l'essence du gascon. Hans Goebl comme Henri Guiter ont poursuivi leurs travaux dialectométriques et c'est sans doute H. Goebl qui a su internationaliser cette problématique et pousser au développement informatisé de cette méthode quantitative. Peu de chercheurs doutent absolument de la valeur heuristique de la DM comme outil de l'analyse diatopique et le nombre d'applications informatiques ne fait qu'augmenter.

La dialectométrie a été l'une des avancées les plus significatives de la géolinguistique du XXème siècle. Sans doute l'une des avancées les plus importantes. La possibilité d'utiliser une grande quantité de données pour l'étude de la variation linguistique permet de dépasser la situation endémique de la dialectologie dès ses débuts, c'est-à-dire des études basées uniquement sur des critères et des variables linguistiques peu nombreuses. L'étude des frontières dialectales est beaucoup plus accessible et plus probante avec cette nouvelle méthodologie.

Dans un autre ordre de choses, les travaux de l'atlas parlant ou sonore ont dû attendre que la technologie de l'étude du signal audio soit d'accès facile pour les dialectologues. Les premiers travaux remontent aux années 1990, comme le sait fort bien Roland Bauer, un des pionniers dans le domaine.

Nous avons nous-même eu connaissance du projet d'atlas sonore en 1992 à l'occasion d'un congrès organisé au sein d'Euskaltzaindia / Académie de la langue basque et dont nous étions les responsables académiques. Lors de ce congrès, le Professeur H. Goebl donna deux conférences, l'une sur les atlas sonores et l'autre sur la DM.

Dans sa conférence sur les atlas sonores intitulée "L'atlas parlant dans le cadre de l'Atlas linguistique du ladin central et des dialectos limitrophes (ALD)" (cfr. Goebl 1992), il avait montré la possibilité de présenter les données directement depuis la réalisation

du locuteur interrogé jusqu'à l'oreille des enquêteurs-transcripteurs ou du grand public. Quiconque pouvait dorénavant, depuis son bureau ou son appareil, écouter les données audiovisuelles brutes de l'enquête. C'était là un pas de géant pour les études géolinguistiques et qui rendait possible la modernisation du savoir-faire dialectologique des atlas linguistiques.

Cette avancée souhaitée par les responsables techniques de l'Atlas *Euskal Herriko Hizkuntz Atlasa* (EHHA) basque n'a pas été possible en son temps au sein de l'Académie de la langue basque, mais nous avons pu développper d'autres projets qui prennent en compte cette dimension, comme le projet EDAK.

L'atlas EHHA poursuit sa publication de documents écrits (cartes, responsaires, index) et quatre volumes ont paru à ce jour dont le numéro 4 en septembre 2012, et nous travaillons à l'élaboration des cartes du tome 9. L'atlas est publié en version papier (en peu d'exemplaires) et en format numérique. La version *en ligne* se trouve à l'adresse web de l'Académie basque (www.euskaltzaindia.net/dialektologia). La matière des livres peut être téléchargée en pdf mais on peut y faire des recherches de cartes par réponses ou par lemme. Actuellement, une difficulté est à signaler aux non bascophones: l'atlas est à ce jour uniquement rédigé et consultable en basque.

La dialectométrie comme les atlas parlants et sonores se sont fondés sur les progrès de la technologie, en tirant profit de ces possibilités techniques : la dialectométrie avec une statistique automatisée et informatisée, les seconds par la facilité d'enregistrement et de digitalisation de la voix humaine.

2 CorpusLem

Le groupe de recherche EUDIA, créé au sein de l'Université du Pays Basque, mais dont font partie des chercheurs d'autres universités (Université de Pau et des Pays de l'Adour (UPPA), Bamberg) a pour objectif l'étude de la variation linguistique (géo-et sociolinguistique) de la langue basque et la création d'outils nécessaires pour cette étude.

L'un de ces premiers outils créé a été "CorpusLem", consultable sur le site http://aholab.ehu.es/CorpusLem/login.html). Cet outil a été créé en collaboration avec le centre IKER (Centre de recherche sur la langue et les textes basques) UMR (Unité Mixte de Recherche) CNRS (Centre national de la recherche scientifique, Bayonne) et a voulu répondre à la nécessité de traiter des textes dialectaux en les transformant en un format "Donnée' lors du projet Bourciez (Aurrekoetxea/Videgain 2009, Aurrekoetxea 2011, Aurrekoetxea/Videgain/Iglesias 2004) puis lors du projet Sacaze (Aurrekoetxea 2011, Aurrekoetxea/Videgain 2012).

Dans les deux projets, nous traitons des textes dialectaux d'origine géographique homogène mais produits par des traducteurs différents et selon des orthographes non normées. Dans le projet Bourciez, il s'agit du texte bien connu de la Parabole de l'enfant prodigue, dans le projet Sacaze de deux légendes: La légende de Barbazan et La légende de Tantugou. Les textes ont été recueillis à la fin du XIXème siècle et concernent 150 localités bascophones, et nous n'avons pas pris à notre charge les textes rédigés en occitan ou catalan.

Notre intention de réaliser une exploitation géolinguistique de ces textes nous obligea à passer du texte original en format texte au format donnée pour introduire le texte dans une base de données et une visualisation cartographique des données. Ce passage du texte vers le format donnée est réalisé à partir du programme CorpusLem. Le programme sert aussi à la création de dictionnaires dialectaux.

L'outil fonctionne en 5 langues (basque, anglais, français, castillan et catalan) et il est d'une grande simplicité puisque ne comportant que deux écrans. Le premier écran sert à introduire l'utilisateur et son mot de passe, le second écran sert à convertir les textes de format *.tst* à une base de données au format MySQL), avec la possibilité de copie de données, élaboration de dictionnaires, etc.

2.1 Ecran de présentation

Il est indispensable d'être enregistré pour utiliser cet outil. La marche à suivre est rapide.

2.2 Ecran de travail

L'écran de travail est partagé en sections (fig. 1):

- Manuel d'instructions (à droite de l'écran)
- Champ de gestion des projets, des utilisateurs, des bases de données.
- Champ des projets (à droite sur l'écran):
- Copie de la base de données du projet
- Projet actuel
- Gestion du projet
- Gestion des règles internes au projet

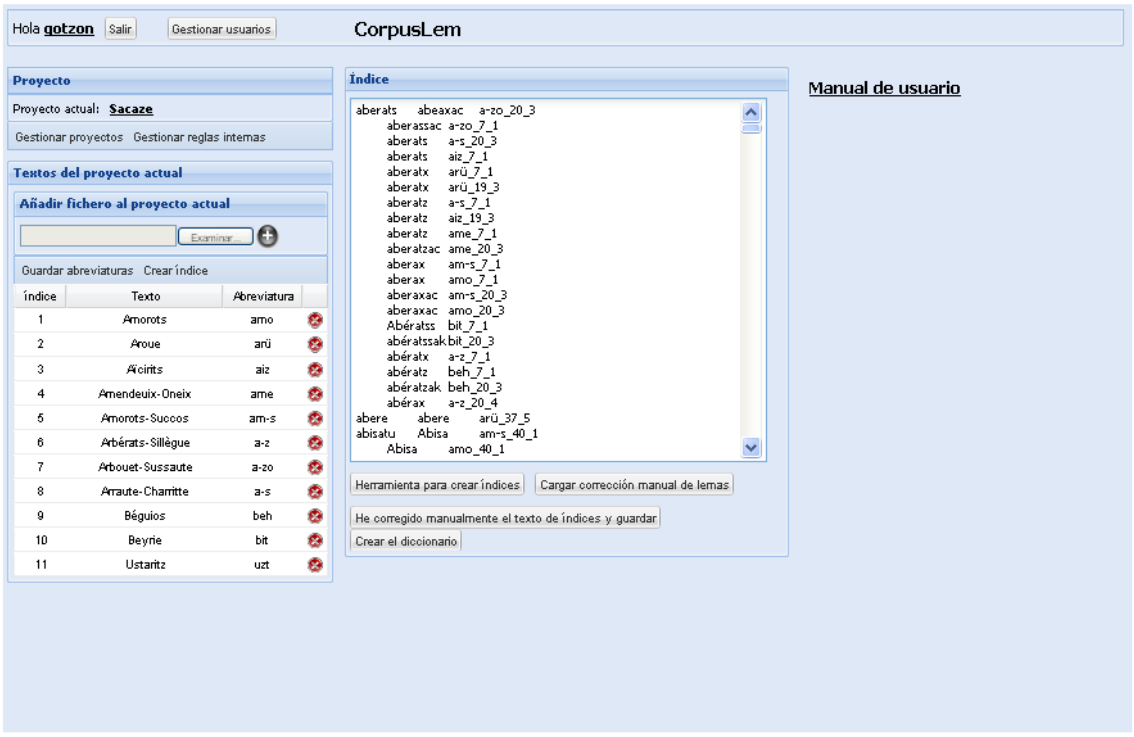


Fig. 1. L'outil *CorpusLem* : ecran de travail

Au moyen de ces règles internes, le chercheur peut adapter la graphie des textes en accord avec les buts du projet. Quatre options sont possibles (fig. 2):

a) 'Equivalent". Sous ce titre, si un vocable possède un grand nombre de variantes qui peuvent être représentées en une seule, le système choisit cette variante comme principale et y réunira les autres variantes. C'est ainsi que sous le factitif basque 'arazi', le système regroupera les variantes 'aazi' (avec perte de la vibrante intervocalique) ou 'aaci' ou 'aasi'.

b) Un autre groupe de règles essaie de réguler les terminaisons, dont on sait qu'elles sont nombreuses de par le système de déclinaisons en langue basque. Sous la variante 'arazi' indiquée ci-dessous, on trouvera donc des formes déclinées comme 'arazia', 'aazia', 'arazten'.

c) "Joindre". Le système joint des termes qui sont séparés dans le texte traité. Ainsi 'aitasso amassoac' ('père et mère') seront réunis par un trait d'union.

d) "Séparer". Ce cas consiste inversement à séparer des termes qui sont écrits d'un seul tenant dans le texte traité. En particulier, les verbes périphrastiques voient souvent l'auxiliaire lié au verbe conjugué et notre règle opère la séparation.

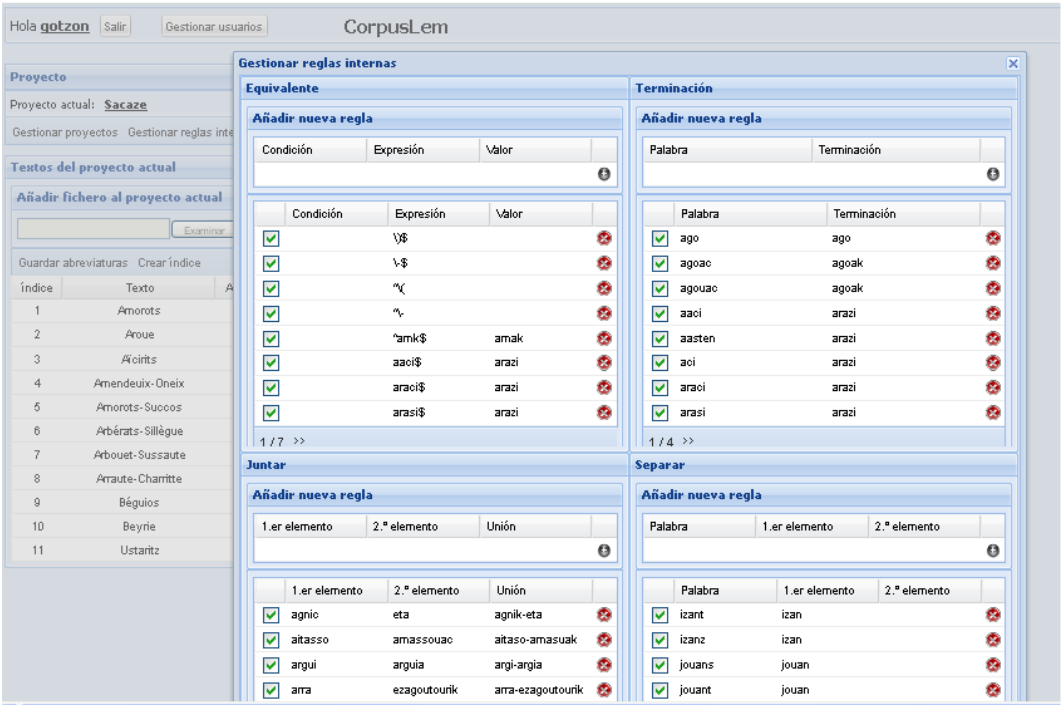


Fig. 2. L'outil *CorpusLem* : gestion des règles internes

- Textes du projet actuel, dans lesquels sont présents les textes actifs de chaque projet, avec la possibilité d'ajouter des documents, d'en effacer, etc.
- Champ de travail (centre de l'écran). C'est le centre de travail, à travers lequel le chercheur peut construire le dictionnaire des formes de tout le texte. Le chercheur peut le faire à partir de l'outil lui-même ou le charger en format *xls*, travailler en local et ensuite l'introduire de nouveau dans l'outil à partir du bouton "charger la correction manuelle des lemmes", et ensuite créer le dictionnaire des formes à partir de l'option "créer dictionnaire".
- Enfin, l'outil crée une base de données MySQL avec les données des textes.

3 Le corpus EDAK: l'atlas sonore basque

Le deuxième instrument créé par le groupe EUDIA est une base de données transcrites et sonores. Le corpus EDAK est un projet mené par le groupe de recherche EUDIA dont l'objectif est l'analyse de la variation. Cette analyse ne se limite pas à la variation géolinguistique, mais aussi la variation sociolinguistique et stylistique, chaque fois que la nature des données le permet.

L'un de ses projets est le "Corpus oral du basque dialectal-EDAK/Euskara Dialektalaren ahozko corpusa" hébergé sur le site http://aholab.ehu.es/edak/2/.

3.1 Caractéristiques du corpus

Ce corpus contient actuellement près de 100 000 registres. C'est le seul corpus oral en basque dialectal accessible sur le Web. Le corpus est assez réduit mais d'une grande richesse et ses caractéristiques essentielles sont les suivantes :

- 403 questions ainsi réparties:
- 120 questions relevant du lexique commun.
- 62 questions de morphologie nominale et verbale.
- 20 questions de syntaxe.
- 179 questions sur l'accent dans les mots monosyllabiques, bissyllabiques, trisssyllabiques, selon la déclinaison à l'indéfini, au singulier ou au pluriel.
- 22 questions sur l'intonation, 12 dans des phrases énonciatives, 6 interrogatives totales, 4 interrogatives partielles.
- Les points d'enquête sont au nombre de 100, 75 dans la partie péninsulaire, 25 dans la partie continentale.
- Les informateurs sont classés en adultes ou jeunes, donc de deux générations différentes, hommes pour le lexique, la morphologie, la syntaxe, et femmes pour la prosodie. On sait que l'étude de la variation sociolinguistique est d'autant plus importante que la langue basque est en situation de nivellement linguistique fort, depuis la création d'un basque unifié ou standard en 1968 et l'implantation récente d'un système d'enseignement de la langue basque, soit dans le réseau des ikastola (écoles privées immersives) soit dans le réseau public, sans oublier le rôle des médias en langue basque. Les diverses générations ont donc un profil fort différent; les adultes ont souvent été scolarisés selon le modèle espagnol ou français (scolarité exclusivement en espagnol ou français), tandis qu'une bonne part des jeunes sont passés par une scolarisation en langue basque à un degré variable. Le Pays Basque est donc un remarquable laboratoire pour l'étude du nivellement linguistique et les mécanismes qu'il génère mais aussi pour observer les formes et variétés plus conservatrices.
- Caractéristiques de la base de données :
    - Format: MySQL
    - L'information est retranscrite selon l'alphabet IPA à partir d'un écran incrusté dans la base de données, de telle sorte que l'utilisateur n'a pas

à saisir un signe phonétique mais à le choisir sur l'écran et à les copier dans la base.

L'information acoustique, sur l'annotation-étiquetage du signal sonore se fait à travers les programmes *SFSWin*. L'information acoustique peut être écoutée à partir d'un programme de son (Praat) et disposée pour être utilisée dans des analyses acoustiques.

Toute l'information indiquée ici est disponible sur le réseau à l'intention de tous les utilisateurs, sans nécessité de s'identifier dans l'accès à la base de données (creative common licencia)

La base de données EDAK est un outil de gestion des données, au format MySQL, qui supporte diverses bases de données et qui, à partir des outils de gestion des données, offre deux nouveaux modules: le module d'aide à la lemmatisation des données, et le module de conversion du format base de données au format corpus (format TEI).

Le module d'aide à la lemmatisation est nécessaire pour une exploitation linguistique ou géolinguistique des données: la lemmatisation permet une cartographie aréale, en utilisant divers outils de visualisation. Dans notre cas, nous avons recours de préférence à la couleur, mais d'autres options sont possibles.

Le module de conversion du format MySQL au format TEI est indispensable aujourd'hui. Il permet de passer d'un format privé, individuel à un format socialisé international. La dialectologie ne peut se dispenser de franchir ce palier en ayant recours à cette technologie.

Cette base de données héberge aussi bien des données écrites que des données audio, si bien que l'utilisateur peut lire ou écouter la réponse dans la base de données.

Les études facilitées par cette bases sont présentées dans les colloques ou congrès et publiées dans les diverses revues spécialisées.

4 L'outil dialectométrique DiaTech

L'outil cartographique et dialectométrique *DiaTech* utlise divers modules: module de données propres (Base de données EDAK), modèle statistico-quantitatif et module cartographique. Ce dernier propose l'option de préparer un atlas parlant et l'option de cartographie de faits synthétiques ou dialectométriques.

4.1 Menu principal

Tout au long de l'application Web, le menu principal reste visible en haut de l'écran et fournit les principaux champs de l'application. Outre ces différents champs, la possibilité est donnée de changer de langue de travail, de changer le mot de passe de l'utilisateur et celle de quitter le programme.

Les champs principaux sont les suivants:

- Début: dans le champ "début", est fourni l'objectif de l'application Diatech sur le Web.
- Projets: dans le champ "projets", apparaît le nom des projets à gérer et utiliser
- Aide: dans ce champ, est donné le guide d'utilisation de Diatech
- Compte: ce champ attribue un compte à chaque utilisateur
- Langues: le programme peut être utilisé en plusieurs langues, basque, espagnol, anglais, choix à enrichir ultérieurement.

4.2 Les projets

Le champ "Projet" est le champ principal du programme. Dans ce champ chaque utilisateur peut sélectionner le projet qu'il gère et en créer de nouveaux.

A chaque projet créé est attribué nécessairement un nom. Ce nom doit être propre au projet et avant d'être retenu, il faut que le système vérifie qu'il ne peut être confondu avec un autre projet existant.

Il faut aussi préciser si le projet est ouvert ou non en ce sens que si le projet est dit "ouvert" il donne la possibilité à tous les utilisateurs du programme Diatech d'utiliser ce projet mais non pas de le gérer. Ces utilisateurs pourront avoir accès à la recherche d'informations dans le projet, à l'élaboration de statistiques mais ne pourront pas modifier les données du projet.

Enfin est donnée la possibilité de fournir une description succincte du projet et de ses objectifs.

En bas de l'écran apparaissent les projets des différents utilisateurs, donc les projets créés par les utilisateurs, ceux ouverts et auxquels les utilisateurs peuvent avoir accès, et ceux soumis à autorisation pour accès et utilisation. Une table permet de détailler cette problématique pour rechercher, filtrer et sélectionner les divers programmes, connaître le nom du programme, le nom de son créateur ou propriétaire, les divers types d'autorisation accordés. De plus, pour chaque projet, des raccourcis sont mis en place. Un champ permet de chercher les réponses, un autre d'établir des statistiques, un autre de gérer la base de donnée, un autre de gérer le projet et enfin un autre d'effacer la sélection du programme.

4.3 Gestion du projet

Ce champ est consacré à la gestion du projet. Apparaissent le nom du projet, l'information sur le fait que le projet soit ouvert ou non, et la possibilité de modifier la description du projet. Il est possible d'importer le projet, de gérer les invitations au projet, et d'entrer dans un champ réservé aux commentaires produits par différents utilisateurs du projet.

4.4 Importation de la base de données.

Quand un utilisateur crée un projet, la base de données est vide. Il peut y introduire les données l'une après l'autre dans le programme ou bien les importer à partir de la base de données dans son ensemble si cette dernière est déjà créée.

Chaque projet crée une base de données adaptée à ses données. En conséquence, les diverses bases de données de l'application peuvent être différentes. Aussi pour favoriser leur importation, les fichiers texte au format CSV sont acceptés. Quatre fichiers peuvent être acceptés dans un seul fichier comprimé par ZIP (fig. 3).

Chaque fichier texte est ainsi organisé.

4.4.1 Localisation.csv:

On donne ici la liste des communes relatives aux données. Outre les données propres à chaque lieu d'enquête, on pourra donner leurs coordonnées géographiques. Le fichier comprendra les lignes suivantes:

- **Id**: identificateur de la commune, signalé par un code numérique
- **Localisation**: nom de la commune

- **Latitude**: coordonnée de la latitude de la commune, en chiffre jusque la décimale
- **Longitude**: coordonnée de la longitude de la commune, en chiffre jusque la décimale

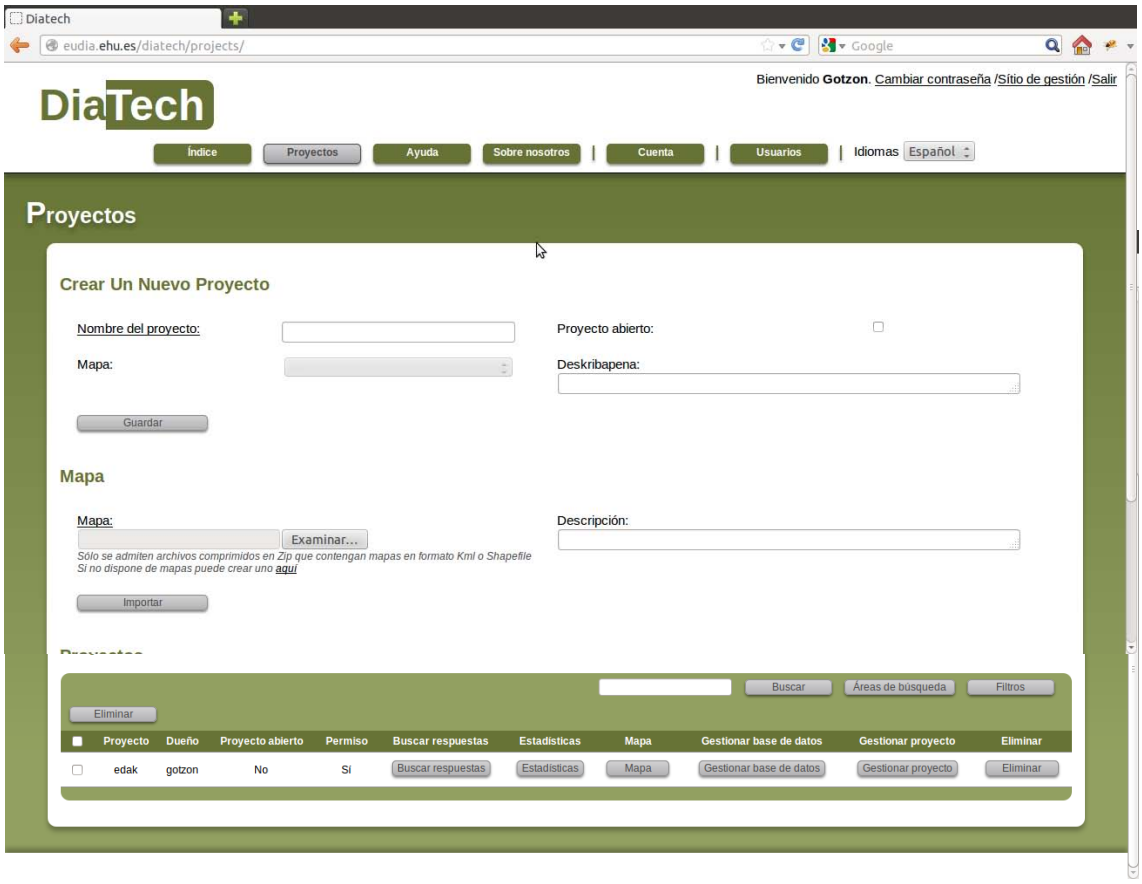


Fig. 3. L'outil *DiaTech:* Gestion de nouvels projects.

4.4.2 Informant.csv (locuteur):

Ici on donne la liste des personnes ayant servi de témoins d'enquêtes, avec l'identification de leur commune. Les lignes suivantes sont à renseigner:

- **Id**: identificateur du locuteur, signalé par un code numérique
- **Nom**: patronyme du locuteur
- **Prénom**: prénom (s) du locuteur
- **Localisation**: identificateur de la commune d'origine du locuteur, en code numérique (qui doit être le même que dans location.csv).
- **Sexe**: sexe du locuteur, signalé en chiffres. 0: ignoré. 1: masculin. 2: féminin.
- **Classe d'âge:** catégorie d'âge, en chiffres (0: non choisi. 1: jeune. 2: adulte. 3: âgé.

4.4.3 Question.csv

Ici on donne les informations relatives aux questions posées au locuteur. Outre les identificateurs et le domaine linguistique, les traductions en chaque langue doivent apparaître. Les lignes suivantes sont donc à construire en fonction de chaque langue et sont:

- **Id**: identificateur de la question, signalé par un code numérique
- **Champ_linguistique**: domaine linguistique choisi, signalé par code numérique. 0: non choisi. 1. phonologie. 2. morphologie nominale. 3. morphologie verbale. 4. syntaxe. 5. lexique.
- **Domaine linguistique**: basque, anglais, castillan... Une ligne est consacrée à la traduction dans la langue concernée. Le titre de la ligne sera le nom de la langue traduit en anglais.

4.4.4 Réponse.csv

Ici on donne les informations relatives aux réponses fournies par les locuteurs. Outre les informations relatives à la réponse, apparaissent aussi les identificateurs du locuteur et de la question. Les lignes suivantes sont à renseigner.

- **Id**: identificateur de la réponse, signalé par code numérique.
- **Question**: identificateur de la question, signalé par code numérique (doit être le même que celui donné dans Réponse.csv).
- **Locuteur**: identificateur du locuteur, signalé par code numérique (doit être le même que celui donné dans Locuteur.csv).
- **Contexte**: contexte de la question
- **Réponse orthographique**: réponse en code orthographique
- **Réponse-phonétique**: réponse en code phonétique
- **Lemme**: lemme de la réponse.

4.4.5 Exportation de la base

Le gestionnaire du projet a la possibilité de faire une copie de sécurité de la base de données et de la déposer sur son ordinateur. L'information sera tenue dans quatre fichiers comprimés par Zip. Les signaux audio et leurs liens avec les réponses ne seront pas sauvegardés sur cette copie.

4.4.6 Localisation.csv:

Ici on donne la liste des communes relatives aux données. Outre les données propres à chaque lieu d'enquête, on pourra donner leurs coordonnées géographiques. Le fichier comprendra les lignes suivantes:

- **Id**: identificateur de la commune, signalé par un code numérique
- **Localisation**: nom de la commune

- **Latitude**: coordonnée de la latitude de la commune, en chiffre jusque la décimale
- **Longitude**: coordonnée de la longitude de la commune, en chiffre jusque la décimale

4.4.7 Informant.csv (locuteur):

Ici on donne la liste des personnes ayant servi de témoins d'enquêtes, avec l'identification de leur commune. Les lignes suivantes sont à renseigner:

- **Id**: identificateur du locuteur, signalé par un code numérique
- **Nom**: patronyme du locuteur
- **Prénom**: prénom (s) du locuteur
- **Localisation**: identificateur de la commune d'origine du locuteur, en code numérique (qui doit être le même que dans location.csv).
- **Sexe**: sexe du locuteur, signalé en chiffres. O: ignoré; 1: masculin. 2: féminin
- **Classe d'âge:** catégorie d'âge, en chiffres (0: non choisi. 1. jeune. 2. adulte 3. âgé.

4.4.8 Question.csv

Ici on donne les informations relatives aux questions posées au locuteur. Outre les identificateurs et le domaine linguistique, les traductions en chaque langue doivent apparaître. Les lignes suivantes sont donc à construire en fonction de chaque langue et sont:

- **Id**: identificateur de la question, signalé par un code numérique
- **Champ_linguistique**: domaine linguistique choisi, signalé par code numérique. 0: non choisi. 1. phonologie. 2. morphologie nominale. 3. morphologie verbale. 4. syntaxe. 5. lexique.
- **Domaine linguistique**: basque, anglais, castillan.... Une ligne est consacrée à la traduction dans la langue concernée. Le titre de la ligne sera le nom de la langue traduit en anglais.

4.4.9 Réponse.csv

Ici on donne les informations relatives aux réponses fournies par les locuteurs. Outre les informations relatives à la réponse, apparaissent aussi les identificateurs du locuteur et de la question. Les lignes suivantes sont à renseigner:

- **Id**: identificateur de la réponse, signalé par code numérique.
- **Question**: identificateur de la question, signalé par code numérique (doit être le même que celui donné dans Réponse.csv).
- **Locuteur**: identificateur du locuteur, signalé par code numérique (doit être le même que celui donné dans Locuteur.csv).
- **Contexte**: contexte de la question
- **Réponse orthographique**: réponse en code orthographique

- **Réponse-phonétique**: réponse en code phonétique
- **Lemme**: lemme de la réponse

4.5 Invitations

Le gestionnaire du projet peut procéder ici à des invitations. La liste des utilisateurs apparaît ainsi que celle des autorisations. L'utilisateur sélectionné peut être invité.

4.6 Commentaires

Un espace est réservé ici aux commentaires que le ou les gestionnaires du projet souhaitent formuler. Quand un commentaire est rédigé, il est ensuite envoyé par mail aux utilisateurs du projet.

4.7 Gestion de la base

On peut ici sélectionner les possibilités de sauvegarder les données, de les actualiser ou de les supprimer. Cet espace est divisé en quatre parties: communes, locuteurs, questions, réponses.

4.8 Communes

On peut ici définir la latitude et longitude de la commune. Une carte permet d'obtenir ces coordonnées.

Sur cette carte apparaissent les communes qui ont déjà été saisies dans la base de données mais on peut y ajouter un symbole pour intégrer de nouvelles communes. En activant ce symbole on peut saisir les coordonnées de la commune correspondante.

Si les coordonnées sont déjà connues, la commune sélectionnée peut être représentée sur la carte en actionnant le bouton. Au-dessous de la carte apparaît la liste des communes à gérer. Des communes peuvent être aussi supprimées.

4.9 Locuteurs

Pour identifier un locuteur nouveau, il faut informer les champs relatifs à la commune, le nom, le prénom, le sexe et la classe d'âge du locuteur puis les sauvegarder.

Sur la table en bas d'écran apparaît la liste des locuteurs de la base. Des locuteurs peuvent aussi être supprimés.

4.10 Questions

Pour identifier une question, il faut informer le domaine linguistique. On le saisit ensuite dans la langue de travail choisie.

Sur la table en bas d'écran apparaît la liste des questions. Des questions peuvent aussi être supprimées.

4.11 Réponses

Avant de saisir une réponse, il est possible d'enregistrer l'enregistrement sonore correspondant. Cet enregistrement peut être un extrait dans lequel est intégrée la réponse. Dans ce cas il faut saisir tout l'enregistrement. A la réponse sont assignés le nom du locuteur, la question, la réponse orthographique, la réponse phonétique, le contexte et le lemme. Si on veut ajouter un extrait de l'enregistrement sonore, il faut choisir parmi la

liste des enregistrements qui ont été saisis celui dont on veut obtenir l'importation. Par le jeu des boutons, on fixe le début et la fin de chaque enregistrement correspondant à la réponse. Ensuite on saisit la réponse. L'enregistrement sélectionné peut être supprimé.

Sur la table en bas d'écran apparaît la liste des réponses de la base. Des réponses peuvent aussi être supprimées.

4.12 Recherche

Ici on peut opérer une recherche parmi les informations de la base. Pour opérer cette recherche on peut jouer sur les éléments suivants: question, réponse orthographique, réponse phonétique, commune, sexe, classe d'âge. Quand on lance la recherche, les réponses apparaissent en bas d'écran.

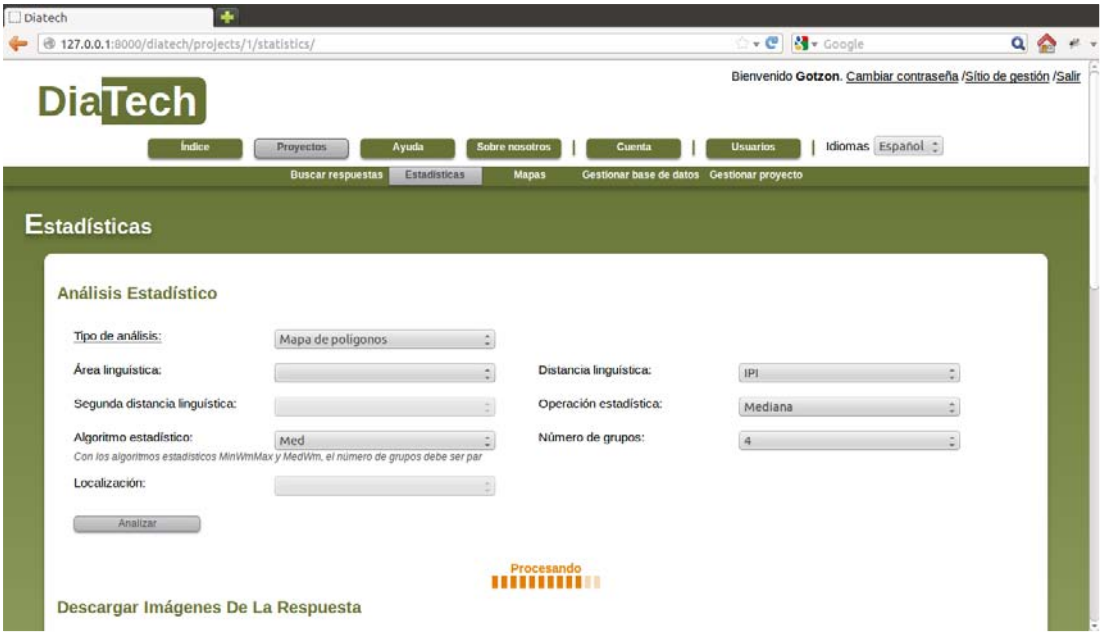


Fig. 4. L'outil DiaTech: statistiques

4.13 Statistique

Ici on peut opérer un travail statistique reposant sur quatre éléments: cartes à polygones, carte à barycentres, carte des limites et arbre. Il faut préciser les choix à opérer pour réaliser ces opérations. Ces choix dépendent pour une part d'entre eux du type d'analyse recherchée, d'autres sont indépendants de ce choix (fig. 4) :

- Champ linguistique: dans toutes les opérations il est possible de choisir un domaine linguistique ou d'analyser toutes les réponses fournies.
- Indice de distance: on peut chercher les distances IPI, IRI, et Levensthein.
- Opérations statistiques: on peut chercher dérive standard, asymétrie, moyenne, maximale, minimale, et corrélation.
- Deuxième champ linguistique: si la corrélation est sollicitée comme opération statistique, un deuxième domaine linguistique peut être activé.
- Algorithmes statistiques: on peut étudier les algorithmes Med, MinWmMax, MedWm.

- Analyse multidimensionnelle: on peut construire un arbre à partir des choix de Ward, Average et Complete.
- Nombre des groupes; on peut choisir de un à dix groupes.

4.14 Compte

Ici chaque utilisateur peut modifier son compte, et donc modifier son nom, son prénom et son e-mail. Le statut de l'utilisateur en relation avec ses autorisations d'accès est indiqué, un utilisateur ne pouvant s'attribuer davantage d'accès que ce qui lui est assigné par les autorisations qu'il a obtenues.

Cet outil fournit une solution au problème des réponses multiples, à ce jour casse-tête de bien d'applications informatiques en dialectologie. Nous pouvons cartographier les cas où plus d'une réponse a été fournie sur un point d'enquête et y apporter un traitement statistique. D'autre part, cet outil permet d'utiliser les bases de données d'autres projets géolinguistiques de telle sorte que n'importe quel chercheur puisse charger ses données dans le système, et puisse cartographier ces données ou en réaliser un traitement dialectométrique qu'il peut ensuite conserver sur son propre ordinateur.

5 L'avenir de la dialectologie

Il est peu concevable que, dans les années qui viennent, la dialectologie puisse se passer du support informatique et n'ait pas recours aux divers outils qui sont mis en place dans le domaine.

Le dialectologue doit nécessairement créer les outils indispensables à ce champ de la connaissance. Ce travail devrait prendre la voie de grandes équipes de telle sorte que nous puissions fournir des outils de travail satisfaisants pour les situations les plus diverses, à partir des outils actuels et en réunissant les forces créatrices capables de les améliorer.

Nous pensons que la phase de groupes réduits peut laisser place à des consortiums plus puissants pour développer des outils de plus grande ampleur et relativement meilleur marché, ce qui n'est pas négligeable en situation économique difficile.

Aussi avons-nous décidé de présenter un projet européen à partir d'un consortium auquel nous invitons tous les chercheurs intéressés à participer autour des points suivants:

a) Création d'un programme informatique sur la base de ceux existant sur le marché.

b) Engagement à importer des données de projets propres dans la base de l'outil à créer.

Nous remercions d'avance tous ceux et celles qui voudront participer à cet évènement.

Bibliographiques

Aurrekoetxea, G. (2011) : "CorpusLem" una herramienta para la conversión de corpus textuales en datos". In M.L. Carrió Pastor & M. A. Candel Mora (eds.), *Las TIC: Presente y futuro en el análisis de Corpus*, Valencia: Universitat Politècnica de València, 611-618. http://www.upv.es/upl/U0547372.pdf (http://alfpro.cc.upv.es:8080/alf

resco/d/d/workspace/SpacesStore/189a8fff-c6da-4c79-bfc1-ad645b17ac38/index.html#/611/zoomed.

Aurrekoetxea, G./Videgain. X. (2004) : *Seme Prodigoaren Parabola Ipar Euskal Herriko 150 Bertsiotan [La Parábola del hijo pródigo en 150 versiones vascas recogidas en el País Vasco-francés]*, ASJUren gehigarriak, EHU, Bilbo (2004) [también en http://klasikoak.armiarma.com/testuak/testuakBourciez001.htm].

Aurrekoetxea, G./Videgain, X. (2009) : "Le projet Bourciez: Traitement géolinguistique d'un corpus dialectal de 1895", in : *Dialectologia* 2, 81-111. (http://www.publicacions.ub.es/revistes/dialectologia2/).

Aurrekotxea, G./Videgain, X./Iglesias, A. (2004): *Bourciez Bildumako Euskal Atlasa (BBEA-1): 1. Lexikoa. [El atlas lingüístico Bourciez: 1. Léxiko]*, ASJU 38:2 [ed. 2007].

Aurrekotxea,G./Videgain, X/Iglesias, A. (2005) : *Bourciez Bildumako Euskal Atlasa (BBEA-2): 2. Gramatika. [El atlas lingüístico Bourciez: 2. Gramática]*, ASJU 39-1 [ed. 2008].

Aurrekoetxea, G. et al. (2012): "DiaTech": A New Tool for Dialectology" (aceptado para su publicación en *Literary and Linguistic Computing*).

Aurrekoetxea, G., Sánchez, J./Odriozola, I. (2009): "EDAK: A Corpus to Analyze Linguistic Variation", in : Cantos Gómez, P./Sánchez Pérez, A. (arg.), 2009, *A Survey on Corpus-based Research Panorama de investigaciones basadas en corpus,* Asociación Española de Lingüística del Corpus, 489-503. (http://www.um.es/lacell/aelinco/contenido/pdf/34.pdf).

DiaTech: http://eudia.ehu.es/diatech/login/?next=/diatech/index/.

EDAK: http://aholab.ehu.es/edak/2/.

GABMAP: http://www.gabmap.nl/.

Goebl, Hans (1992): "L'atlas parlant dans le cadre de l'Atlas linguistique du ladin central et des dialectes limitrophes (ALD) ", in : Aurrekoetxea, G./Videgain, X. (eds.) : *Nazioarteko Dialektologia Biltzarra*, Iker, 7, Bilbao.

Goebl, Hans (2010): "Introdución a los problemas y métodos según los principios de la Escuela Dialectométrica de Salzburgo (con ejemplos sacados del 'Atlante Italo-Svizzero', AIS)", in : Aurrekoetxea, G./Ormaetxea, J. L. (eds.), *Tools for Linguistic Variation*, Bilbao: UPV/EHU, 3-39.

MySQL: http://www.mysql.com/.

Praat: http://www.fon.hum.uva.nl/praat/.

SFSWin: http://www.phon.ucl.ac.uk/resource/sfs/.

TEI: http://www.tei-c.org/release/doc/tei-p5-doc/en/html/TS.html.

Reflexionen und Streiflichter zum Einsatz der EDV beim ALD-II (Ladinienatlas, 2. Teil)

Hans Goebl & Bernhard Schauer & Heidemarie Beer & Agnes Staudinger, Universität Salzburg

1 Bericht des Projektleiters (*Hans Goebl*)

Wiewohl im Folgenden explizit zum 2. Teil des Ladinienatlasses berichtet werden soll, muss dennoch immer auf frühere Phasen des Projektes ALD zurückgegriffen werden, da die hier besprochenen Probleme ohne genauere Kenntnisse des Gesamtprojektes ALD bzw. des Ineinandergreifens der beiden Projektteile nur teilweise verstanden werden können.

1.1 Entstehung und Realisierung des ersten Projektteiles (ALD-I)

Die Idee zu einem mit dem Kürzel ALD (für „Atlante linguistico ladino“) zu bezeichnenden Sprachatlas Nordostoberitaliens entstand im Juli 1972[1] an einem Wirtshaustisch in St. Martin / S. Martin de Tor im Gadertal. Damals traf ich erstmals – und zwar auf explizite Empfehlung durch verschiedene von mir zu sprachlichen Dingen befragte Einheimische – mit Lois Craffonara, dem späteren Gründer und langjährigen Direktor des Istitut ladin „Micurà de Rü“ (IlMdR), zusammen und konnte mit diesem nach einer sehr kurzen Zeit des ersten Beschnupperns eine weitestgehende Übereinstimmung zu diversen Streitpunkten der sattsam bekannten „Questione ladina“ herstellen. Die damals schon sehr detailliert ventilierte Idee war, diesen Sprachatlas nicht nur auf die eigentliche Ladinia zu beschränken, sondern das Untersuchungsgebiet so weit als möglich auf den ganzen Zentralbereich des *Rätoromanischen* (nach Th. Gartner) bzw. der *zona ladina* (nach G. I. Ascoli) auszudehnen und dabei generös das südliche Vorfeld einzubeziehen.

In den anschließenden Jahren (1973-1982) konnte ich im Rahmen einer Assistentenstelle an der Universität Regensburg durch zahlreiche Sondierungsenquêten vor Ort das Thema vertiefen und auch Ideen entwickeln, wie die Daten eines solchen Atlasses zum einen im Feld gesammelt und zum anderen mit den damals immer operationeller werdenden Mitteln der „Computerei“ weiterverarbeitet und publiziert werden könnten. Angedacht war damals bereits aus arbeitspraktischen und ökonomischen Gründen eine Zweiteilung des Projekts (ALD-I: Phonetik und elementare Morphologie; ALD-II: „der Rest“); zudem stand auch schon ab 1978 das später explorierte Gebiet fest, das mit seinen 25.000 km² der Fläche Siziliens entspricht.[2]

Dieser sondierenden Phase folgte nach meiner Berufung an die Universität Salzburg (1982) jene der praktischen Umsetzung. In Zusammenarbeit mit Dieter Kattenbusch, meinem Nachfolger auf der Regensburger Assistentenstelle, und dessen Freund Thomas Stehl[3] konnte das für den ALD-I vorzusehende Fragebuch erarbeitet werden. Lois Craf-

[1] In der „Introductio“ des ALD-I wird auf Seite VII dieses erste Zusammentreffen mit Lois Craffonara auf den Monat *September* des Jahres 1972 gelegt. Das ist nicht richtig. Ich konnte jüngst anhand der damals gemachten Photographien korrigierend feststellen, dass es sich um den Monat *Juli* gehandelt hat.

[2] Über alle Arbeitsfortschritte bei ALD-I und ALD-II wurde regelmäßig und sehr genau berichtet; dies geschah meistens in der Zeitschrift „Ladinia“. Eine vollständige Auflistung der betreffenden Publikationen findet man am Ende der „Introductiones“ von ALD-I und ALD-II und auf den Homepages zu ALD-I und ALD-II.

[3] Thomas Stehl ist seit geraumer Zeit Professor für Romanische Philologie an der Universität Potsdam, Die-

fonara wiederum sorgte als seit 1977 amtierender Leiter des IIMdR für die finanzielle Bedeckung der im Herbst 1985 von Dieter Kattenbusch im Fassatal begonnenen regulären Enquêten. Diese erstreckten sich dann – unter Beteiligung weiterer vier Exploratoren (darunter zwei Herren und zwei Damen) – bis zum Jahr 1992.

Der schon bei der Konzeption des Fragebuchs vorgesehene Einbezug der EDV konnte an der Universität Salzburg dank des Einsatzes zunächst von Roland Bauer und dann von Edgar Haimerl in sehr effizienter Weise realisiert werden. So wurde für die EDV-gestützte Eingabe der Transkripte der 217 Fragebücher bereits eine richtiggehende Erfassungs- und Korrekturschiene in Betrieb genommen, innerhalb derer alle zu verrichtenden Arbeitsschritte Gegenstand genauer Planung und Evaluierung waren. Edgar Haimerl hat dann seit den frühen 1990er Jahren ein exakt dazupassendes *DOS*-Programm namens *CARD* („*Cartography and Retrieval of Dialect Data*“) konzipiert und perfektioniert, womit nicht nur die Ersteingabe und alle nachfolgenden Korrekturen der Transkripte, sondern auch deren Einspeisung in eine relationale Datenbank und die Herstellung der eigentlichen Sprach-Karten sowohl für die Phase der Karten-Ausarbeitung und -Redaktion als auch für jene des eigentlichen Drucks bewerkstelligt werden konnten.

Dazu musste an der Salzburger ALD-Forschungsstelle namens „ALD-Archiv“ eine komplexe Kooperationskette organisiert waren, deren Glieder – bestehend aus philologisch und informatisch arbeitenden Mitarbeitern (*semper utriusque sexus*) –möglichst reibungslos und effizient zusammenarbeiten sollten, ja mussten. Die damit verbundenen fachlich-romanistischen, logistischen, menschlichen und – last but really not least – finanziellen Probleme konnten – Deo maximas gratias! – alle bemeistert werden, so dass der ALD-I im Jahr 1998 unter dem Schirm des Dr. L. Reichert-Verlags bei einer Stuttgarter Druckerei unter Verwendung der von uns gelieferten PDF-Dateien produziert und noch im selben Jahr in den Handel gebracht werden konnte.

Ich habe ganz bewusst auf eine Optimierung der drucktechnischen und buchbinderischen Qualität des Papier-Produkts ALD-I hingearbeitet, da schon damals klar war, dass ein derartiges *opus* die Eigenschaft *aere perennius* haben sollte, um deutlich länger (über)leben zu können als die parallel dazu in Umlauf gesetzten drei CD-ROM, die man leider – wiewohl (ebenso: *leider*) vorhersehbar – *tels quels* heute in keinem der üblichen PC mehr zum Laufen bringen kann.

Diese drei CD-ROM enthielten zunächst alle Daten des ALD-I in drei graphischen Versionen: in der vollen Lautschrift („*ALD-standard*“), in einer simplifizierten Version dazu („*ALD-light*“) und in normalen Buchstaben. Zusätzlich befand sich darauf ein von Edgar Haimerl geschaffenes Suchprogramm namens *IRS* („Index Retrieval System“), das es gestattete, die in *ALD light* gefassten Daten nach vom User einzugebenden Kriterien alternativ vom Anfang oder vom Ende eines Worts her zu klassifizieren. Es war das ein Tribut an die in Salzburg intensiv praktizierte Dialektometrie[4] und die in ihrem Rahmen notwendige Zusammenfassung („Taxierung“) etymologisch identischer, lautlich aber variierender Formen in als linguistisch konvergent gedachten Gruppen (“Taxatareale“).

Auf in den frühen 1990er-Jahren in Salzburg vom Stapel gelassene EDV-technische Basteleien, die in Kooperation mit einer in Essen (Deutschland) angesiedelten Soft-

ter Kattenbusch dasselbe an der Humboldt-Universität Berlin.

4 Die Daten des ALD-I wurden am Beginn des laufenden Jahrtausends von Roland Bauer dialektometrisiert: cf. dazu Bauer 2009. Allerdings ist dabei m. W. diese Möglichkeit nicht oder kaum zum Einsatz gekommen. Überdies befanden sich auf den fraglichen CD-ROM auch animierte Demo-Sequenzen, die die Handhabung von IRS sehr anschaulich vorstellen.

warefirma[5] begonnen und etappenweise perfektioniert wurden, geht eine weitere informatische Komponente zurück, die sich anfangs ebenso auf einer der drei zitierten CD-ROM befand und in späterer Folge auf eine eigene DVD überwechselte: der „Sprechende Sprachatlas", kurz „der Sprechende" genannt.

Die hinter dem Sprechenden stehende Idee war eine reine Linguisten-Geburt: es sollte dem *Ohr* des Fachmanns die Möglichkeit geboten werden, die lautlichen Äquivalente der auf den Karten festgehaltenen Transkriptionen in akustisch optimaler Form zu perzipieren. Und zwar in genau jener isoliert-diskreten Weise, in der sich die auf den Atlas-Karten befindlichen Transkriptionen dem *Auge* des Linguisten darbieten. Dahinter steckte also nicht nur ein *heuristisches*, sondern auch ein *sprachtheoretisches* Programm, das im Übrigen auch der geolinguistischen Gesamtkonzeption des ALD zu Grunde liegt. Dieser diente ab ovo keineswegs zur Erhebung völlig unbekannter oder besonders interessanter Materialien oder gar der Erkundung diastratischer Probleme (wie des Nebeneinanders von Basi-, Meso- und Akrolekten in den Köpfen der Einheimischen), sondern – unter Fortsetzung und Schärfung des von Jules Gilliéron bei der Schaffung des ALF (Datensammlung: 1897-1901, Publikation: 1902-1910) angewandten Prinzips – einer mit sprachlichen Mitteln (hier: exklusiv mittels der Elizitation von basilektalen Responses) vorzunehmenden Vermessung („Geodäsie") eines bestimmten Raumes (hier: der 25 000 km² der ALD-Zone). Man kann daher ruhig zur Charakterisierung dessen, was hier geschehen ist und noch betrieben wird, das Etikett einer „Glotto-Geodäsie" verwenden.[6]

Die angepeilte hohe akustische Qualität konnte nur über Nacherhebungen erreicht werden, die mit einem digitalen (und eben nicht mehr analogen!) Magnetophon[7], einem exzellenten Mikrophon und im Rahmen einer maximal „entspannten" Aufnahmesituation[8] vorgenommen wurden. Angesichts des damit verbundenen hohen Aufwands wurde die dafür herangezogene Zone auf die 21 Messpunkte der brixnerisch-tirolischen Ladinia (= ALD-PP. 81-101) beschränkt.

In formaler Hinsicht entsprach die solcherart erhobene Datenmenge einer zweidimensionalen Matrix aus 21 Messpunkten und rund 900 Antwortreflexen, d. h. also aus rund 19 000 in Salzburg elektronisch aus den Aufnahmegesprächen herauszuschneidenden Segmenten, die nach erfolgtem Schnitt ihrerseits in eine eigene Datenbank eingespeist wurden.

Die von wirklich aufopferungsbereiten Mitarbeitern (Brigitte Rührlinger und Slawomir Sobota) herauspräparierten 19 000 Tonsegmente sind noch immer hochaktuell: im Jahr 2005 wurden sie in eine dem Unternehmen VIVALDI entlehnte Netz-Technologie übernommen, wodurch der heute noch klaglos funktionierende „Netz-Sprechende" entstand. In der zweiten Hälfte des Jahres 2012 hat sich Bernhard Schauer,

---

5 Diese stand und steht unter der Leitung des Trierer Computerlinguisten Reinhard Köhler.

6 Zur sprachlichen Aufdröselung der drei altgriechischen Bestandteile dieses Wortes: *glótta* (etc.) „Sprache", *geo-* „zur Erde gehörig", *dáis* „Teil", *dáiomai* „teilen" (entspricht dem lateinischen Wort *pars*).

7 Damals handelte es sich um einen *DAT*-Rekorder. *DAT: Digital Audio Tape.*

8 Diese Aufnahmesituation war damit nicht das, was die traditionelle Dialektologie unter „natürlich" versteht. Für mich als einen in den Dimensionen von Quantität und Messung denkenden Sprachgeographen hat das Konzept der Natürlichkeit keine wie immer geartete Relevanz. Keine Messung ist „natürlich", wo immer auch gemessen wird. Überall wird anhand theoretisch vorfixierter Konzepte und auch unter „Molestierung" der zu vermessenden Objekte beurteilt. Im vorliegenden Fall ist aus methodischer Hinsicht noch das Prinzip der „Interkomparabilität" sehr wichtig. Die an den 217 Messpunkten gesammelten Daten müssen untereinander voll vergleichbar sein und damit über denselben messtheoretischen Status verfügen. Nur dadurch ist es möglich, bei der Auswertung der Sprachatlas-Daten (geolinguistische) Aussagen zu tätigen, die für alle Teile des Gesamtgebiets gleichermaßen relevant sind.

der Chef-Informatiker des Projektes ALD-II, mit einem eigens dazu angeheuerten Ferialpraktikanten dieser Daten mit dem Ziel angenommen, einen neuen netz- und DVD-basierten „Sprechenden“ zu kreieren: der heute dazu schon vorliegende Prototyp ist sehr vielversprechend und wird bald in seine definitive Form übergehen.

Der „Sprechende“ ist seit 1998/1999 mit seinem gerade für Ladiner sehr attraktiven Inhalt und seinen ausgeklügelten Funktionen immer wieder in Unterricht, im Museumsbetrieb und in der Forschung zur Anwendung gekommen.

In toto dauerten die Arbeiten zum ALD-I also rund 13 Jahre (1985 bis 1998), innerhalb derer ein organismusähnliches System bis zur Erreichung des gesteckten Zieles unter Dauerspannung gehalten werden musste.

1.2 Entstehung und Realisierung des zweiten Projektteiles (ALD-II)

Im Jahr 1998 standen mir als Projektleiter angesichts der herrschenden Gesetzeslage, die meine Emeritierung ab dem 1.10.2012 vorsah, noch rund 14 Jahre an verplanbarer Arbeitszeit zur Verfügung. Daher musste hinsichtlich der ein zweites Mal zu durchlaufenden Arbeitsetappen (Erstellung des Fragebuchs, Probe-Enquêten, reguläre Feldarbeit, Aufbau einer neuen EDV-Schiene, EDV-gestützte Erfassung der gesammelten Daten etc.) ohne Zeitverlust gehandelt werden.

Dabei kam es, was auch angesichts der finanziellen Rahmenbedingungen nicht verwunderlich ist, zur Bildung eines personell weitestgehend neu zusammengesetzten Teams, in dem nur wenige Köpfe schon beim ALD-I mitgearbeitet hatten: zu diesen zählten neben Edgar Haimerl (der durch seinen erneuten Einsatz einen im Jahre 2001 mit einer externen Programmierfirma erlittenen EDV-technischen Schiffbruch reparierte) nur noch die beiden (steirischen) Exploratoren Helga Böhmer und Tino Szekely. Ähnlich wie beim ALD-I kamen die anderen Mitarbeiter aus verschiedenen Ecken Österreichs (vornehmlich aus Salzburg), Deutschlands und Italiens.

Bei der Erstellung des Fragebuchs und den damit in Ladinien durchgeführten Explorationen tat sich ganz besonders der aus Enneberg (= ALD-P. 81) stammende und später in Innsbruck und Eichstätt zum Romanisten ausgebildete Linguist Paul Videsott[9] hervor. Unter den Exploratoren nahm ganz unzweifelhaft die in weiterer Folge mehrfach zitierte Brigitte Rührlinger auch deshalb eine Sonderstellung ein, weil sie nach der Exploration von 32 Ortschaften noch zahlreiche Sonderaufträge für den ALD-II übernommen und bravourös erledigt hat, die sich von Nachforschungen auf Friedhöfen, über Nachenquêten bis hin zu höchste Präzision erfordernden Tagging-Arbeiten an den Sound-Daten des ALD-I erstreckten: auch dazu wird man in der Folge (v. a. im Abschnitt 3.) Näheres lesen können. Br. Rührlinger wurde für die Belange des ALD im Rahmen des ersten von mir zu diesem Thema an der Universität Salzburg gehaltenen Seminars „entflammt“ und hat dem Gesamtprojekt bis heute unverbrüchlich die Treue gehalten.

Während beim ALD-I die insgesamt 217 Enquêten in den Händen von fünf Exploratoren lagen, wurde die analoge Arbeit mit dem nunmehr 1 063 z. T. recht komplexe Fragen enthaltenden Questionnaire von insgesamt zehn Exploratoren (davon sechs Männer und vier Frauen) erledigt. Die Regel-Enquêten fanden zwischen 2001 und 2007 statt und umfassten somit sieben Jahre.

Insgesamt wurden dabei 833 Gewährspersonen befragt (und deren Äußerungen akustisch dokumentiert), denen beim ALD-I nur 488 Informanten gegenüberstanden.

[9] Paul Videsott ist derzeit Professor für Romanische Philologie an der Universität Bozen (mit Sitz in Brixen).

Der Inhalt des neuen Fragebuchs betrifft die linguistischen Kategorien *(elaborierte) Morphologie, Syntax* und *Lexikon* und versteht sich als Supplement zu jenem des ALD-I. Klarerweise sind in dessen Konzeption und auch in die layout-technische Umsetzung die beim ALD-I gemachten Erfahrungen eingeflossen, wobei diese erneut auf den optimierten Einsatz der EDV bei der Erfassung und Weiterarbeitung der Daten abzielten.

Aus einsichtigen Gründen mussten beim ALD-II dasselbe Netz und dasselbe Lautschriftsystem zum Einsatz kommen. Weitere *Ähnlichkeiten* zum ALD-I: ethnophotographische Dokumentation jeder Ortschaft; in den ersten Jahren der Feldarbeit: gemeinsame Schulungsseminare („*Seminario di trascrizione*"-*SETRA*) für die Exploratoren.

Projekt- und zeitspezifische *Unterschiede* zum ALD-I waren: nur einmalige Abfrage des Questionnaires pro Ortschaft; durchgehender Einsatz von digitalen (und damit eben nicht mehr von analogen) Aufnahmegeräten.

Angesichts der um die Jahrtausendwende vollzogenen Ablösung des Betriebssystems *MS-DOS* durch *Microsoft Windows* war klar, dass für alle einschlägigen EDV-Belange des Projektes ALD-II in *Ersetzung* (und nicht bloß: *Reparatur*) des veralteten ALD-I-Programms *CARD* eine neue Generalsoftware zu erstellen war. Dies geschah ab etwa 2002 durch das von Edgar Haimerl konzipierte und lauffähig gemachte Programm *DMG* („*Dialect Map Generator*").

Die Funktionalitäten dieses Programms übertrafen natürlich jene von *CARD*: dies betraf ganz besonders den Bereich der Produktion der Karten, die angesichts der großen optischen Bedeutung, inhaltlichen Komplexität und vor allem Anzahl derselben in größtmöglichem Umfang automatisiert werden sollte. Dies hatte zur Voraussetzung, dass Programm-Module geschaffen werden mussten, die imstande sein sollten, die überschneidungsfreie Verteilung von Transkriptionen (samt allen Sub- und Superskripten) auf den beiden Kartenhälften des ALD-Netzes fehlerfrei vorzunehmen. Dieses Problem wurde von insgesamt vier Informatikern (Edgar Haimerl, Fabio Tosques, Andreas Wagner und Bernhard Schauer) bearbeitet und schlussendlich in wirklich überzeugender Weise gelöst.

Nur nebenbei: von den vier genannten Personen waren bzw. sind zwei (Andreas Wagner, Bernhard Schauer) ausschließlich informatisch tätig, während Edgar Haimerl und Fabio Tosques von ihrer Ausbildung her interdisziplinäre Doppelnaturen darstellen. Abgesehen von der Informatik hat sich Edgar Haimerl mit Skandinavistik, Mathematik und Philosophie befasst, während Fabio Tosques die Informatik mit der Romanistik kombiniert hat.

Wie schon erwähnt, sind beim ALD-II die Enquêten im Jahr 2007 zu Ende gegangen. Schon zwei Jahre darnach existierte eine als „fehlerfrei" zu betrachtende Datenbank mit dem gesamten transkribierten Feldertrag, so dass die Produktion der Probe-Karten und damit die erste Stufe von deren philologisch-linguistischer Bearbeitung anlaufen konnte.

B. Schauer, H. Beer und A. Staudinger beschreiben die Details dieser Arbeit in den Kapiteln 2.-4. In toto waren an dieser zeitlich und technisch ungemein anspruchsvollen Arbeit im Wesentlichen nur vier Personen beteiligt: die drei zuvor Genannten und meine Wenigkeit. Insgesamt gab es drei volle und eine halbe Korrekturphase, wozu jeweils der gesamte Besatz an Karten und den dazugehörenden (sehr nützlichen) Liste ausgedruckt werden musste. Agnes Staudinger gibt dazu in Kapitel 4. numerische Hinweise.

Von unschätzbarem Wert war die ab 2009 mögliche Einbindung einer Sound-Datenbank in die Redaktionsarbeit. So war es möglich, jeden irgendwo an der Korrektheit einer Transkription aufgekommenen Zweifel durch sofortiges (und beliebig ausdehnbares) Hineinhören in die fragliche Stelle auszuräumen.

Da die Konzeption des ALD-II nicht nur die Produktion von *Karten*-Bänden, sondern auch jene von zwei kleinformatigen *Zusatz*-Bänden (Supplement-Band [„Volumen supplementarium"] und Index-Band [„Index generalis"]) vorsah und auch dafür möglichst elegant und zugleich arbeitsökonomisch zu produzierende Druckvorlagen erstellt werden sollten, hat Bernhard Schauer dazu ein weiteres Spezial-Programm („SuBIReS") kreiert.

Von der für alle diese Zwecke und Programme geschaffenen EDV-Architektur bzw. Server-Landschaft berichtet B. Schauer im nachfolgenden Kapitel.

Für den ALD-II wurde angesichts der seit der Zeit des ALD-I erfolgten EDV-technischen Fortschritte auf die Produktion eines auf Papier zu publizierenden vor- und rückwärts alphabetischen Index verzichtet und stattdessen einem netzbasierten „*Index Retrieval System*" (*IRS*) neuer Art der Vorzug gegeben. Dieses kann nicht nur – erneut auf der Grundlage von *ALD-light* und den *Normalbuchstaben* – die erwähnten alphabetischen Indizes erstellen, sondern auch nach bestimmten positionellen Vorgaben (betreffend Anfang, Mitte und Ende eines Wortes) feststellen, ob eine bestimmte Zeichensequenz im Gesamtdatenbestand vorkommt. Derzeit gibt es diese sehr schnelle Suchmaschine für beide Teile des ALD. Über das *IRS* des ALD-II kann man auch an die PDFs ganzer Karten und aller dazugehörenden Listen herankommen.

Unter Ausnützung von für unser Projekt günstigen Neuerungen (Stichwort: *open access*) bei der Politik der Druckförderung vonseiten der österreichischen Forschungsförderungsorganisation FWF war es möglich, alle Etappen des eigentlichen Druckvorgangs in der Hand zu behalten und damit die anfallenden Kosten zu minimieren. Dabei kam es zu einer technisch recht komplexen, prozedural und menschlich gleichwohl aber stets sehr harmonischen Kooperation mit der Trentiner Druckerei Alcione (gelegen in Lavis, auf halbem Weg zwischen den ALD-Messpunkten 66 [S. Michele all'Adige] und 121 [Trient]).

Drucker und Herausgeber (also meine Wenigkeit) konnten dabei – vor allem bei der Reparatur von gar nicht wenigen, während der Druckvorbereitung entdeckten Fehlern – von der ungemeinen Effizienz der unter der Obhut von B. Schauer stehenden (und entstandenen) EDV-Maschinerie profitieren.

Dieser Effizienz, die auch auf die nicht-informatischen Mitarbeiter(innen) übergesprungen ist, war es schlussendlich zu verdanken, dass die zwei für den ALD-II geschaffenen zentralen Tools – *SDB* und *IRS* – auch für die Daten des ALD-I geöffnet worden sind, so dass diese – abgesehen von ihrer ungeschmälerten Fortexistenz im (und als) Druckwerk – qua EDV einen neuen Impuls erhalten haben.[10]

1.3 Zukunftsperspektiven

Vom Druckwerk ALD-II wurden 330 Exemplare produziert und davon 90% für den Verkauf freigegeben. Dieser erfolgt im Rahmen des Non-Profit-Verlages „Editions de Linguistique et de Philologie" (mit Sitz in Straßburg und Paris) und hat nach zwei von Salzburg aus gestarteten Werbe-Campagnen (via Post und via Mail) im Dezember 2012 begonnen.

[10] Die Gesamtmenge der von ALD-I und ALD-II der Fachwelt zur Verfügung gestellten Sprachkarten (SK) beträgt 1950 Einheiten. Innerhalb jener Kartenmengen, die man in den anderen romanischen Regionalatlanten findet, ist das ein absoluter Spitzenwert. Nur der korsische Atlas ALEIC (publiziert: 1933-1942, 9 Karten-Bände) und der gaskognische Atlas ALG (publiziert: 1954-1974, 4 Karten-Bände) haben mit 2001 bzw. 2531 SK einen größeren Kartenbestand. Wenn man aber das Produkt aus der Anzahl der Messpunkte (MP) und jener der Karten (SK) bildet, dann liegt der ALD als Ganzes mit dem Wert 423 150 an der absoluten Spitze (ALEIC: 49 MP mal 2001 SK = 98 049; ALG: (174 MP + 155 MP)/2 mal 2531 SK= 416 349,5). Ist es angesichts dieser Fakten bloß ein historischer Zufall, dass der Autor des datenmäßig sehr umfangreichen ALG – Jean Séguy (1914-1973) – zum Begründer der *dialectométrie* geworden ist?

Dabei wird die eigentliche Verkaufsarbeit durch einen in Baden-Württemberg angesiedelten Distributor geleistet, der nicht nur die Lagerhaltung und den Versand, sondern auch die Entgegennahme der Bestellungen und das Inkasso der ausgestellten Rechnungen besorgt.

Durch die schon erwähnte Akkumulierung günstiger Umstände – wozu auch die hohe Performance der ALD-Informatik zählt – war es möglich, einen konkurrenzlos niedrigen Verkaufspreis (von nur 200 Euro!!!) für die sieben in exzellenter Qualität produzierten Bände des ALD-II festzusetzen. Die Zukunft des papierenen ALD-II scheint damit für lange Zeit gesichert zu sein.

Doch wie schaut es mit den netzbasierten Tools[11] aus, über die man über unsere zwei neuen Homepages[12] herankommt? Da mit 31.12.2012 die ALD-Mannschaft in alle Winde zerstoben ist und bedauerlicherweise für das Jahr 2013 bereits zugesagte Fördermittel für die Aufarbeitung noch unerledigter Projekt-Arbeiten und die fortdauernde Pflege der EDV-Struktur wieder zurückgezogen worden sind, ist auf dem EDV-Sektor die Zukunft des ALD-II deutlicher weniger rosig als auf jenem des Papiers.

Ich werde versuchen, trotz weitestgehender Austrocknung der (zwischen 1999 und 2012 in sehr großzügiger Weise geflossenen) finanziellen Mittel von Fall zu Fall die sich als nötig erweisenden Sanierungen zu organisieren. Mit sehr großer Dankbarkeit habe ich die diesbezüglich geäußerte Hilfsbereitschaft der Mitautoren dieses Berichts zur Kenntnis genommen.

Immerhin geht es hier um so etwas wie die „Pflege des eigenen Kindes". Und zwar eines rundum sehr gut geratenen Kindes...

Die über das Ende von 2012 noch offen verbliebenen Agenden sind:

- Inhaltliche und sprachliche Komplettierung der beiden neuen Homepages;
- Einrichtung der Bilddatenbank und Tagging der 17 000 Dias;
- Fertigstellung der neuen Version des „Sprechenden" zum ALD-I;
- Fortsetzung und Vollendung des Fein-Taggings der Daten der Sound-Datenbank des ALD-II.

Alle diese Agenden könnten bei Vorhandensein entsprechender Fördermittel von den ehemaligen Mitarbeiterinnen des ALD-II „nebenbei" erledigt werden.

Noch ein Wort zu den derzeit der geolinguistischen Fachwelt zur Verfügung gestellten Funktionalitäten von ALD-I und ALD-II:

Papierversionen:

Traditionelle Auswertung von 1950 Volltext-Karten (mit feiner Transkription) durch direkte Konsultation mit (oder ohne) parallele(r) Anwendung von „stummen Karten" (die über die beiden neuen Homepages verfügbar sind)

[11] Überdies findet man die PDFs des Inhalts aller sieben Bände des ALD-I in der E-Book-Library „Phaidra" des FWF unter dem folgenden Link: https://e-book.fwf.ac.at/search_object. Man kann sich dort in der Tat – das Vorhandensein eines guten Druckers (für Farbe und das Format A2) und von sehr viel Geduld vorausgesetzt – die sieben Bände des ALD-II völlig „frei" herunterladen und ausdrucken. Auch hier bleibt offen, wie lange dies möglich sein wird. Doch ist davon auszugehen, dass der FWF als öffentliche Institution alles daransetzen wird, um die EDV-technische Lesbarkeit des Inhalts seiner E-Book-Library auf lange Sicht sicherzustellen.

[12] ALD-I (neu): http://ald1.sbg.ac.at/; ALD-I (alt): http://ald.sbg.ac.at/ald/; ALD-II: http://ald2.sbg.ac.at/.

Elektronische Tools:
SDB: akustische Nachkontrolle aller Transkriptionen, sowohl punktgenau (via Eingabe von Messpunkt x [1-217] und Frage y [1-1063]) als auch durch Anhörung stundenlanger Aufnahmegespräche
IRS:
Zugriff auf die transkriptorisch auf *ALD-light* reduzierten Daten nach den folgenden Kriterien:

- vorwärts alphabetisch;
- rückwärts alphabetisch;
- freie Suche (via Definition eines frei [in *ALD-light* oder in Normalbuchstaben] definierbaren Such-Nexus) nach der Existenz eines Such-Nexus in den Gesamtdaten von ALD-I oder ALD-II in den drei folgenden Positionen eines „Worts": Anfang, Mitte, Ende.

Beim ALD-II kann die zuletzt genannte Prozedur bis zur Auffindung (und zum nachfolgenden Ausdruck) der betreffenden Sprach-Karten und der dazugehörenden Listen verlängert werden. Damit können genuin philologisch-linguistisch orientierte Interessen von Sprachgeographen optimal zufriedengestellt werden.

Anders als beim ALD-I[13] wird aber beim ALD-II ein Bereich nicht bedient: jener von datenanalytisch orientierten Computerlinguisten, die daran Interesse hätten, direkt auf die Transkriptionen (ob nun in *ALD-standard*, *ALD-light* oder in Normalbuchstaben) zuzugreifen. Diese Interessen können derzeit nicht durch die Konsultation der beiden Homepages bedient werden. Wer mit den Transkriptionen in diesem Sinn „rechnen" möchte, muss sich an mich oder Bernhard Schauer wenden, damit er solcherart Zugriff auf unsere einschlägige Datenbanken erhält.

2 Zum EDV-Einsatz beim ALD (*Bernhard Schauer*)

2.1 Zum EDV-Einsatz beim ALD-I

Zur Eingabe und Korrektur von Transkriptionen sowie zum Druck der Sprachkarten wurde in der Mitte der 1990er Jahre von Edgar Haimerl eine Software namens CARD entwickelt und erfolgreich eingesetzt. In weiterer Folge wurde dieses DOS-Programm durch das Programm *IRS* („Index Retrieval System") ergänzt, das das Durchsuchen der Transkripte in einer vereinfachten Lautschrift („ALD light") erlaubte, die entsprechenden Vorkommen auf den Karten anzeigte und auch zu Zwecken der Klassifikation (samt nachfolgender Visualisierung der Resultate) der Daten einzelner Karten des ALD-I verwendet werden konnte.

Um dem ALD-I auch eine hörbare Komponente beizugeben, wurde gegen Ende des letzten Jahrtausends der „Sprechende Sprachatlas" (kurz: der „Sprechende") zuerst als CD-, dann als DVD- und schließlich – mit einer dem Projekt VIVALDI entlehnten Technologie – als Internet-Version entwickelt. Zu diesem Zweck wurden in Ladinien nach den Regel-Explorationen zusätzlich Schälle mit besonders hoher Qualität („Edelsounds") – also unter idealen Bedingungen aufgenommene Ton-Daten – erhoben und der Allge-

[13] Beim ALD-I war der skizzierte Zugriff auf die Gesamtdaten über die erwähnten drei CD-ROM möglich. Diese Möglichkeit wurde von mir im Jahr 2008 in Kooperation mit Thomas Zastrow (Universität Tübingen) benützt, um eine dialektometrische Verrechnung der Daten des ersten Bandes des ALD-I mittels der Levenshtein-Distanz vorzunehmen.

meinheit über den „Sprechenden“ zur Verfügung gestellt. Datenseitig handelte es sich dabei nicht um die 217 Messpunkte des Gesamtnetzes, sondern nur um die 21 Messpunkte Ladiniens (ALD-PP. 81-101).

Beim ALD-II wurde im Jahr 2010 mit zwei Digitalisierungs-Campagnen begonnen: diese betrafen zuerst die während Regel-Enquêten bereits mit digitalen Aufnahmegeräten erstellten Ton-Aufnahmen und ab 2012 die bei der Feldarbeiten aufgenommenen Dias („Ethnophotographie“). Weiters wurde im Sommer 2012 die Software für den DVD-Sprechenden des ALD-I neu entwickelt, da ab Windows 7 64bit das auf die Zeit vor der Jahrtausendwende zurückgehende, alte Programm nicht mehr funktionstüchtig war.

2.2 Zum EDV-Einsatz beim ALD-II

Die Erfassung und Korrektur der Transkriptionen sowie der Druck der Sprachkarten beim ALD-II erfolgte über das Softwarepaket *DMG* („Dialect Map Generator“). Auch beim ALD-II wurde dieses Programm durch ein *IRS* („Index Retrieval System“) flankiert, das aber in diesem Falle über das Internet funktioniert und das gezielte Durchsuchen des Datenbestandes anhand der schon erwähnten reduzierten Transkription („ALD light“) erlaubt. In diesem Fall geht es also nicht mehr um das Klassifizieren der Daten des ALD-II.

Zusätzlich entstand beim ALD-II eine eigene Software („SoundDatenBank“ oder „Sound-Datenbank“, *SDB*) zur Verwaltung der im Feld während der Exploration erhobenen Ton-Daten. Die *SDB* fand unter anderem schon bei der Korrektur und Redaktion der zu veröffentlichenden Daten Anwendung; dabei war es wesentlich, dass Ton-Daten auf ganz kurzem Weg aufgefunden und abgehört werden konnten.

Eine weitere Zielsetzung beim ALD-II war, die rund 17 000 Dias von ALD-I & II nach deren Digitalisierung über das Internet allgemein verfügbar zu machen. Die dazu entstehende *BDB* („BildDatenBank“) konnte allerdings bis Projektende (31.12.2012) leider nicht zur Gänze fertig gestellt werden.

3 Details zu den Strukturen der beim ALD-II verwendeten EDV

In diesem Kapitel soll ein vertiefter Einblick in die technische Struktur des ALD-II und der einzelnen Programme vermittelt werden. In einem weiteren Punkt soll auf die jeweiligen Erkenntnisse aus der Programmerstellung eingegangen werden. Um hier detailgenau informieren zu können, ist es erforderlich, das einschlägige Fachvokabular zu verwenden. Es wird aber versucht, alle verwendeten Abkürzungen und Begriffe in Fußnoten zu erklären und auch auf weiterführende Literatur hinzuweisen.

3.1 Infrastruktur

Um die technische Infrastruktur funktional zu halten und dabei möglichst hardwareunabhängig zu bleiben, wurde gemeinsam mit dem Fachbereich Romanistik der Universität Salzburg ein physikalischer Server angeschafft, auf dem die Virtualisierungslösung *VMWare ESXi*[14] (Version 4) eingesetzt wird: siehe dazu die Visualisierung in Abbildung 1. Die in der Wolke zusammengefassten Server laufen dabei innerhalb der Virtualisierungslösung parallel.

[14] Aktueller Produktname, ab Version 5: *VMWare VSphere Hypervisor*.

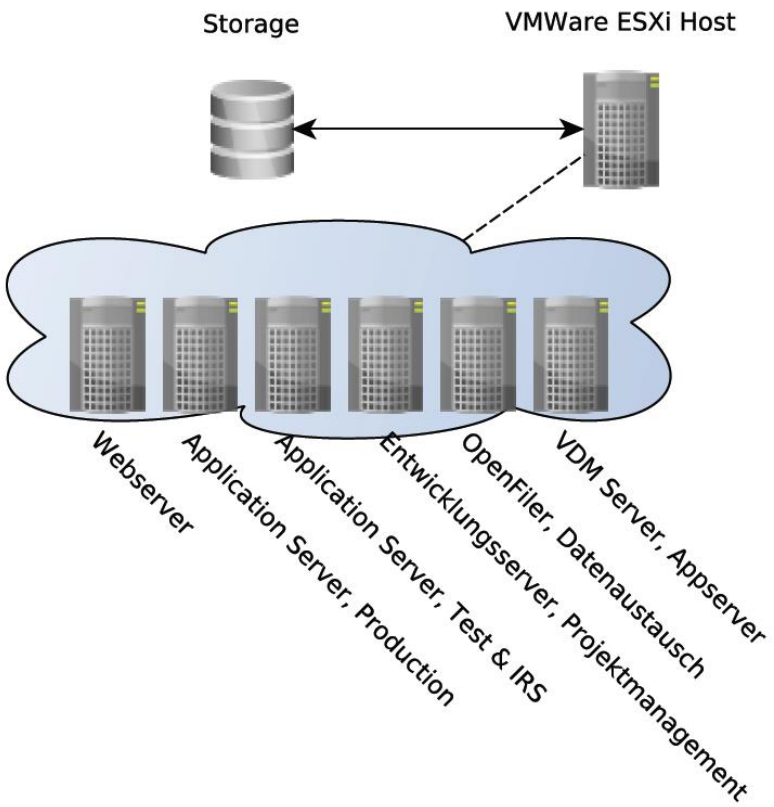


Abbildung 1: Überblick über die EDV-Infrastruktur des ALD-II sowie über die virtuellen Server

Ein großer Vorteil der Virtualisierung der Server besteht darin, dass diese einfach auf andere *VMWare ESXi* Server umgesiedelt werden können, und auch, dass die Universität Salzburg eine derartige Infrastruktur aufgebaut hat. Alternativ dazu ist es möglich, die virtuellen Server in die „Cloud" zu übersiedeln, beispielsweise zu *Amazon EC2* oder ähnlichen Angeboten. In unserem Fall kommt dabei eine sogenannte „*private Cloud*" zum Einsatz: siehe Mell et al. 2011.

Die digitalen Daten des ALD-I und II werden von den ITServices der Universität Salzburg gehostet und intern über *Windows Shares*[15] bzw. *iSCSI*[16] zur Verfügung gestellt. Im technischen Sinn handelt es sich dabei um ein *Network Storage*, das Daten sowohl als *SAN*[17] bzw. *NAS*[18] zur Verfügung stellen kann.

Die Aufteilung auf die einzelnen virtuellen Server[19] basiert auf speziellen Funktionalitäten, wobei versucht wurde, deren Dienste anhand der gestellten Anforderungen zu separieren.

[15] Darunter wird hier die Anbindung über *CIFS* verstanden. Die Daten sind dabei über die „Netzwerkumgebung" unter Windows bzw. vergleichbare Dienste unter *MacOS* bzw. *Linux* zu erreichen.

[16] Es ist das der Name des Protokolls, um Festplatten „blockweise" über ein Netzwerk ansprechen zu können. D.h. es handelt sich um keinen dateibasierten Zugriff.

[17] „*Storage Attached Network*", ist der Oberbegriff für Systeme, die Daten *block*-basiert zur Verfügung stellen (beispielsweise über *iSCSI*).

[18] „*Network Attached Storage*", ist der Oberbegriff für Systeme, die Daten *datei*-basiert zur Verfügung stellen (beispielsweise über *CIFS* bzw. *NFS*).

[19] Ab hier wird in diesem Text mit dem Begriff *Server* immer ein „virtueller Server" bezeichnet.

Der Webserver ist der einzige Server der IT Infrastruktur der Universität Salzburg, der von außen erreichbar ist; er stellt unter anderem die Projekt-Homepages (zu: ALD-I, ALD-II, ALD-I Web-Sprechender, Dialektometrie) zur Verfügung. Darüber hinaus wird er auch als „*Reverse Proxy*[20]“ für die dahinter liegenden Server verwendet.

Weiters stehen dem Projekt zwei Anwendungsserver zur Verfügung, die einerseits die Produktivumgebung für den laufenden Betrieb und andererseits eine umfassende Testumgebung bereitstellen.

Ein weiterer Server dient dem Projektmanagement, der Quellcodeverwaltung sowie als *Build/CI*-Umgebung: er wird als *Entwicklungsserver* bezeichnet.

Um die Tondateien und die Daten der digitalisierten Dias gleichzeitig mehreren Servern zur Verfügung stellen zu können, wurde eine *OpenFiler*[21]-Instanz installiert, die den Zugriff darauf über *NFS* gewährleistet.

Da der Zugriff auf die Daten nur über einen einzelnen Punkt erfolgt, lässt sich dieser besser absichern. Zudem ist es auch einfacher, im Hintergrund Server auszutauschen, da diese nur *intern* und nicht *weltweit* bekannt sein müssen. Zusätzlich nimmt diese Aufteilung auch Rechenlast vom Webserver, da diese Zugriffe nur weitergeleitet werden. Aufgrund der Virtualisierung wäre es hier möglich, den/die *Anwendungsserver* auf physikalisch andere Server umzusiedeln, um mehr Rechenleistung zur Verfügung zu haben.

3.2 Softwarearchitektur

Die für den ALD-II entwickelte Software basiert grundsätzlich auf einer *Client-Server*-Architektur mit *Datenbank-Backend*, es handelt sich also um eine *3-Tier*-Architektur (*Client, Anwendungsserver, Datenbank*): siehe dazu die Abbildung 2. Hier erkennt man auch den Unterschied zwischen internen und externen Zugriffen auf die Struktur. Serverseitig wird beim Forschungsprojekt ALD-II ein *JBoss AS*[22] bzw. für die Webservices ein *Apache Tomcat Servlet Container*[23] verwendet.

[20] Dabei wird der interne Server nur über den Webserver indirekt angesprochen. Abfragen könnten beispielsweise damit auch am Webserver zwischengespeichert werden.

[21] Das ist eine Softwarelösung, die als lokales *NAS* für die Server operiert. Die Datenspeicherung erfolgt in unserem Fall über das *SAN*.

[22] *JBoss AS*: *Java EE Application Server*.

[23] Es handelt sich dabei – vereinfacht ausgedrückt – um ein Produkt, um Java Web-Anwendungen in einer „leichtgewichtigeren Umgebung“ betreiben zu können, als dies ein *Java EE Server* wäre.

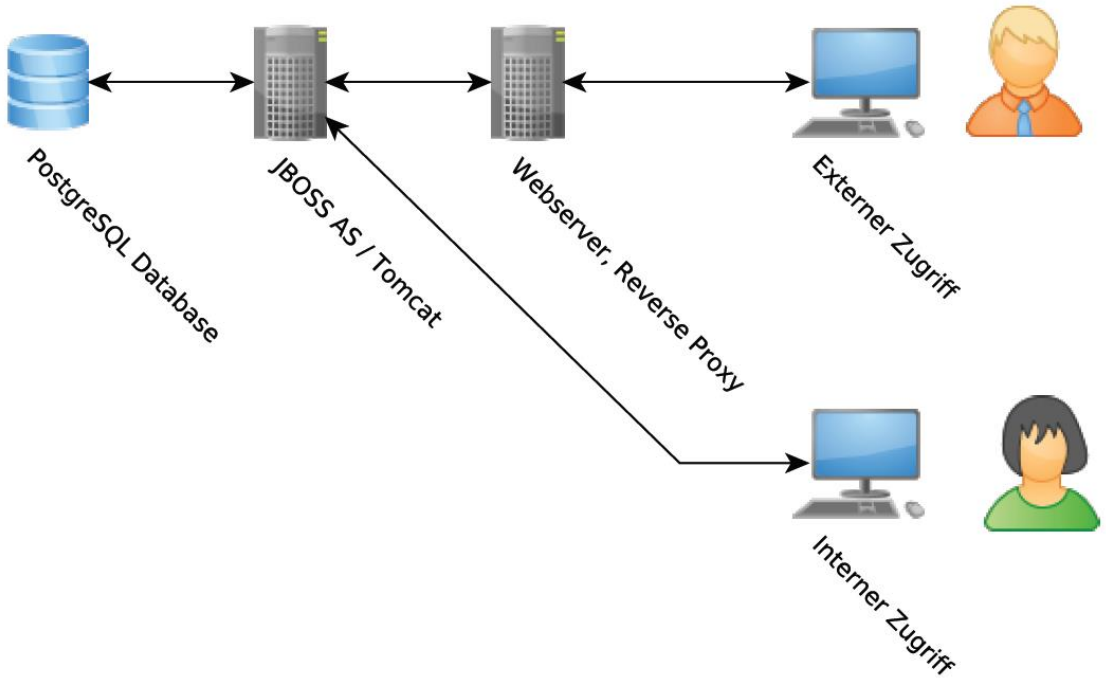


Abbildung 2: Softwarearchitektur mit Zugriffsvarianten

Die eigentlichen *Daten* werden in der dahinter liegenden Datenbank abgelegt, die unter anderem auch deren referentielle Integrität sicherstellt.

3.3 Verwendung von Software, Tools & Bibliotheken

Bei der Auswahl der zu verwendenden Software wurde auf mehrere Gesichtspunkte Rücksicht genommen: primär wurde, soweit dies möglich war, auf *Open-Source*-Lösungen abgezielt (*Apache Tomcat, JBoss AS, PostgreSQL, Apache Webserver*, usw.). Wenn dies nicht möglich oder auch nicht genug effizient war, wurde die Palette um kostenfrei verwendbare Produkte erweitert, was aber nur für die Virtualisierungslösung notwendig war. *VMWare ESXi* lizenziert die Serversoftware alleine kostenfrei; das Monitoring bzw. die Verwaltung mehrerer Server usw. müssten bei Bedarf kostenpflichtig nachgerüstet werden.

Ein weiterer Punkt, der bei der Softwareauswahl zu berücksichtigen war, sah vor, dass nach Möglichkeit versucht wurde, immer weit verbreitete Produkte zu verwenden, da davon eine größere Langlebigkeit erwartet werden konnte. Dazu kommt, dass weit verbreitete (Open Source) Produkte in aller Regel besser gepflegt werden, da quantitativ ein größeres Interesse am Funktionieren dieser Software besteht.

3.4 Kurzvorstellung der einzelnen Programme

3.4.1 DMG – Dialect Map Generator

Die „Hauptsoftware“ des Projektes erlaubt die Erfassung und Korrektur von Transkriptionen, sowie den Druck der Sprachkarten. Weiters wurden damit die Daten zu den Informanten und den explorierten Orten erfasst.

Die Entwicklung dieser Software sollte ab 2001 von einer externen Firma durchgeführt werden, was zu keinem Resultat führte. Daher erfolgte umgehend und unter direkter Federführung von Edgar Haimerl der „Neustart" der Entwicklung dieser Software, die dabei durch den als *Map Generator* bezeichneten Teil, der die Kartenerzeugung unterstützt, komplettiert wurde.

Technisch gesehen erfolgte die Entwicklung auf der Grundlage von *Java EE,* im speziellen von *EJB 2.*1. Ein großes Manko dieser Software ist, dass sie nicht von Anfang an auf Nutzungseffizienz getrimmt war. Nähere Details findet man in den nachfolgenden Abschnitten 3. und 4.

3.4.2 SDB – SoundDatenBank

Um auf die während der Exploration aufgenommenen Tondaten schnell und punktgenau zugreifen zu können, entstand ab Frühjahr 2012 die zweite Version der Sounddatenbank (*SDB2*). Die erste Version (*SDB1*) war eine Behelfslösung, die vor allem mit dem Ziel der raschen und einfachen Umsetzung entstanden ist und keine vollständige Abbildung der vorhandenen Daten auf die Datenstruktur erlaubte.

Diese neue Software entstand als monolithische Anwendung und enthielt von Anfang an die folgenden Funktionen: die Neuerfassung von Tondateien, die Segmentierung längerer Ton-Strecken durch Taggen zwecks Erzeugung eigener Responses, die damit verbundenen Korrekturen sowie das Anhören der getaggten Responses über das Netzwerk. Die Ton-Daten werden dabei nach Bedarf ganz oder ausschnittsweise übertragen. Zur besseren Orientierung wird dem Benutzer eine Schalldruckkurve präsentiert, mit deren Hilfe er innerhalb ganzer Ton-Dateien bestimmte Bereiche markieren und anhören kann, ohne dass deswegen die vollständigen Daten auf den PC des Benutzers übertragen werden müssen.

Serverseitig wurde hier auf *Webservices* gesetzt, die mit *Apache Axis 2* umgesetzt wurden und in Richtung Datenbank durch *Hibernate* ergänzt werden. Dabei handelt es sich um ein Framework, das den Zugriff über *SOAP*[24] übernimmt.

Aufgrund der Erfahrungen mit dem weiter beschriebenen Programm *SuBIReS* und mit der dort verwendeten *Netbeans RCP*[25] wurde die *SDB* mit der Version 2.1 auf diese Plattform portiert, in Module zerlegt und an die sich in der Zwischenzeit geänderten Anforderungen angepasst.

Aufgrund des großen Erfolges der *SDB2* beim ALD-II wurde beschlossen, diese auch auf die Daten des ALD-I auszuweiten. Dieser Vorgang konnte im November 2012 abgeschlossen werden. Seitdem ist es möglich, in der *Version 2.1* der *SDB* frei zwischen ALD-I und II hin- und herzuschalten.

Die aktuelle Version ist über das Internet verfügbar. Der dazu erforderliche Client kann von der Projekthomepage heruntergeladen und unter *Windows*, *MacOS* als auch *Linux* verwendet werden.

---

[24] *SOAP*: *Simple Object Access Protocol*: vereinfacht gesagt handelt es sich um ein Netzwerkprotokoll, um Daten über eine allgemein gültige Beschreibung unabhängig von der Programmiersprache zu übertragen. Abfragen und Daten werden dabei in *XML* gepackt und übermittelt.

[25] *RCP*: *Rich Client Platform*: in grober Näherung handelt es sich dabei um eine Programmbasis, die dem Entwickler Arbeit abnimmt, da beispielsweise Abhängigkeiten zwischen Modulen, deren Update, Inter-Modul-Kommunikation usw. von der Plattform bereits fertig implementiert angeboten werden.

3.4.3 SuBIReS – Supplementary Book and Index Retrieval System

Zur Erstellung der Druckvorlagen – vor allem für die geplanten Papier-Indices zum ALD-II – wurde eine weitere Softwarelösung benötigt, die es gestattet, bei der Vorbereitung der Druckvorlagen die einzelnen Dokumente aus der Datenbank möglichst automatisiert erzeugen zu können.

In layout-technischer Hinsicht wurde auf das automatisch gut erzeugbare *(La)TeX* gesetzt. *SuBIReS* erzeugt damit *LaTeX*-Dokumente und erlaubt es auch, diese kleinweise (d. h. Verzeichnis um Verzeichnis) nach *PDF* umzuwandeln. Damit konnten zum einen der „Supplementband", der jene Daten enthält, die auf den bestimmten Karten des ALD-II keinen Platz mehr hatten, und zum anderen der Indexband (mit vorwärts- und rückwärts alphabetischen Indices zu den italienischen Kartentiteln) erzeugt werden.

Zusätzlich zu den Karten, die weiterhin mittels *DMG* entstanden, wurden zu jeder Atlas-Karte mehrere Listen gedruckt, deren Erstellung im Zuge der Entwicklung von *SuBIReS* ebenfalls automatisiert werden konnte. Diese Listen enthalten zu jeder Karte die vollständigen Transkriptionen inklusive aller Versionen in verschiedener Abfolge: sortiert nach Orten, tokenisiert sowie in vorwärts- und rückwärtsalphabetischer Anordnung.

SuBIReS basiert wiederum auf *Netbeans RCP* für die Anwendung, auf *Apache Axis 2* als Serveranbindung und auf *Hibernate* für den Datenbankzugriff.

3.4.4 IRS2 – Index Retrieval System des ALD-II

Um die Transkripte online durchsuchen zu können, wurde ab Frühjahr 2011 eine auf *Adobe Flex*[26] basierende Web-Anwendung erstellt. Aufgrund der verwendeten Technologie muss auf dem PC des Betrachters die jeweils aktuellste Version von *Flash Player* installiert sein.

Das System erlaubt es, den Datenbestand anhand der reduzierten Transkription (*„ALD light"*) zu durchsuchen. Dabei gibt es zwei Abstufungen der Reduktion: *ALD light* (als simplifizierte Lautschrift) und *A-Z* (als Normalbuchstaben). Im zweiten Fall werden alle Zeichen des Transkriptionssystems auf die entsprechenden Grund-Buchstaben des Alphabets reduziert. Seit dem Frühjahr 2012 ist das System produktiv über die Projekthomepage erreichbar und auch mit den Daten des ALD-II direkt verlinkt, wobei zusätzlich Grafiken, Listen und sogar ganze Sprach-Karten gezeigt werden können. Seit dem Sommer 2012 ist *IRS2* auch für die Daten des ALD-I verfügbar, jedoch ohne die Option direkter Links zu den eben erwähnten Grafiken, Listen und Sprach-Karten.

Serverseitig wurde hier auf *Blaze DS* anstatt auf *SOAP* gesetzt. Die Anbindung an die Datenbank erfolgte wie bei allen Webservice-Anwendungen des ALD-II über *Hibernate*.

3.4.5 BDB – Bilddatenbank

Im Sommer 2012 begannen die Arbeiten an der Bilddatenbank (*BDB*), die die Digitalisate aller bei ALD-I und ALD-II gemachten Dias verwalten und über das Web verfügbar halten sollte.

Die Bilddatenbank sollte serverseitig sowohl über *SOAP* als auch über *PHP* erreichbar sein, um die Integration der Daten in die Projekthomepage zu vereinfachen. Da die Arbeiten zur Nachbearbeitung und Kategorisierung der digitalisierten Dias nicht – wie ursprünglich geplant – schon im Herbst des Jahres 2012 begonnen werden konnten, war es

[26] Jetzt: *Apache Flex*.

nicht mehr möglich, die *BDB* bis zum Projektende fertig zu stellen. Eine Vollendung zu einem derzeit noch nicht feststehenden Zeitpunkt ist aber angedacht.

3.4.6 Homepages

Im Jahr 2011 begannen die Arbeiten an der Neugestaltung der Homepages der beiden Teile des Gesamtprojektes ALD. Sowohl der ALD-II als auch der ALD-I haben damit ein neues Gesicht bekommen. Die neuen Homepages bestehen nicht mehr aus statischen *HTML*-Seiten, sondern werden über ein *Content-Management-System (CMS)* namens *Concrete 5* verwaltet. Damit ist es auch für „Nicht-Techniker" möglich, einzelne Elemente der Homepage zu ändern bzw. in verschiedene Sprachen zu übersetzen, was sich sehr positiv auf die Wartbarkeit auswirkt.

Der letzte Schritt, auch die neue Version des Web-Sprechenden des ALD-I in die neue Homepage des ALD-I zu integrieren, konnte bis Projektende leider nicht mehr realisiert werden.

3.5 Erkenntnisse

Es hat sich gezeigt, dass es sehr vorteilhaft ist, dass der EDV-Techniker seine Arbeit inmitten des Projektteams verrichtet. Auf der einen Seite erhält er dadurch ein direktes Feedback zu den eigenen Lösungen; andererseits lassen sich dadurch neue Ideen und die damit verbundenen Programme im wechselseitigen Austausch rasch entwickeln bzw. erstellen.

Die Verwendung einer Projektmanagement-Lösung wie beispielsweise *Redmine*, das bei uns zum Einsatz kam, war unverzichtbar, um rechtzeitig steuernd eingreifen zu können und die entwickelten bzw. zu entwickelnden Softwarelösungen und deren Änderungen verwalten zu können. Dass zur Quellcode-Verwaltung ein Versionsverwaltungssystem (*VCS*) wie beispielsweise *Subversion* eingesetzt wurde, war dabei eine Notwendigkeit.

Ein weiterer wichtiger Punkt war auch, die Software so *modular* wie möglich zu entwickeln, um die Wiederverwertung von Quellcodes zu vereinfachen bzw. überhaupt erst zu ermöglichen. Klare und einfache Konzepte auch im Sinne des *KISS*-Prinzips („*Keep it small and simple*") haben sich hier immer bezahlt gemacht. Dazu gehörte aber auch, dass Ziele möglichst früh definiert werden mussten, wenngleich dies bei einem so komplex strukturierten Forschungsprojekt natürlich schwierig war, da sich laufend neue Notwendigkeiten ergeben haben bzw. bestehende Lösungsansätze abgeändert werden mussten.

4 Zukunftsperspektiven

Mit dem Ende des Jahres 2012 endete das Forschungsprojekt ALD-II. Zu diesem Zeitpunkt wurden alle Softwarekomponenten auf dem erreichten Stand eingefroren und werden hinfort – zumindest von Seiten der Universität Salzburg – nicht weiter betreut. Es steht zu erwarten, dass, solange die Hardware des *VMWare ESXi Servers* funktionsfähig bleibt, die Server-Dienste auf dem aktuellen Stand erhalten bleiben. Danach sollten diese nahtlos in die IT-Struktur der Universität Salzburg eingegliedert werden.

Dass die Software lauffähig bleibt, scheint zumindest für die Dauer der nächsten Jahre wahrscheinlich, da *Java* als etablierte Umgebung auf fast jedem PC anzutreffen ist. Dagegen ist – zumindest für die nächsten Versionen – *Flashplayer* als integraler Bestand-

teil des Browsers *Google Chrome* in *plattform-unabhängiger* Form nur eingeschränkt verfügbar.

Die Webseiten und auch die damit verbundenen *HTML5*-Inhalte dürften, wenn man ihnen dieselbe Lebensdauer wie den heute immerhin schon 15 Jahre alten *HTML*-Dokumenten zubilligt, am längsten verfügbar bleiben.

5 Zur praktischen Anwendung der für den ALD-I und ALD-II entwickelten EDV-Werkzeuge: die Sound-Datenbank (SDB) (*Heidemarie Beer*)

5.1 Die für den ALD-II entwickelte Sound-Datenbank

Neben den Transkripten und den bildlichen Daten (Diapositiven) gibt es beim ALD-II auch recht umfangreiches Tonmaterial, das während der Enquêten erfasst wurde.

Es sind dies die zu den 217 Messpunkten des ALD-II und dessen 1 063 Fragen erhobenen akustischen Daten. Sie wurden von 10 Exploratoren bei 833 Informanten erhoben. Diese durchwegs in digitaler Form vorliegenden Aufnahmen umfassen insgesamt mehr als 20 000 Stunden bzw. ein Datenvolumen von rund 145 Gigabyte.

Von Anfang an war geplant, alle Tonaufnahmen in einer Sound-Datenbank (*SDB*) zur Verfügung zu stellen, die es erlaubt, bestimmte Schallereignisse in ihrer originalen Form möglichst exakt und rasch aufzufinden und abzuhören.

Mit der Einführung einer neuen Version der *SDB* im Frühjahr 2012 (*SDB2*) wurde es viel leichter, sich im Programm zu orientieren, die gewünschte Fragenummer zu finden und abzuhören.

Die Abbildung 3 zeigt die verschiedenen Felder des Programmfensters der *SDB*.

Zunächst gibt man die gewünschte Ort- bzw. Fragenummer in die entsprechenden Felder ein und klickt auf „Query“. Nun werden im rechten oberen Feld alle Sounddateien aufgelistet, die es zu diesem Ort gibt.

Wählt man (rechts oben) eine Sounddatei aus, so erscheinen im Feld darunter alle darin vorhandenen Fragenummern mit den Positionsangaben. Außerdem erscheint in der Mitte unten das Schalldruckdiagramm. Damit kann man Segmente, die man abhören möchte, markieren und sich nach Bedarf vor- oder zurückbewegen. Das Schalldruckdiagramm bildet die jeweilige Tonstelle visuell ab und bietet damit die Möglichkeit, sich noch genauer zu orientieren und das Gehörte zu vertiefen

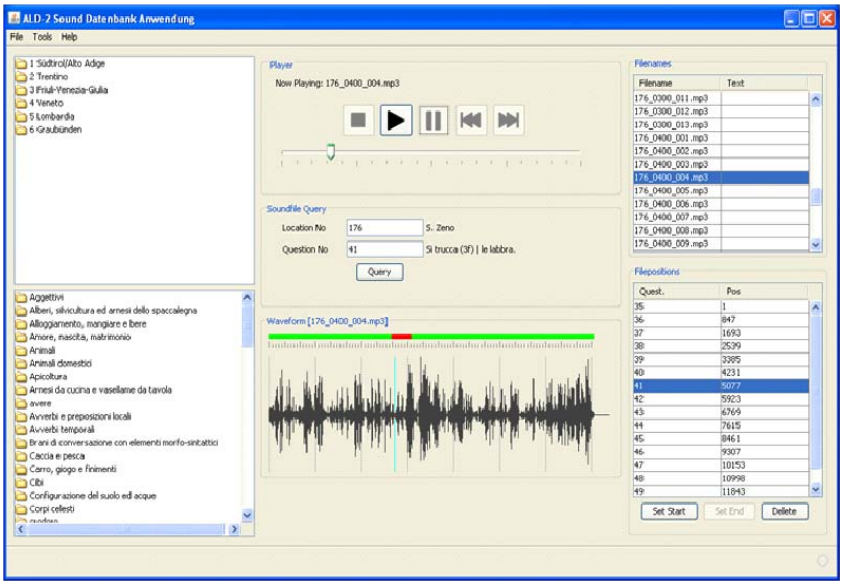


Abbildung 3: Die *SDB2* des ALD-II in Betrieb (Abhörbeispiel: ALD-Ortschaft 176, Frage 41)

In den beiden linken Feldern sehen wir zusätzliche Einträge, nach denen man die Suche ebenso ausrichten kann. Im unteren Bereich sind dies die Regionen der ALD-Zone, darüber die nach Sachthemen gruppierten Fragen des ALD-II-Questionnaires. Wenn man nun beispielsweise den Ordner *Trentino* auswählt, so klappen alle dem Trentino zugehörigen Orte des ALD auf. Ebenso verhält es sich bei den anderen Regionen.

Letztlich entscheidet aber der interessierte Benutzer am besten für sich selber, welche Suchmethode (1. Suche im regionalen Ordner, 2. Eingabe der Nummern der Messpunkte und der Fragen) er zur Ingangsetzung der *SDB2* bevorzugt.

Die *SDB* des ALD-II war für uns auch während der Redaktion der Atlas-Karten von großer Bedeutung, so vor allem für die Rekonstruktion von Teilen komplexer Fragen des Fragebuchs. Zum einen handelte sich es sich dabei um die Rekonstruktion der Antworten auf Stimuli, die nach den Vorgaben des Fragebuchs bei der Enquête nicht transkribiert werden sollten. Zum anderen mussten während der letzten Korrekturdurchgänge redaktionelle Unsicherheiten abgeklärt werden. So saßen der Projektleiter und ich nicht nur einmal in unserem ALD-Büro im alten Gebäude (Akademiestraße 24) und hörten uns gemeinsam die eine oder andere Frage an, um nochmals zu überprüfen, welche Transkription im speziellen Fall nun die richtige wäre.

5.2 Die beim ALD-II getätigten Nachaufnahmen

An einigen Messpunkten des Befragungsgebietes verblieben nach dem Ende der Regel-Enquêten (im Jahr 2007) Datenlücken. Es handelte sich dabei um insgesamt 25 Ortschaften, in denen die ALD-II-Exploratorin Brigitte Rührlinger im Herbst 2010 sowie im Frühjahr und Sommer 2011 in bewährter Weise Nachenquêten durchführte.

Bei diesen Enquêten fielen erneut akustische Daten an, die anschließend weiterbearbeitet werden mussten. Es war nun meine Aufgabe, diese Daten abzuhören und zu schneiden. Dies bedeutete, dass die auf Speicherkarten vorliegenden und oft recht langen Dateien in etwa 5 Minuten lange Segmente zerlegt werden mussten. Nebenbei waren

Listen anzufertigen, in die einzutragen war, welche Fragenummern sich in der jeweils neu geschnittenen mp3-Datei befanden.

Teilweise wurden diese von Brigitte Rührlinger neu erfassten Daten auch in das Programm *DMG* eingegeben und – soweit möglich bzw. notwendig – sogar noch in die Korrektur und Redaktion der zum Druck bestimmten Karten einbezogen.

Nach dem Taggen galt es, die neuen mp3-Dateien in die schon bestehende Sound-Datenbank einzupflegen. Dazu waren vor allem die vorhin erwähnten systematisch angelegten Listen von Bedeutung.

Für das Taggen selbst verwendete ich das Audio-und Schneideprogramm *Audacity*, mit dem man im Allgemeinen sehr gut arbeiten und sogar Mängel in der akustischen Qualität der Daten sehr effektiv ausgleichen oder beseitigen kann (s. dazu Abbildung 4).

5.3 Die für den ALD-I nachträglich erstellte Sound-Datenbank

Nach den äußerst positiven Erfahrungen mit der Sound-Datenbank des ALD-II kam bald der Wunsch nach einer analogen Möglichkeit für die Tondaten des ALD-I auf.

Vom ALD-I gab es systematisch aufbereitetes akustisches Material zunächst nur in der Form des „Sprechenden“, der jedoch nur die 21 Messpunkte der brixnerisch-tirolischen Ladinia (ALD-PP. 81-101) abdeckt. Bei der Erstellung der Sound-Datenbank des ALD-I und dem damit verbundenen Taggen wurden wegen ihrer wesentlich besseren Qualität die dem „Sprechenden“ zu Grunde liegenden DAT-Aufnahmen und nicht die bei den Regel-Enquêten gemachten (und mit einem deutlich hörbaren Grundrauschen behafteten) C90-Kassetten verwendet.

Sämtliche akustischen Daten des ALD-I lagen ursprünglich in analoger Form auf mehr als 1 200 C90-Kassetten vor. Diese nicht unbeträchtliche Daten-Menge wurde im Sommer 2010 von Brigitte Rührlinger zunächst in ziemlich mühseliger und zeitfressender Kleinarbeit digitalisiert und entrauscht. Bei insgesamt 217 Orts-Aufnahmen kam schließlich ein Datenvolumen von insgesamt etwa 95 Gigabyte zusammen, das nach dem Schneiden immerhin noch 88 Gigabyte umfasste.

Im Herbst 2010 begann Brigitte Rührlinger – zunächst im Alleingang – die zuvor digitalisierten Daten zu taggen, wobei die jeweils 45 Minuten umfassenden Vorder- und Rückseiten der C90-Kassetten in etwa 5 Minuten lange Segmente zerlegt werden sollten. Das Abhören und Schneiden einer Kassetten-Seite dauerte durchschnittlich 40 Minuten und führte zu durchschnittlich neun Schnitten.

Allerdings waren die bei dieser Arbeit erzielbaren Fortschritte sehr unterschiedlich, da manche Aufnahmen ein rascheres, andere wiederum ein weniger zügiges Vorankommen erlaubten. Hilfreich und arbeitsökonomisch war in einer solchen Situation die Technik des „Springens“ von einer Frage zur nächsten. Ein weniger zügiges Vorankommen ergab sich dann, wenn die Gewährspersonen langatmige Digressionen zum Besten gaben, von denen die interessierenden Frage-Antwort-Komplexe sozusagen „zugeschüttet“ waren. In solchen Fällen war ein genaueres Abhören nötig: man konnte nur selten „springen“.

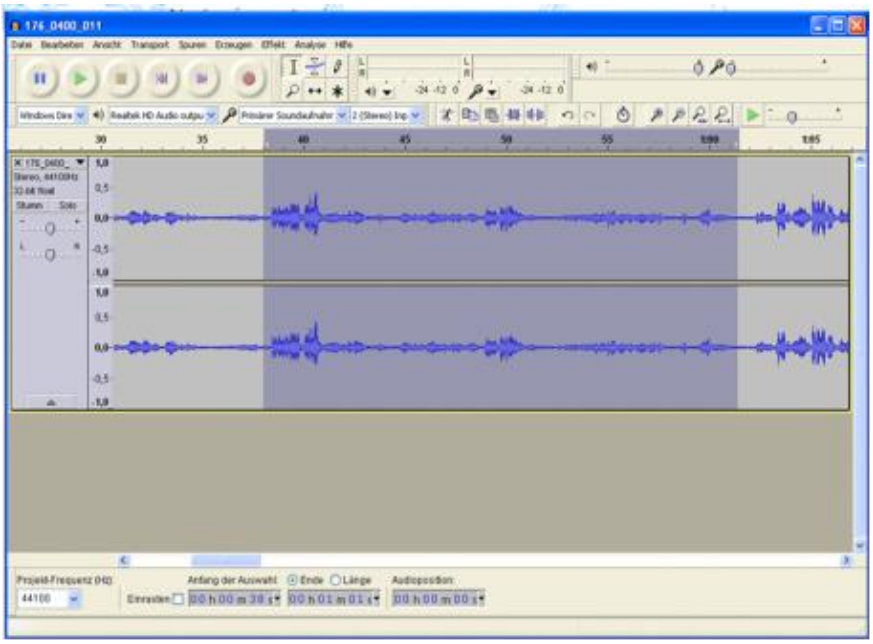


Abbildung 4: Taggen der Tondaten mit *Audacity*

Angesichts der großen Datenmenge und der Kürze des zur Verfügung stehenden Zeitrahmens – die Arbeiten sollten auf jeden Fall vor dem Ende des Jahres 2012 beendet sein – wurden zusätzliche Mitarbeiter in diese Tätigkeit einbezogen. Neben Brigitte Rührlinger waren dies in weiterer Folge Stefanie Holzner, die beiden Italienerinnen Ilaria Adami und Tiziana Gatti sowie meine Wenigkeit. Ilaria Adami war – wie Brigitte Rührlinger – ebenfalls Exploratorin beim ALD-II. Bei den ladinischen Orten unterstützte uns eine Zeitlang auch der Salzburger Studienassistent Christoph Hülsmann. Die Koordination der in und um Salzburg ablaufenden Arbeiten war meine Aufgabe, während die in Italien zu erledigenden Arbeiten von Brigitte Rührlinger (von ihrem dortigen Wohnort Sabbio Chiese aus) überwacht wurden. Allerdings wurden gar manche Details im direkten Gespräch zwischen Ilaria Adami sowie Tiziana Gatti und meiner Wenigkeit besprochen bzw. geklärt.

Insgesamt dauerten die Tagging-Arbeiten 1 ¼ Jahre, was bei einer 30-Stunden-Woche einem Personenarbeitsjahr entspricht. Ab dem Frühjahr 2012 waren wir über mehrere Monate sogar zu sechst mit der „Belauschung" der Tondaten des ALD-I beschäftigt, wobei wiederum das Schneide- und Akustik-Programm *Audacity* zum Einsatz kam. Die Bearbeitung der letzten 20 Orte lag schließlich in den Händen von Brigitte Rührlinger, Stefanie Holzner und mir selber.

Neben dem Programm *Audacity* schien auch die Verwendung eines weiteren Programms sinnvoll zu sein, das den Arbeitsfortschritt dokumentiert und jedem Mitarbeiter einen Überblick über das Geschehen – zu seiner eigenen Tätigkeit und jener der anderen – gibt. Bernhard Schauer adaptierte dazu das Verwaltungsprogramm *Redmine*.

Die Abbildung 5 zeigt ein leeres Ticket, worin wir Mitarbeiter verschiedene Daten einzutragen hatten, so etwa:

- Zuweisen des Messpunktes an den entsprechenden Mitarbeiter;
- Anfang und Ende der Arbeit;

- Arbeitszeit (Stunden, die man für das Taggen aufgewendet hat);
- Prozentsatz der erledigten Arbeit eines Messpunktes;
- eventuelle Kommentare, wenn Auffälliges zu vermerken war.

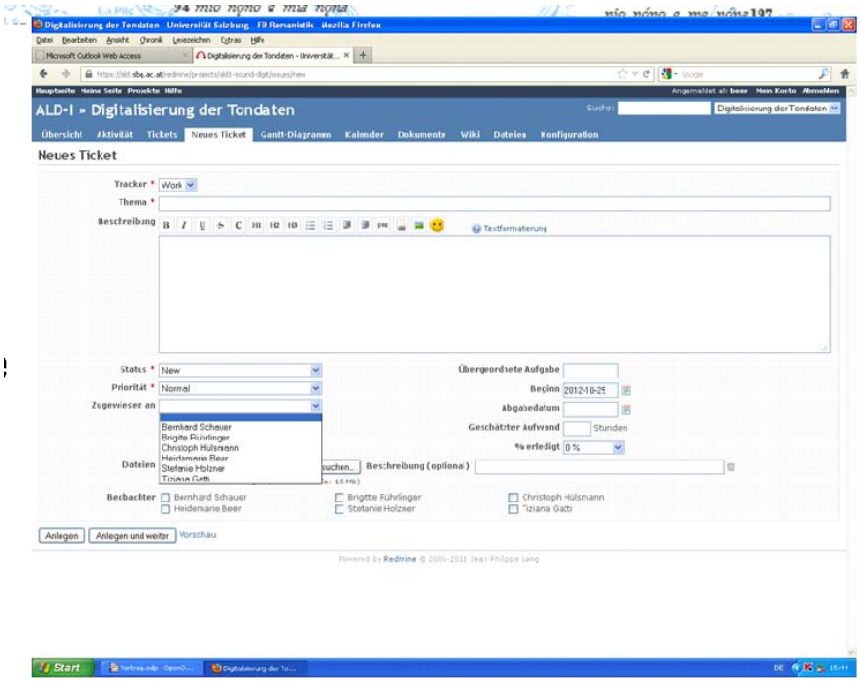


Abbildung 5: Beispielsseite aus Redmine

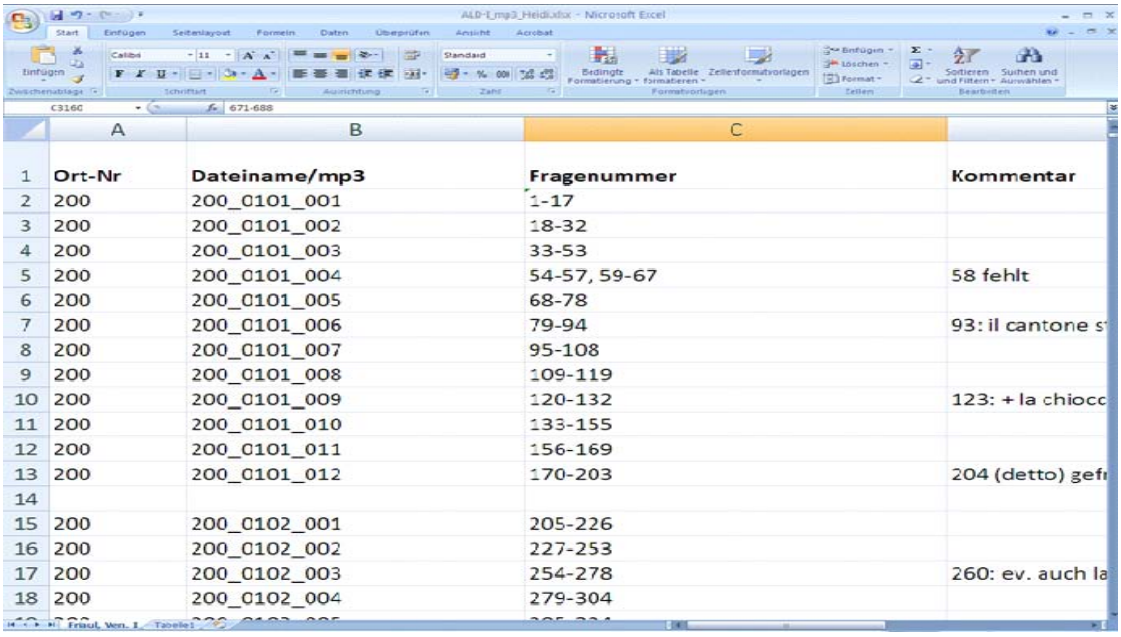
Für jeden der 217 Messpunkte wurde ein eigenes Ticket angelegt. Diese Tickets wurden – entsprechend dem Arbeitsfortschritt – laufend von den jeweiligen Mitarbeitern aktualisiert.

Ende September 2012 konnten die Schneidearbeiten abgeschlossen werden.

Für das nachfolgende Einpflegen aller neu geschnittenen mp3-Dateien in die Sounddatenbank waren wiederum sehr systematisch aufgebaute Listen unentbehrlich, die über den Messpunkt, die jeweilige Datei und die darin vorhandenen Fragenummern genaue Auskunft gaben.

Die Abbildung 6 zeigt ein Detail aus einer solchen Begleitliste.

Zuletzt wurde bei der Berliner Demonstration ein Hörbeispiel aus der SDB des ALD-II zum Besten gegeben: und zwar zum ALD-Messpunkt 176, San Zeno, der von Brigitte Rührlinger exploriert worden war. An diesem Messpunkt – wo auch eine Nachenquête stattfand – stand Brigitte Rührlinger ein besonders guter Informant zur Verfügung. Zudem ist die Tonqualität der Aufnahme sehr gut.



	A	B	C	
1	Ort-Nr	Dateiname/mp3	Fragenummer	Kommentar
2	200	200_0101_001	1-17	
3	200	200_0101_002	18-32	
4	200	200_0101_003	33-53	
5	200	200_0101_004	54-57, 59-67	58 fehlt
6	200	200_0101_005	68-78	
7	200	200_0101_006	79-94	93: il cantone s
8	200	200_0101_007	95-108	
9	200	200_0101_008	109-119	
10	200	200_0101_009	120-132	123: + la chiocc
11	200	200_0101_010	133-155	
12	200	200_0101_011	156-169	
13	200	200_0101_012	170-203	204 (detto) gefi
14				
15	200	200_0102_001	205-226	
16	200	200_0102_002	227-253	
17	200	200_0102_003	254-278	260: ev. auch la
18	200	200_0102_004	279-304	

Abbildung 6: Beispiel einer Begleitliste zur Kontrolle der einzupflegenden Sound-Dateien

Es wurde eine Passage vorgeführt, die auch Amüsantes beinhaltet. Man erkennt daraus, wie unterhaltsam sich manche Enquêten gestalten können. Das vorgeführte Gespräch bezog sich auf einen Teilbereich der Frage 44 des ALD-II-Questionnaires – *il moccio* („der Nasenschleim“) –, den man nach Eingabe der Nummern 176 (für die Ortschaft) und 44 (für die interessierende Frage) sowie unter Konsultation der Schalldruckkurve leicht auffinden kann. Mit der Maus können nach Belieben kleinere oder größere Segmente zur Abhörung ausgewählt werden. Man kann aber auch – wenn gewünscht – rasch zu einer anderen Frage übergehen.

Jeder, der die Sounddaten des ALD-II – seit kurzem gesellen sich ja auch jene des ALD-I dazu – anhören möchte, kann jederzeit nach eigenem Gutdünken selbst in die Originalaufnahmen hineinhören. Dafür stehen fortan sämtliche Daten im Netz (mit Installations- und Bedienungsanleitung) zur allgemeinen Verfügung:

ALD-I: http://ald1.sbg.ac.at/a/index.php/de/daten/sound-datenbank/.

ALD-II: http://ald2.sbg.ac.at/a/index.php/de/daten/sound-datenbank/.

6 Zur praktischen Applikation der für ALD-II und ALD-I entwickelten EDV-Werkzeuge: das Programm DMG und die Homepages (*Agnes Staudinger*)

6.1 DMG (Dialect Map Generator)

Nach Abschluss der Aufnahmen vor Ort wurden die in den Fragebüchern deponierten Trankskriptionen in Salzburg händisch in *DMG* übertragen. Die Eingabe der betreffenden Lautzeichen erfolgte alternativ durch Anklicken in einem auf dem Bildschirm sichtbaren Alphabetikum oder durch die Applikation von Tastenkombinationen auf der Tastatur. Die Verwendung der Tastenkombinationen erwies sich dabei als weniger zeitintensiv, allerdings unter der Voraussetzung, dass man zuvor das System gut durchschaut hatte bzw. sogar auswendig kannte: siehe dazu die Abbildung 7.

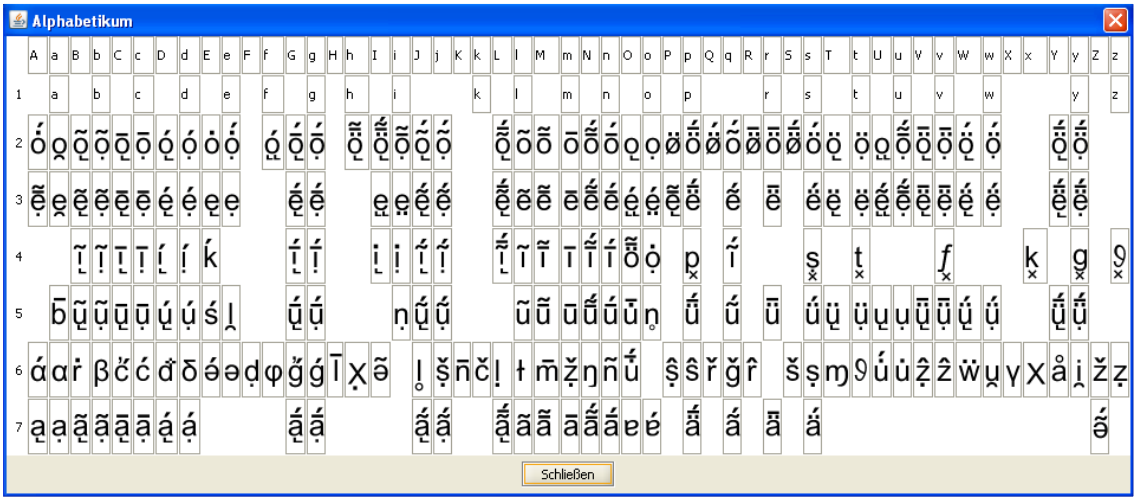


Abbildung 7: Alphabetikum der beim ALD (Teil I und Teil II) verwendeten Lautschrift

So entspricht eine *Ziffer* (1-7: als *Zeilen* organisiert) hauptsächlich den verschiedenen Varianten eines Vokals: *Zeile* 2: o-Laute, *Zeile* 3: e-Laute, *Zeile* 4: i-Laute, usw. Die *Buchstaben* (Aa-Zz; als *Spalten* organisiert) definieren dagegen den Typ: *Spalte* D: offen und betont, *Spalte* d: geschlossen und betont, usw.

Drei konkrete Anwendungsbeispiele: 2d: geschlossenes betontes o; 3d: geschlossenes betontes e; 4d: geschlossenes betontes i, etc.

Auf Abbildung 8, links oben, sieht man die Frage 2, die aus vier Teilfragen (*Dove...saranno...mio nonno e...mia nonna?*) besteht, als Baum. Es wird gerade die Transkription der dialektalen Antwort auf die Frage „Dove“ eingegeben. Man hat nun (z. B. bei zwischen Frage und Antwort divergierender Syntax) die Möglichkeit, bei der Eingabe die Reihenfolge der zu erfassenden Teil-Antworten/Fragen zu vertauschen, und kann außerdem (bei Doppel- oder Mehrfachantworten) eine weitere Version hinzuzufügen.

Abbildung 8: Eingabefeld von DMG zur Frage 2 des ALD-II-Fragebuchs

Nach der Eingabe und Korrektur aller Transkriptionen konnte der erste Ausdruck von Probe-Karten realisiert werden. Diese wurden daran anschließend aus philologisch-linguistischer Perspektive korrigiert. Die Erzeugung solcher Karten mittels *DMG* war

relativ komplex und damit sehr zeitaufwändig. So waren für die Produktion einer Karte bis zu 11 Mausklicks nötig, um ein ausdruckbares PDF zu erzeugen.

Bei den beiden ersten der insgesamt drei Campagnen zur Produktion von Probe-Karten wurden bei komplexeren Fragen fast alle Möglichkeiten der Kombination von Einzel- und Teilfragen (als „Kombi-" [KK] und „Doppel"-Karten [DK]) ausprobiert. Dadurch kamen pro Produktions-Campagne für die 1 063 Fragen des ALD-II-Questionnaires insgesamt 2 026 Karten zusammen.

In den Jahren 2009-2012 wurden die PDFs einerseits für drei intermediäre Produktions-Campagnen (zu je 2 026 Karten) und andererseits für den finalen Durchgang (mit 1 066 zur Publikation ausgewählten Atlas-Karten) erstellt: das sind in toto (= 3 mal 2 026 + 1 066) 7 144 Karten-PDFs, für deren Generierung 22 286 Mausklicks nötig waren. Rechnet man dies auf die dafür benötigte Zeit um (das Erstellen und Abspeichern eines Karten-PDFs dauert ca. 1,5 Minuten), dann kommt man auf 178,6 Stunden reiner Arbeitszeit bzw. auf 7,14 Personenarbeitswochen (bei 30 Stunden/Woche).

Der Ausdruck einer Karte auf DIN A2 großem Papier nahm mit unserem Vierfarbendrucker *HP designjet* durchschnittlich vier Minuten in Anspruch. Alle vier Durchgänge wurden sowohl auf Prüfpfad-Vorlagen, deren Verwendung die Korrektur erleichtert, als auch auf Blaudruck-Vorlagen, die dem definitiven Erscheinungsbild mit blauem Kartenhintergrund entsprechen, realisiert. Somit kommt man auf ca. 14 300 Karten, deren Ausdruck 953 Stunden benötigte: dem entsprechen bei 30 Stunden/Woche umgerechnet 38,1 Personenarbeitswochen.

Parallel zu den Kartenausdrucken wurden zu Korrekturzwecken verschiedene Begleitlisten angefertigt. Auf ihnen wurden die Karteninhalte z.B. in vorwärts alphabetischer oder in rückwärts alphabetischer Form bzw. auch nach Orten sortiert (1 bis 217) präsentiert. Auch gab es Listen, die jene Informationen beinhalteten, die später im Supplementband („Volumen supplementarium") veröffentlicht wurden, und solche, die nur die Anmerkungen und die Zusatzinformationen einer Karte enthielten.

Die PDFs dieser Listen wurden wie jene der Karten ebenso mit *DMG* erzeugt, was pro Liste ca. 30 Sekunden oder 13 Mausklicks in Anspruch nahm. Bei den im Verlauf von zwei Campagnen manuell erzeugten 9 556 Listen kamen wir so auf 124 228 Klicks bzw. eine Arbeitszeit von 80 Stunden.

Den Ausdruck der Listen führten wir bei den ersten beiden Durchgängen noch am ALD-Archiv durch: dies betraf insgesamt 74 496 Seiten. Da unser *HP LaserJet* ca. 15 Seiten/Minute schaffte, ergab sich daraus eine reine Druckzeit von 82,8 Stunden.

Da der projekt-eigene Drucker damit seine Leistungsgrenze erreicht hatte und zu „schwächeln" begann, beschlossen wir, die maschinell um vieles besser ausgerüstete Kopierstelle der Universität Salzburg mit dem Ausdruck der definitiven Listen zu betrauen, die nach dem Ausdruck archiviert wurden. Diese letzte Erzeugung von listenartig organisierten PDFs wurde außerdem nicht mehr mittels, sondern mit dem neuen, von Bernhard Schauer erstellten Programm *SuBIReS* („*Supplementary Book and Index Retrieval System*") durchgeführt. Für die Generierung aller 4 778 Listen eines kompletten Durchgangs benötigte *SuBIReS* ca. 2,5 Stunden, wobei es im Hintergrund lief und man sich daneben anderen Dingen widmen konnte.

Der Start von *SuBIReS* dauerte pro Listenart ca. 10 Sekunden. Diese Aktion und die nachfolgende PDF-Erzeugung mussten insgesamt vier Mal repetiert werden. Daraus ergab sich eine Gesamtzeit von etwa einer Minute für den Start. Für die Erzeugung und den Druck der Listen gelangten wir somit auf 162,8 Stunden reiner Arbeitszeit, also auf 6,5 Personenarbeitswochen (bei 30 Stunden/Woche).

Insgesamt ergab sich damit ein ganzes Personenarbeitsjahr (bei 30 Stunden/Woche) für die Erzeugung und den Druck aller Karten und Listen.

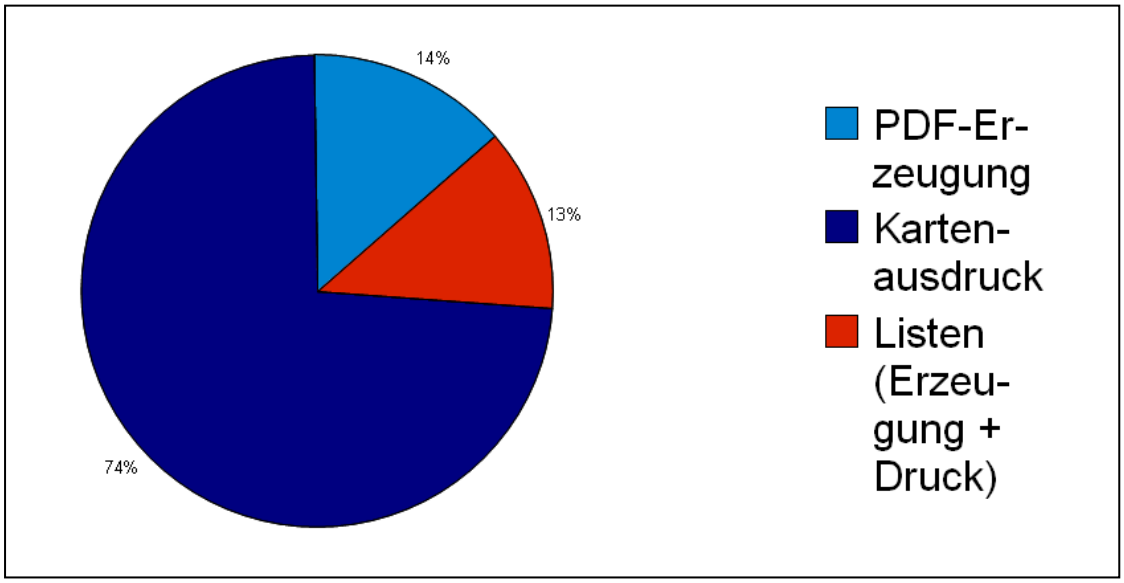


Abbildung 9: Aufteilung eines Arbeitsjahrs nach Tätigkeiten

In dieser Arbeitszeit sind allerdings das Falten der A2 großen Kartenausdrucke, das Zusammenheften der (A4 großen) Listen, das Abstempeln beider sowie deren Einordnung in die A3 großen Hängeordner nicht enthalten.

6.2 Ethnophotographische Daten

Die Exploratoren fertigten im Zuge der Enquêten auch Dias von den besuchten Orten an. Darauf wurden unter anderem die Gewährspersonen selber, die jeweilige Dorfkirche samt Turm und Innenraum, der Friedhof, Denkmäler und das Straßenbild festgehalten. Beim Friedhof war es für uns von besonderem Interesse, ob es dort nur Erdgräber (ital. *tombe*) oder auch Nischengräber (ital. *loculi*) gab. Die Bestattung in *loculi* hat in den südlichen Teilen der ALD-Zone Tradition, breitet sich aber immer mehr nach Norden hin aus und bringt somit eine Änderung der Begräbnis- und Grabkultur mit sich.

An den 217 Messpunkten wurden sowohl beim ALD-I als auch beim ALD-II jeweils 30-40 Aufnahmen gemacht, wodurch es möglich wurde, einerseits zwei synchrone Bild-Dokumentationen aller Aufnahmeorte zu erhalten und andererseits die diachrone Entwicklung jedes Ortes über jene 15 Jahre hinweg, die zwischen ALD-I und ALD-II liegen, zu beobachten.

Insgesamt gibt es von beiden Atlasteilen ca. 17 000 Aufnahmen. Davon bestand der Großteil aus analogen Diapositiven, die von einer Mitarbeiterin (Stefanie Holzner) mittels Scan digitalisiert wurden. Unser Plan war, bis Ende des des Jahres 2012 eine (grob getaggte) Bilddatenbank aufzubauen und mit Daten zu befüllen, die es Außenstehenden erlauben sollte, darin unter Eingabe der Nummern der 217 Ortschaften und verschiedener Oberbegriffe (wie „Kirche“, „Friedhof“, „Informant“ etc.) zielorientiert oder enzyklopädisch zu suchen. Leider konnte dieses Ziel nicht erreicht werden.

6.3 Homepages

Die Homepage des ALD-II wurde mit einem *CMS* erstellt, dessen Vorteil darin besteht, dass Personen ohne spezielle Programmierkenntnisse Inhalte hinzufügen und bearbeiten können. Es erschien uns dabei wichtig, der Fachwelt neben grundsätzlichen Informationen zum ALD-II auch konkretes Datenmaterial zugänglich zu machen. So kann sich jeder das komplette Fragebuch, stumme Karten zum Untersuchungsnetz und sogar alle 1 066 Karten des ALD-II als PDFs herunterladen und bei Bedarf auch ausdrucken. Außerdem gibt es auf der Homepage Statistiken zu den Informanten, den Exploratoren und zu den enquêtierten Ortschaften.

Man kann nach einzelnen Karten und auch nach Wörtern, die in den (italienischen) Kartentiteln vorkommen, suchen. Über die *SDB2* des ALD-II lassen sich die 217 im Feld gemachten Aufnahmen punktgenau anhören und die 1 066 Sprachkarten mittels des digitalen Index (IRS2) in sehr gezielter Weise absuchen. Auch findet man auf unserer Homepage ein Flußdiagramm zu den über die Jahre gemachten Arbeitsfortschritten sowie eine ständig à jour gehaltene Übersicht über den jeweiligen Stand einzelner Arbeitsschritte.

Abbildung 10: Homepage des ALD-II, deutsche Version

Um der Homepage des ALD-II eine größtmögliche internationale Sicht- und Lesbarkeit zu verschaffen, gibt es davon fünf sprachliche identische Versionen: und zwar neben Deutsch auch in Ladinisch (Ladin dolomitan), Italienisch, Englisch und Französisch.

Abbildung 11: Homepage des ALD-II, ladinische Version

Analog zur Homepage des ALD-II wurde auch die Homepage des ALD-I in einem neuen Design erstellt, da die auf die 1990er-Jahre zurückgehende alte Homepage weder optisch noch inhaltlich à jour war und eigentlich nur mehr historischen Wert besaß. Allerdings stellte sie bis Mitte 2012 die einzige netzbasierte Informationsquelle zum ALD-I dar. Sie wurde nunmehr komplett überarbeitet, sodass die beiden neuen Homepages einander nicht nur im Design, sondern auch in der Menüführung gleichen.

Am Ende des Jahres 2012 war die neue Homepage des ALD-I vorerst in den Sprachen Deutsch, Italienisch und Französisch verfügbar. Die restlichen zwei Sprachen (Ladin dolomitan und Englisch) sollen später eingearbeitet werden.

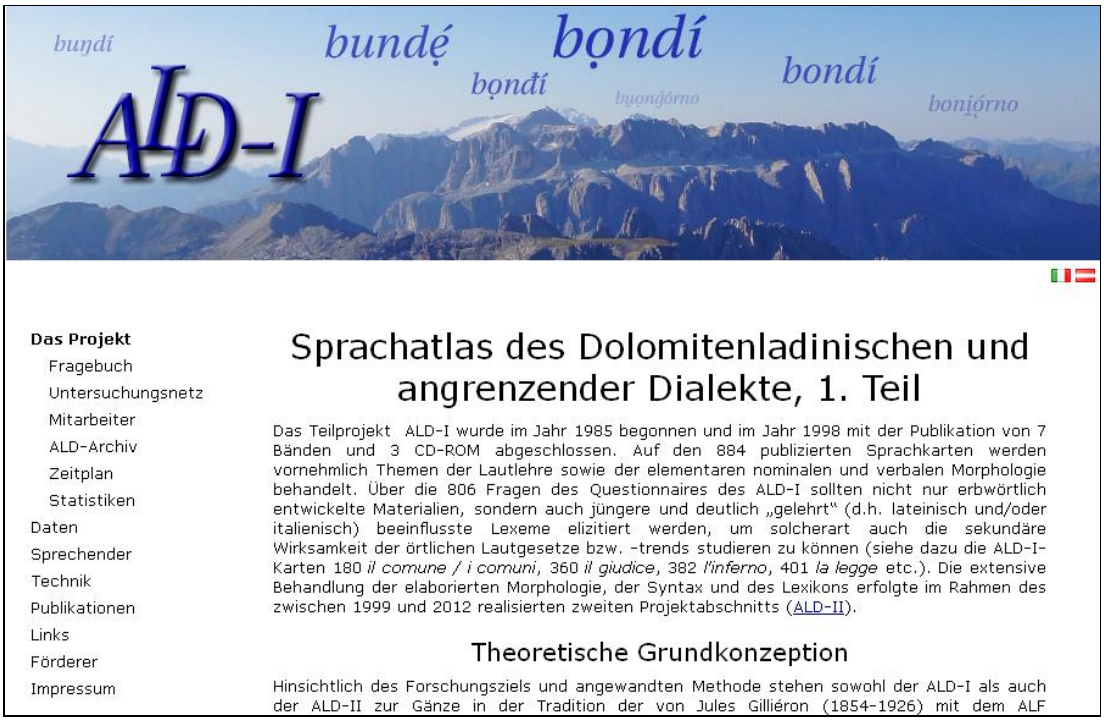


Abbildung 12: Homepage des ALD-I, deutsche Version

6.4 Texmaker

Das Vorwort und die Deckblätter der Kartenbände sowie die Indices und der Supplementband wurden mit dem *LaTeX*-Programm *Texmaker* erstellt. Der Vorteil dieses Programms besteht in einer relativ leichten Handhabung bei gleichzeitigem professionellem Design, wobei es zu keinerlei Verrutschen oder anderen Formatierungsproblemen kommen kann, wie sie beispielsweise bei Word nicht selten sind.

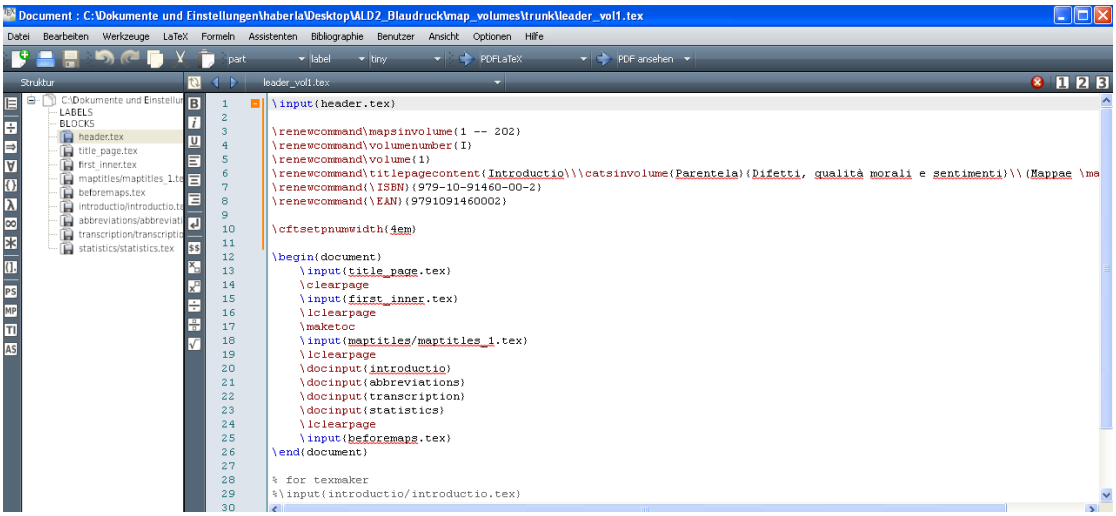


Abbildung 13: ein mit LaTeX erstelltes Dokument

Links sieht man den Aufbau des Dokuments, in der Mitte das Dokument selbst, oben lässt sich mit einem Mausklick ein PDF erzeugen („*PDFLaTex*") und ansehen („*PDF ansehen*"). Außerdem gibt es, ähnlich wie bei Word, im Menü viele nützliche Hilfen, so dass

man nicht alle Befehle selbst auswendig kennen muss. So ist es beispielsweise möglich, den Text mit einem Mausklick kursiv zu setzen, ohne dazu eigens den Befehl „`\textit{}`“ einfügen zu müssen.

Bibliographie

ALF: Gilliéron, Jules/Edmont, Edmond (Hg.): Atlas linguistique de la France, Paris: Champion, 1902-1910, 10. vol. (Neudruck: Bologna: Forni, 1968).

ALD-I: Goebl, Hans (Hg.), unter Mitarbeit von: Bauer, Roland/Haimerl, Edgar/Böhmer, Helga/Gislimberti, Silvio/Kattenbusch, Dieter/Perini, Elisabetta/Szekely, Tino/Dautermann, Irmgard/Heißmann, Susanne/Hofmann, Ulrike/Kozak, Anna/Pamminger, Heide Marie/Rössler, Judith: *Atlant linguistich dl ladin dolomitich y di dialec vejins, 1ª pert. Atlante linguistico del ladino dolomitico e dei dialetti limitrofi, 1ª parte. Sprachatlas des Dolomitenladinischen und angrenzender Dialekte, 1. Teil*, Wiesbaden: Dr. L. Reichert, 1998, 4 vol. Karten (vol. I: 1-216; vol. II: 217-438: vol. III: 439-660; vol. IV: 661-884), 3 vol. Indizes (Index alphabeticus, 823 S.; Index alphabeticus inversus; 833 S.; Index etymologicus, 177 S.) mit 3 CD-ROM (Salzburg 1999-2000), 1 DVD (Salzburg 2002f.).

ALD-II: Goebl, Hans (Hg.), unter Mitarbeit von: Haimerl, Edgar/Schauer, Bernhard/Tosques, Fabio/Wagner, Andreas/Adami, Ilaria/Böhmer, Helga/Heinemann, Axel/Jodl, Frank/Klinger, Liza/Rando, Daniele/Rührlinger, Brigitte/Strauß, Walter/Szekely, Tino/Videsott, Paul/Beer, Heidemarie/Staudinger, Agnes: *Atlant linguistich dl ladin dolomitich y di dialec vejins, 2a pert. Atlante linguistico del ladino dolomitico e dei dialetti limitrofi, 2a parte. Sprachatlas des Dolomitenladinischen und angrenzender Dialekte, 2. Teil*, Strasbourg: Éditions de Linguistique et de Philologie, 2012, 5 vol. Karten (vol. I: 1-202; vol. II: 203-420: vol. III: 421-635; vol. IV: 636-850; V: 851-1066), 2 vol. Indizes (Volumen supplementarium 174 S.; Index generalis, 213 S.).

Bauer, Roland (2009): *Dialektometrische Einsichten. Sprachklassifikatorische Oberflächenmuster und Tiefenstrukturen im lombardisch-venedischen Dialektraum und in der Rätoromnia,* San Martin de Tor.

Beer, Heidemarie/Staudinger, Agnes (2012): *Die praktische Applikation der für den ALD-II entwickelten EDV-Werkzeuge.* [Online] 2012. [Zitat vom 12. 18. 2012.] http://www2.hu-berlin.de/vivaldi/tagung/folien/Staudinger_Beer.pdf.

Mell, Peter M./Grance, Timothy (2011): *SP 800-145. The NIST Definition of Cloud Computing*. Gaithersburg, MD: National Institute of Standards & Technology.

Zur Dialektometrisierung des ALD (I und II): Ein Arbeits- und Erfahrungsbericht 2000–2012

Roland Bauer, Universität Salzburg

Im Zentrum des Beitrags steht der im Jahr 2012 erschienene zweite Teil des dolomitenladinischen Sprachatlasses (ALD-II). Dieser umfasst, aufgeteilt auf fünf großformatige Kartenbände, insgesamt 1.066 Sprachkarten, die den Bereichen Lexikon, Morphologie und Syntax gewidmet sind. Wie bereits für den auf die Phonetik zentrierten ersten Teil des ALD geschehen, sollen nun auch sämtliche auf den ALD-II-Karten publizierten Dialektmaterialien analysiert oder taxiert werden, wobei dieser Taxierungsertrag in eine Datenbank einfließen wird, die ihrerseits die Grundlage für die weitere dialektometrische Bearbeitung darstellt. In diesem Zusammenhang können anhand eines lexikalischen Kleincorpus bereits erste Werkstatt-Ergebnisse in Form ausgewählter Arbeitskarten und provisorischer Ähnlichkeitsprofile gezeigt werden. Die Präsentation stützt sich dabei besonders auf das in Salzburg entwickelte Softwaretool *VDM (Visual DialectoMetry)*.

1 Vorbemerkung

Der dolomitenladinische Sprachatlas ALD ist bereits kurz nach der Publikation der ersten Kartenbände (1998) auch Gegenstand einer detaillierten dialektometrischen Untersuchung geworden. Dem Motto des vorliegenden Tagungsbandes "20 Jahre digitale Sprachgeographie" entsprechend, möchte ich daher mit einem Rückblick auf die in diesem Zusammenhang bisher geleistete Projektarbeit beginnen, dann auf die derzeit laufenden Arbeiten eingehen und schließlich kurz die Zukunftsperspektiven skizzieren.

2 Rückblick (2000-2004)

Im Jahr 2000 startete das Forschungsprojekt ALD-I-DM, das hauptsächlich dank der Förderung durch den österreichischen Wissenschaftsfonds (FWF) durchgeführt werden konnte.[1] Ab dem Jahr 2008 konnten weitere Projektmodule realisiert werden, die u.a. den Einbau standardladinischer Datensätze zum Ziel hatten. Im Frühsommer 2012 wurde nun auch die dialektometrische Bearbeitung der Materialien des ALD-II in Angriff genommen. Für die Durchführung dieser Arbeiten durfte und darf ich dankenswerterweise auf materielle Unterstützung seitens des ladinischen Kulturinstituts *Micurà de Rü* in St. Martin in Thurn (Südtirol), seitens des österreichischen Unterrichtsministeriums in Wien und seitens der Universität Salzburg zurückgreifen.

Arbeitstechnisch ging es in der ersten Projektphase um die Auswertung von nahezu allen, also weit über 800 originalen Sprachatlaskarten, die in den vier Bänden des ALD-I publiziert worden waren. Wie man der entsprechenden Übersichtskarte entnehmen kann,[2] sind in dieser sprachgeographischen Dokumentation neben den drei Kerngebieten der Rätoromania (also neben Graubünden, Dolomitenladinien und Friaul) auch die Dialektlandschaften der Lombardei, des Trentino und des Veneto erfasst. Zudem wurden im Rahmen der Dialektometrisierung an dieses 217 Messpunkte bzw. deren Dialekte umfassende Untersuchungsnetz drei so genannte Kunstpunkte angedockt, die die Standardsprachen Französisch, Italienisch und Dolomitenladinisch repräsentieren. Ende des Jah-

[1] FWF-Projekt Nr. P14566-G01.
[2] Siehe dazu Karte 1.

res 2009 konnten die gesammelten Ergebnisse dieses Projektteils in Form einer Monographie veröffentlicht werden.[3]

Bevor ich auf ausgewählte Aspekte daraus eingehe, möchte ich einige grundlegende Bemerkungen zur Methode der Dialektometrie machen, deren Namen in seiner französischen Ausprägung *(dialectométrie)* bekanntlich auf den Tolosaner Dialektologen Jean Séguy (1973) zurückgeht, und deren methodologischer Ausbau seit den späten 1970er Jahren untrennbar mit dem Namen Hans Goebl verbunden ist.[4] Nach einer gängigen Definition versteht man unter Dialektometrie eine Kombination aus Sprachgeographie und numerischer Klassifikation. Das zentrale Forschungsinteresse ist so ausgelegt, dass durch die detaillierte Analyse und durch die kollektive Vermessung einer möglichst großen Anzahl von in aller Regel aus Sprachatlanten stammenden Karten bzw. Daten Raummuster und Arealstrukturen sichtbar gemacht werden, die sonst in den Massendaten tausender Atlaskarten verborgen blieben. Das im Rahmen der eben angesprochenen Kartenanalyse zu erstellende Corpus umfasst für eine Anzahl N von Objekten (hier von sprachlichen Messpunkten bzw. Ortsdialekten) eine Anzahl p von Attributen (hier sprachlicher Merkmale bzw. so genannter Arbeitskarten). Sämtliche auf der Basis dieser so genannten Datenmatrix aufsetzenden Verrechnungsmethoden sind dabei aus der numerischen Klassifikation bekannt. Auf der Seite der Ergebnispräsentation eröffnet sich die Möglichkeit, durch den Einsatz verschiedener klassifikatorischer Verfahren und unterschiedlicher Visualisierungsoptionen jeweils andere Aspekte derselben räumlichen "Ur"-Struktur sichtbar zu machen.

Dabei wird (zumindest in der hier beschriebenen österreichischen Ausprägung bzw. der so genannten Salzburger "Schule"[5] der Dialektometrie) nach folgender Verfahrenskette vorgegangen:[6] Zunächst müssen die originalen Sprachatlaskarten nach bestimmten innerlinguistischen Kriterien vermessen oder, wie wir im Fachjargon sagen, taxiert werden. Betrachten wir zunächst ein Beispiel für eine lexikalisch ausgerichtete Taxierung.[7] Dabei werden jene 217 auf der originalen Sprachatlaskarte mit dem Titel *l'amico* "der Freund" eingetragenen Antwort-*Tokens* zu vier onomasiologischen bzw. etymologischen Antwort-*Types* zusammengefasst, so dass etwa Antworten wie [amí], [amíko] oder [amís] unter den vlat. Typ AMÍCU subsummiert werden können, während die dialektalen Bezeichnungen [kọmpáñ], [kompáño] oder [kumpáni] zu unserem Typ 2, mlat. COMPÁNIO, zusammenfallen. Daneben finden wir hier noch den weniger frequenten SÓCIU- und den nur einmal auftretenden COLLÉGA-Typ. Wie man der räumlichen Verteilung anhand der Arbeitskarte entnehmen kann, dominiert der Typ AMÍCU mit 188 Okkurrenzen eindeutig und ist als Mehrheitstyp nicht nur in weiten Teilen unseres Beobachtungsraums, sondern auch in den drei durch eingefärbte Kreise repräsentierten hochsprachlichen Kunstpunkten zu finden, während der zweithäufigste, gelb signierte COMPÁNIO-Typ nur 18-mal auftritt und dabei auf die nördliche Dolomitenladinia und vereinzelt auf zentraltrentinische Dialekte beschränkt bleibt. Mit dieser lexikalischen Arbeitskarte tritt uns ein erstes, evi-

[3] Cf. Bauer 2009.
[4] Cf. Goebl 1975.
[5] Der Terminus der dialektometrischen "Schule" wurde von mir 2003 (33) erstmals eingeführt und hat sich seither bezeichnungstechnisch etabliert, um etwa die Salzburger Schule von anderen dialektometrischen Forschungsausrichtungen und -zentren (Groningen/Niederlande, Georgia/USA) zu unterscheiden (cf. dazu auch Bauer 2009, 85).
[6] Siehe dazu Abb. 1.
[7] Siehe dazu Karte 2.

denterweise noch sehr oberflächliches Raummuster gegenüber, das in seiner heuristischen Aussagekraft auf eine Ebene mit etwa einer traditionellen Isoglossenkarte zu stellen ist. Man denke beispielsweise an die gut bekannten Wortkarten in Gerhard Rohlfs' *Romanischer Sprachgeographie*.[8]

Originale Sprachatlaskarten können natürlich auch nach anderen innerlinguistischen Kriterien analysiert werden, wie dies anhand unseres zweiten Beispiels exemplifiziert wird.[9] Es geht dabei um eine phonetisch ausgerichtete Taxierung der ALD-Karte *la voce* "die Stimme", wobei im vorliegenden Fall nur der konsonantische Anlaut V- aus lateinisch VÓCE berücksichtigt wird. Wie wir der Arbeitskarte entnehmen können, kommt es auch hier wieder zu einer Reduktion von 220 Antwort-*Tokens* auf vier Antwort-*Types*, wobei zwei phonetische Typen dominieren, nämlich 1. jener mit Beibehaltung des lat. Anlautfrikativs wie etwa bei italienisch [vǫ́će] oder venedisch [vǫ́ze] (mit 103fachem Auftreten) und 2. jener mit Aphärese des Anlautfrikativs wie etwa in ladinisch [ūš], ostlombardisch [us] oder solandrisch [ǫš], der 99 Mal vorkommt. Auch bei dem auf Karte 3 erkennbaren Raummuster handelt es sich um nicht mehr (aber auch um nicht weniger) als eine von potentiell tausenden Oberflächenstrukturen mit einer jeweils eigenen Geschichte, die uns im Rahmen der gerade behandelten sprachlichen Merkmale auf den einzelnen Arbeitskarten gegenübertreten. Die zuletzt gezeigte Taxierung stellt im Übrigen nur eine von vier Möglichkeiten dar, aus der Originalkarte *la voce* phonetische Arbeitskarten zu ziehen. Genauso gut kann man auch die Entwicklung des Haupttons -Ó-, des nachtonigen Konsonantismus oder des Auslauts -E aus lat. VÓCE untersuchen und somit vier verschiedene phonetische Analysen mit einer einzigen Karte durchführen.

Aus dem bisher Gesagten wird ersichtlich, dass die Anzahl der im Rahmen der Taxierung erstellten Analysen bzw. Arbeitskarten jene der originalen Sprachatlaskarten deutlich übersteigt. Im Falle meines Projekts kommen auf jede der insgesamt 845 ausgewerteten ALD-I-Karten im Schnitt fünf Arbeitskarten, wobei immer nur Erstantworten berücksichtigt wurden, und jede Merkmalsausprägung minimal raumbildend sein, also zumindest an drei Messpunkten vorkommen sollte. Zudem wurden nur jene Originalkarten taxatorisch behandelt, die zumindest zu 210 der insgesamt 217 ALD-Messpunkte Informationen liefern konnten.

Auf diesen Prämissen aufbauend ergeben sich im Projekt ALD-DM bisher folgende numerische Strukturen: das auf der Auswertung des ALD-I basierende Corpus verfügt über gut 4.300 Einzelanalysen, die sich auf folgende systemlinguistische Bereiche verteilen: 70% der Arbeitskarten betreffen die Phonetik, 18% das Lexikon und 12% aller Taxierungen fallen in den Bereich der Morphosyntax. Diese datenseitige Schieflage ergibt sich aus der Grundkonzeption bzw. aus der thematischen Zweiteilung des ladinischen Sprachatlasses und soll durch die in den nächsten Jahren anstehende Dialektometrisierung des ALD-II ausgeglichen werden. Da jede unserer Arbeitskarten Information zu insgesamt 220 Untersuchungsobjekten aufweist, verfügt die entsprechende Datenmatrix über knapp 950.000 Informationseinheiten.

Der nächste im Rahmen der dialektometrischen Verfahrenskette zu setzende Schritt betrifft die Verwandlung der oben angesprochenen Datenmatrix in eine Ähnlichkeitsmatrix. Dabei wird, häufig unter Einsatz des so genannten *Relativen Identitätswertes*

8 Cf. Rohlfs 1971.
9 Siehe dazu Karte 3.

RIW, die relative Anzahl jener sprachlichen Eigenschaften ermittelt, die zwei miteinander verglichene Dialekte (bzw. deren in der Datenmatrix gespeicherte Ortsvektoren) gemeinsam aufweisen. Der solcherart errechnete Ähnlichkeitswert RIW liegt immer zwischen minimal 0 und maximal 100%. Im Rahmen dieser Ähnlichkeitsmessung werden alle N Objekte mit den übrigen $N–1$ Objekten verglichen, so dass schlussendlich zu jedem unserer 220 Ortsdialekte 219 Vergleichswerte zur Verfügung stehen.[10]

Das bisher geschilderte Procedere kann anhand des in Abb. 1 aufscheinenden Schaubildes nochmals nachvollzogen werden. Auf die Auswahl des zu untersuchenden Corpus oder Taxandums A folgt 2. die Wahl eines entsprechenden Messverfahrens, um eine Datenmatrix B zu erzeugen. Aus dieser wird 3. nach Wahl des gewünschten Ähnlichkeitsmaßes die Ähnlichkeitsmatrix (oder eine dazu komplementäre Distanzmatrix) C erzeugt, die 4. ihrerseits als Basis der Visualisierung der taxometrischen Ergebnisse D fungiert. Eine Möglichkeit der damit angesprochenen Ergebnispräsentation besteht nun in der Erstellung so genannter Ähnlichkeitskarten.[11]

Alle in der Folge gezeigten Abbildungen wurden übrigens mit einem sehr mächtigen Software-Tool namens *VDM (Visual DialectoMetry)* generiert, das in den Jahren 1998/99 von unserem ehemaligen Kollegen Edgar Haimerl entwickelt worden war und zu dem seither eine Reihe von *Upgrades* zur Verfügung stehen. Für die Erstellung der vorliegenden Arbeit verwende ich die *VDM*-Version 1.10.5.0 aus dem Jahr 2011.[12]

2.1 Dialektometrische Ähnlichkeitskarten

Das erste Fallbeispiel betrifft die räumliche Verteilung der zwischensprachlichen Ähnlichkeiten von unserem standarditalienischen Kunstpunkt aus gesehen.[13] Dieser wird als so genannter Prüfbezugspunkt, also als jener Ort bezeichnet, dessen interdialektale Ähnlichkeiten mit allen übrigen Vergleichsobjekten visualisiert werden, und ist dabei auf der Karte selbst als nicht weiter eingefärbtes Kreissymbol im Süden unseres Netzes abgebildet, auf das durch einen roten Pfeil verwiesen wird. Alle übrigen Orte bzw. Ortsdialekte sind auf der Ähnlichkeitskarte durch nach dem Sonnenspektrum eingefärbte Flächen (Polygone oder Kreise) repräsentiert, wobei die Aufteilung der Messwerte in sechs verschiedene Farbklassen bestimmten Algorithmen gehorcht, die sich meist an den Polwerten Maximum und Minimum sowie am Mittelwert oder am Median orientieren.[14] Besonders warme Farben stehen für hohe Ähnlichkeiten mit dem Prüfbezugspunkt, hier also mit dem Italienischen. Dies trifft im vorliegenden Beispiel etwa auf die rot eingefärbten Zonen des Veneto zu, deren Dialekte bis knapp 77% Affinität zur italienischen Standardsprache aufweisen. Auf der anderen Seite der Messwerteskala finden wir die mit kalten Farben signierten Gebiete, deren Dialekte in großer Distanz zum Italienischen stehen. Damit ist in unserem Fall der blau eingefärbte Sprachraum des Rätoromanischen angesprochen, dessen drei Teilgebiete (Bündnerromanisch im Westen, Dolomitenladinisch im Zentrum und Friaulisch im Osten) kollektiv auf Distanz zum Italienischen gehen und diesem nur mehr zu 39 bis knapp 47% ähnlich sind. Von dieser Distanzierung ist

[10] Zum methodisch-theoretischen Hintergrund der Ähnlichkeitsmessung cf. Bauer 2009, 91-101.
[11] Siehe Eintrag D3 auf Abb. 1.
[12] Zu *VDM* cf. auch op.cit., 201-205.
[13] Siehe dazu Karte 4.
[14] Für weitere Details cf. op.cit., 102-105.

im Übrigen auch das im Nordwesten unseres Beobachtungsraums platzierte Standardfranzösische betroffen.

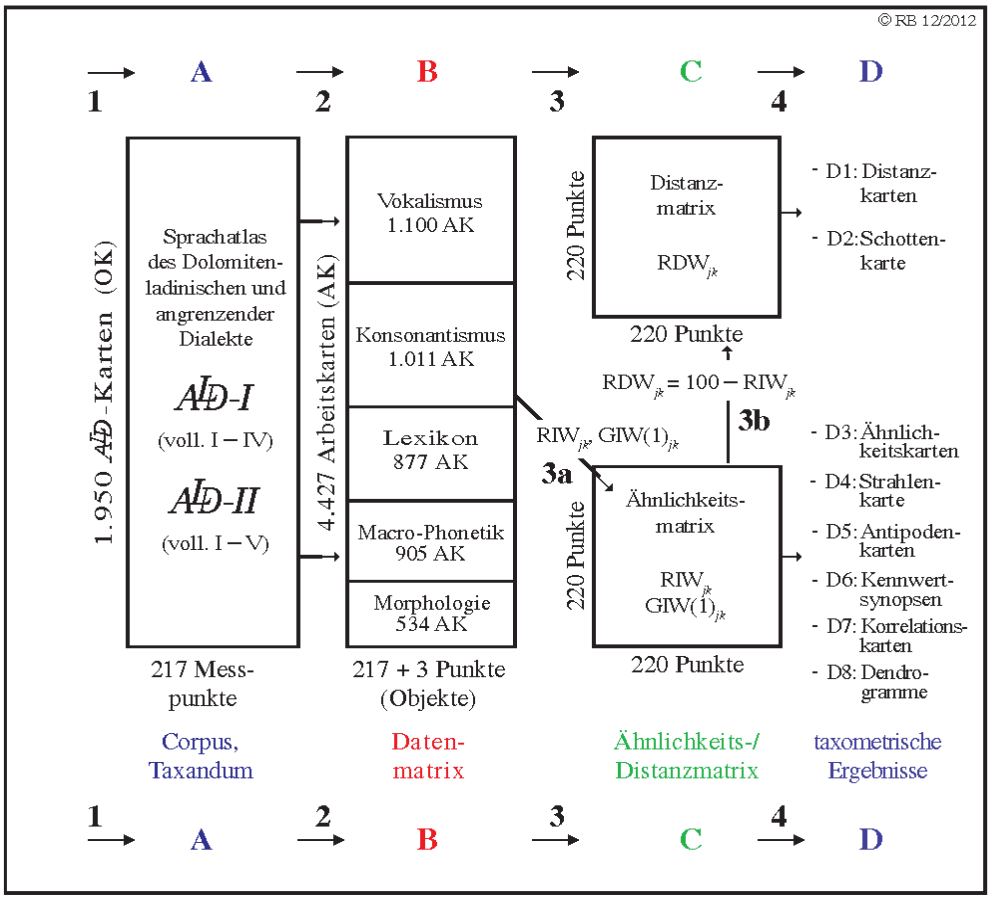


Abb. 1: Die dialektometrische Verfahrenskette (mit Bezug auf das Projekt *ALD-DM*: Stand 12/2012)

1 = Wahl des Sprachatlasses → A = Corpus, Taxandum →
2 = Wahl des Messverfahrens → B = Datenmatrix →
3 = Wahl des Ähnlichkeitsmaßes → C = Ähnlichkeitsmatrix →
4 = Wahl des Klassifikationsalgorithmus → D = taxometrische Ergebnisse

Abb. 1: Die dialektometrische Verfahrenskette (mit Bezug auf das Projekt ALD-DM: Stand 12/2012)

Eine ähnliche Strukturierung unseres Beobachtungsraumes, bei der die Polarisierung in einen nicht rätoromanischen Südteil und einen rätoromanischen, dem restlichen Sprachgebiet besonders unähnlichen, blau eingefärbten Nordteil deutlich ins Auge fällt, ergibt sich nun jedes Mal, wenn wir unseren Prüfbezugspunkt in den Dialektbereich der Lombardei, des Trentino oder des Veneto legen, also die Ähnlichkeitsrelationen nach wie vor von außerhalb der Rätoromania betrachten.[15] Dies gilt selbst dann noch, wenn der Prüfbezugspunkt bereits sehr nahe bzw. unmittelbar am rätoromanischen Sprachgebiet

[15] Siehe dazu Karte 5 mit drei unterschiedlichen, räumlich jeweils weit von der Rätoromania entfernten Prüfbezugspunkten.

zu liegen kommt,[16] wie anhand eines alpinlombardischen Profils, anhand einer exemplarischen nordvenedischen Ähnlichkeitskarte und auch anhand eines Profils ersichtlich wird, das aus einer politisch-territorial gesehen zwar noch friaulischen, sprachlich-dialektal gesehen jedoch bereits venedischen Position heraus erstellt wurde. In allen drei Fällen stehen die Dialekte der Prüfbezugspunkte den räumlich zum Teil bereits weit entfernten Dialekten Oberitaliens bedeutend näher als dem in unmittelbarer Nähe gesprochenen Rätoromanisch, welcher Ausprägung auch immer.

Als erstes Zentralergebnis der Dialektometrisierung des ALD-I ist also festzuhalten, dass alle nicht-rätoromanischen Dialekte des Untersuchungsgebietes gleichsam systematisch auf Distanz zum Rätoromanischen gehen und dabei allen übrigen, in unserem Raumausschnitt erfassten Dialekten näher stehen, als dem Bündnerromanischen, dem Dolomitenladinischen und/oder dem Friaulischen. Vor allem die auf Karte 6 figurierenden Profile suggerieren durch das direkte Nebeneinander von rot und blau eingefärbten Polygonen, also dem Prüfbezugspunkt besonders ähnlichen und besonders unähnlichen Dialekten, das Vorhandensein sprachlicher Barrieren bzw. die Existenz von Sprachgrenzen. Auch die Herausarbeitung solcher mehr oder weniger linearer Strukturen kann mithilfe der Dialektometrie bewerkstelligt werden.

2.2 Quantitative Isoglossenkarten

Die eben angesprochene Abschottung der rätoromanischen von den benachbarten oberitalienischen Mundarten kann mit einem anderen dialektometrischen Heuristikum gut abgebildet werden, das vom Software-Paket *Visual DialectoMetry* bereit gehalten wird. Es handelt sich dabei um die quantitative Isoglossenkarte oder Schottenkarte, die datenseitig auf die bereits angesprochene Distanzmatrix zurückgreift, dabei allerdings nur jene Distanzwerte ausliest, die unmittelbar benachbarte Dialekte betreffen.[17]

Je größer die innerlinguistische Distanz zwischen zwei Nachbardialekten, desto größer fällt auch die Strichstärke des entsprechenden Lineaments (i.e. der Schotte) aus, wobei die Farbgebung diesmal so gestaltet ist, dass kalte Farben für große Distanz und warme Farben für geringe Distanz stehen. Die auf der Schottenkarte[18] deutlich sichtbaren, blau eingefärbten Polygonseiten im Süden der Bündnerromania, im Bereich der Dolomitenladinia und im Westen Friauls markieren also massive Isoglossenbündelungen, die, wie wir der Legende von Karte 7 entnehmen können, auf sprachlichen Unterschieden zwischen rund 42 und knapp 55% beruhen. In anderen Worten bedeutet dies, dass es beim Vergleich aller direkt benachbarten Dialekte auf der Basis von insgesamt 4.300 berücksichtigten Sprachmerkmalen im Extremfall zu unter 50% liegenden Übereinstimmungen kommt, dass also räumlich unmittelbar benachbarte Dialekte äußerst große sprachliche Distanz aufweisen. Dies kann bzw. muss als Hinweis auf die Existenz sprachlicher Barrieren gelesen werden und gilt auf unserer Karte in erster Linie für die Abschottung des Bündnerromanischen gegenüber dem südlich davon vorgelagerten Alpinlombardischen. Ein ähnlicher, wenn auch quantitativ weniger stark ausgeprägter Bruch verläuft auch mitten durch die Ladinia, ein Umstand, den wir kartographisch besser her-

16 Siehe dazu Karte 6 mit drei unterschiedlichen, räumlich jeweils nahe bei der Rätoromania liegenden Prüfbezugspunkten.

17 Siehe dazu Abb. 1, Block C, Eintrag D1; zur Funktionsweise der Schottenkarte und zur Problematik der Eruierung von Sprachgrenzen cf. ferner Bauer 2009, 117-124.

18 Siehe Karte 7.

ausarbeiten können, wenn wir den Beobachtungsraum auf das Dolomitengebiet einschränken.[19]

Dies wird durch relativ einfache Datenbankabfragen bewerkstelligt und stellt somit eine Möglichkeit dar, kleinräumig relevante Arealstrukturen besser aufzuzeigen. Durch eine dicke, blau eingefärbte Polygonseite repräsentiert zeigt sich auf Karte 8 zunächst ein mächtiges Isoglossenbündel, das die beiden dolomitenladinischen Talschaftsvarietäten Gadertalisch (Val Badia) und Ampezzanisch (Anpezo) voneinander trennt und das dabei auf gut 64% innerlinguistischer Distanz beruht. Ferner wird auf unserer Ausschnittkarte der deutliche Bruch zwischen der deutsch überdachten Nordladinia und der mit einem italienischen Sprachdach versehenen Südladinia sichtbar. Die blauen Schotten repräsentieren in diesem Bereich sprachliche Unterschiede von rund 46%.

3 Laufende Arbeiten (2008-2012)

3.1 Standardladinisch

Diese zuletzt angesprochenen sprachlichen Divergenzen zwischen Nord und Süd sind im metasprachlichen Bewusstsein der ladinischsprachigen Bevölkerung deutlich verankert und spielen somit auch in sprachpolitischen Zusammenhängen eine nicht unwesentliche Rolle. Dies gilt beispielsweise für den Versuch, alle fünf bzw. sechs dolomitenladinischen Talschaftsvarietäten (Ennebergisch, Gadertalisch, Grödnerisch, Fassanisch, Buchensteinisch, Ampezzanisch) mit einer gemeinsamen, *Ladin Dolomitan* oder *Ladin Standard* genannten Schriftsprache zu überdachen. Diese wurde zwar bereits vor gut zehn Jahren im Rahmen eines breit angelegten Projekts am Schreibtisch ausgearbeitet und anhand einer Grammatik und eines Wörterbuchs auch entsprechend kodifiziert,[20] ihre Einführung in ausgewählten Bereichen scheiterte jedoch bislang am politischen Willen mancher Verantwortlicher bzw. an der oftmals allzu kampanilistischen Grundhaltung wesentlicher Protagonisten. In diesem Zusammenhang werden dem Neostandard *Ladin Dolomitan* bisweilen zu starke Ähnlichkeiten mit der Nachbarvarietät nachgesagt, während die Kritiker die eigene Varietät im Standard meist unterrepräsentiert finden und durch dessen Einführung einen sprachlichen Identitätsverlust herannahen sehen.[21]

Vor diesem Hintergrund habe ich mich im Jahr 2008 entschlossen, das bestehende dialektometrische Corpus durch einen standardladinischen Datensatz zu ergänzen, um die Ähnlichkeiten und Unähnlichkeiten der in den Tälern gesprochenen Dialekte mit der neuen Schriftsprache messen und somit möglichst objektiv herausarbeiten zu können.

Ein erstes Ergebnis zeigt sich auf einem dialektometrischen Ähnlichkeitsprofil des gesamten Beobachtungsraums aus der Sicht des Prüfbezugspunkts *Ladin Dolomitan*, der als weißes Kreissymbol in der westlichen Peripherie Ladiniens aufscheint und die Messpunktnummer 777 trägt.[22] Es wird ersichtlich, dass das durch eine gepunktete Liniensignatur abgegrenzte ladinische Sprachgebiet mit bis zu 80% Übereinstimmung erwartungsgemäß als besonders standardaffin in Erscheinung tritt, wohingegen die anderen beiden hochsprachlichen Kunstpunkte, nämlich Französisch im Westen und Italienisch

[19] Siehe Karte 8.
[20] Cf. SPELL 2001 und 2002.
[21] Cf. dazu besonders Bauer 2012, passim.
[22] Siehe Karte 9.

im Süden eindeutig als Antipoden des Ladinischen auftreten und diesem nur mehr zu 40 bis 49% ähnlich sind. Von dieser Distanz werden auch weite Teile des – laut Auskunft von Karte 4 – besonders italo-affinen, venedischen Sprachraums erfasst.

Wirft man nun einen genaueren Blick auf die interne Strukturierung der Dolomitenladinia aus der Sicht der neuen Schriftsprache,[23] so fällt die bereits auf der quantitativen Isoglossenkarte erkennbare Zweiteilung auch hier gut ins Auge: man erkennt schon auf den ersten Blick einen durch warme Farben repräsentierten, besonders standardnahen Norden und einen in kälteren Farben gehaltenen, deutlich standardferneren Süden. Die Unterschiede sind dabei beträchtlich: der Dialekt des oberen Gadertals etwa ist, auf der Basis des Gesamtcorpus von hier knapp 3.000 Arbeitskarten, mit annähernd drei Viertel seiner Merkmale im Standard vertreten,[24] während etwa das Ampezzanische oder aber das Unterfassanische dort nur knapp die Hälfte ihrer sprachlichen Besonderheiten repräsentiert sehen.[25] Ein differenzierterer Zugriff auf die Datenbank erlaubt es nun, diese Konstellation auch datenseitig zu hinterfragen.

In diesem Zusammenhang stehen verschiedene innerlinguistische Subcorpora zur Verfügung, die es erlauben, die Ähnlichkeitsverhältnisse bezüglich der Phonetik (getrennt in Vokalismus und Konsonantismus) oder aber jene bezüglich des Lexikons getrennt aufzuzeigen.

Ein erster Vergleich stellt das Gesamtcorpus dem phonetischen Subcorpus, das auf mehr als 2.000 makro- und mikro-phonetischen Taxierungen beruht, gegenüber.[26] Die Spannweite der jeweiligen Ähnlichkeitswerte ist nahezu deckungsgleich und reicht in beiden Fällen von rund 46% minimaler Ähnlichkeit bis zu über 70% maximaler Übereinstimmung. Dies rührt wohl daher, dass das Gesamtcorpus, wie bereits erwähnt, (noch) dominant phonetisch geprägt ist. Auch bezüglich der räumlichen Profilierung fallen keine besonderen Unterschiede ins Auge, wenn man von der Absenz der grün eingefärbten Klasse [3] auf dem phonetischen Ähnlichkeitsprofil (Karte 11) einmal absieht. Dieses Fehlen ist indirekt ein kleiner Hinweis darauf, dass das phonetische Subcorpus deutlicher polarisiert als das Gesamtcorpus, dass es also aus phonetischer Sicht einerseits drei Talschaftsdialekte gibt, die dem Standard überdurchschnittlich nahe stehen, nämlich das Gadertalische der Val Badia, Grödnisch / Gherdëina und Buchensteinisch / Fodóm,[27] während wir es andererseits mit den Dialekten Ampezzos und des Fassatals zu tun haben, die phonetisch gesehen zum Großteil bereits unter 50% Ähnlichkeit aufweisen und somit überaus standardfern erscheinen.[28] Bei dem auf Karte 10 abgebildeten Profil, das die gesamte Datenmatrix berücksichtigt, tritt das Buchensteinische / Fodóm noch als gleichsam zwischen diesen beiden Polen vermittelnd in Erscheinung.

Ob die erwähnte phonetische Polarisierung nun eher durch konsonantische oder aber durch vokalische Merkmale bedingt ist, kann dem Vergleich der beiden folgenden Profile entnommen werden.[29] Bei Betrachtung der Spannweiten aller Ähnlichkeitswerte fällt auf, dass die konsonantischen Affinitäten (Karte 12) stärker streuen und dabei deutlich

[23] Siehe Karte 10.
[24] Siehe dazu die auf Karte 10 rot eingefärbten Polygone.
[25] Siehe dazu die auf Karte 10 blau eingefärbten Polygone.
[26] Siehe Karte 11.
[27] Siehe dazu die auf Karte 11 rot, orange und gelb eingefärbten Polygone der Klassen [4], [5] und [6].
[28] Siehe dazu die auf Karte 11 blau eingefärbten Polygone der Klassen [1] und [2].
[29] Siehe dazu die Karten 12 (Konsonantismus) und 13 (Vokalismus).

über 80% hinausreichen, während der Vokalismus (Karte 13) eine bedeutend geringere Streuung zeigt und dabei maximale Ähnlichkeiten von lediglich 64% erreichen kann. Es gibt also unter den 21 vermessenen Dialekten keinen einzigen, der zumindest zwei Drittel seiner vokalischen Merkmale im Standard wiederfindet. Der Vokalismus der dolomitenladinischen Mundarten scheint aufgrund seiner großen Heterogenität viel schwieriger in einen gemeinsamen Schriftstandard einzubringen zu sein als der Konsonantismus. Diese je nach Talschaftsdialekt bzw. selbst innerhalb der Täler unterschiedliche Ausprägung vokalischer Merkmale tritt dem Betrachter des auf rein vokalischen Merkmalen basierenden Profils von Karte 13 u.a. auch dadurch gegenüber, dass etwa die fünf im unteren Bereich des Gadertals bzw. in Enneberg gelegenen Dialekte (Orte Nr. 81-85) gleich fünf verschiedenen Intervallen bzw. Farbklassen angehören, während sie bei Auswertung des Konsonantismus in lediglich zwei Klassen fallen. Andererseits ist der Vokalismus bedingt durch die relativ geringe Spannweite aller Ähnlichkeitswerte ausgewogener, um nicht zu sagen "gerechter" verteilt als der Konsonantismus. Denn bei letzterem gibt es immerhin zumindest einen Dialekt (nämlich jenen von Kolfuschg im oberen Gadertal), der bezüglich seiner konsonantischen Besonderheiten mit knapp 84% mehr als doppelt so gut abgebildet ist wie der standardfernste Dialekt (also jener von Colle S. Lucia, im Osten Buchensteins) mit unter 42% Ähnlichkeit zum Standardladinischen.[30]

3.2 Taxierung des ALD-II

Die bisher gezeigten Karten sollen nun auch mit einem nach ausschließlich lexikalischen Gesichtspunkten erstellten Ähnlichkeitsprofil verglichen werden. Wie bereits eingangs erwähnt, handelt es sich dabei um einen innerlinguistischen Bereich, der erst im Rahmen der (seit Beginn des Jahres 2012 in Angriff genommenen) Dialektometrisierung des ALD-II zu einem größeren Corpus ausgebaut werden kann. Aus den Originalkarten des ALD-I konnten insgesamt 760 lexikalische Arbeitskarten generiert werden. Darüber hinaus wurden mittlerweile 117 lexikalische Analysen zu ALD-II-Karten erstellt und in die Projektdatenbank eingegeben, so dass wir derzeit (Dezember 2012) über insgesamt 877 einschlägige Arbeitskarten verfügen.

Auf einer Ähnlichkeitskarte zum gesamten Untersuchungsgebiet des ALD werden die innerhalb unseres Beobachtungsraums wirksam werdenden lexikalischen Relationen sehr plakativ wiedergegeben.[31] So tritt der gesamte Dolomitenraum (einmal abgesehen von Cortina d'Ampezzo, P. 92, auf dessen Sonderstellung hier nicht weiter eingegangen werden kann) als überaus kompakte Klasse mit 74 bis 82% lexikalischer Ähnlichkeit zum Standardladinischen in Erscheinung, was hier zwar nicht weiter verwundert, was aber wohl auch die Repräsentativität, um nicht zu sagen die "einende Kraft" des neuen Schriftdachs für den gesamten Dolomitenraum unterstreicht.[32]

Bei Verlagerung des Prüfbezugspunktes in einen beliebigen ladinischen Talschaftsdialekt fällt die Profilierung nämlich weitaus weniger kompakt aus. Dies geht aus einigen exemplarischen Ähnlichkeitskarten deutlich hervor. Im Vergleich dazu soll die Kom-

[30] Siehe dazu die mit weißem Gitterraster hinterlegten Polygone der Punkte 89 (Kolfuschg, rot) und 93 (Colle S. Lucia, blau) auf Karte 12.

[31] Siehe Karte 14.

[32] Siehe dazu die auf Karte 14 rot und orange eingefärbten Polygone der Punkte 81-91 und 93-101 (Werteklassen [5] und [6]). Von dieser Standardnähe sind auch die dem Ladinischen besonders ähnlichen, räumlich jedoch bereits knapp außerhalb der historischen Dolomitenladinia gesprochenen Mundarten von Rocca Pietore (P. 138) und Laste (P. 139) betroffen.

paktheit der Ladinia aus standardladinischer Sicht stets im Blickfeld behalten werden. Das erste Kontrastprofil stammt aus Gröden, wobei zunächst (metaphorisch gesprochen) nur der "Verlust" des Fassanischen aus der ladinischen Kerngruppe auffällt.[33] Wenn wir den Beobachtungspunkt dorthin, also ins Fassatal legen, zerfällt die ladinische Kompaktheit völlig und wir stehen einer Zersplitterung von 21 Ortsdialekten auf fünf Klassen gegenüber.[34] Noch deutlicher fällt die innerladinische Distanzierung ins Auge, wenn man die lexikalischen Ähnlichkeiten aus der Sicht des Ampezzanischen betrachtet.[35] Dieses orientiert sich aus lexikalischer Sicht überhaupt nicht mehr in Richtung Dolomitenraum, der ampezzanische Wortschatz erscheint vielmehr an den im Süden vorgelagerten, periladinischen Raum des nördlichen Veneto angebunden und geht gleichzeitig auf deutliche Distanz zum Lexikon der nördlichen Dolomitenladinia. Allen vier gezeigten Profilen ist jedoch gemeinsam, dass sich das jeweilige Ladinische (ob nun Standard oder realiter gesprochener Talschaftsdialekt) immer besonders deutlich vom Französischen, aber auch vom Bündnerromanischen unterscheidet.[36]

Kehren wir diesbezüglich nochmals zu dem auf Karte 14 abgebildeten standardladinischen Ähnlichkeitsprofil zurück. Die dort sichtbar werdende Kompaktheit des Dolomitenraums wurde bereits ausreichend kommentiert. Andererseits setzen sich hier sowohl das Französische als auch (und das ist durchaus überraschend) das Bündnerromanische mit Ähnlichkeitswerten um die 50% sehr deutlich vom *Ladin Dolomitan* ab, während das Italienische mit gut 60% Ähnlichkeit dem Standardladinischen doch bedeutend näher steht.[37]

Bezogen auf die so genannte "questione ladina", also auf die bekanntlich bereits seit dem Beginn des 20. Jahrhunderts schwelende Streitfrage, ob denn das Rätoromanische als eigene Sprachgruppe existiere und dabei Bündnerromanisch, Dolomitenladinisch und Friaulisch umfasse, ist dieses Resultat insofern nicht ohne Brisanz, als dadurch erkennbar wird, dass sich die Näherelationen zwischen den drei genannten rätoromanischen Schwesteridiomen offensichtlich markant verschieben können, wenn man den Merkmalsraum der Phonetik verlässt und die Analyse beispielsweise auf den Wortschatz einschränkt.[38]

[33] Siehe dazu das auf Karte 15 oben abgebildete Profil zum Prüfbezugspunkt St. Christina (P. 87), auf dem die fassanischen Dialekte (PP. 97-101) nicht mehr in die ladinischen "Top-Klassen" (rot, orange) fallen.

[34] Siehe dazu das auf Karte 15 in der Mitte abgebildete Profil zum Prüfbezugspunkt Moncion (P. 99).

[35] Siehe dazu das auf Karte 15 unten abgebildete Profil zum Prüfbezugspunkt Cortina d'Ampezzo (P. 92).

[36] Siehe dazu die auf den Karten 14 und 15 blau eingefärbten Polygone bzw. den entsprechend signierten Kreis (P. 888) im Nordwesten des Beobachtungsraumes.

[37] Siehe dazu den auf Karte 14 gelb signierten Kreis (P. 999) am Südrand des Beobachtungsraumes.

[38] Wie uns hier nicht weiter in Betracht gezogene, dialektometrische Korrelationskarten (cf. Bauer 2009, 143-148) zeigen, betrifft die klassifikationstechnisch wirksam werdende Corpusabhängigkeit im Raumausschnitt des ALD v.a. das Westfriaulische, innerhalb der Dolomitenladinia gilt dies in erster Linie für das Fassanische und den Neo-Standard. – Bezüglich der Relevanz der jeweils berücksichtigten Corpora verweise ich ferner auf den bei dieser Tagung gehaltenen Vortrag von Roseanu (et al., *Razones históricas en la mezcla prosódica del catalán de l'Alguer*), der bei der clusteranalytischen Verarbeitung systemlinguistisch unterschiedlicher Subcorpora zum sardischen Sprachraum ähnliche Erfahrungen gemacht und dabei u.a. festgestellt hat, dass eine ausschließlich auf Prosodie-Daten beruhende Klassenbildung anders ausfällt und somit andere Relationen im Raum aufzeigt (nämlich die gemeinsame Klassifizierung des Katalanischen von Alghero, des Sardischen und des auf Sardinien gesprochenen Regionalitalienischen in einem einzigen Cluster), als das bei Verrechnung eines phonetischen, lexikalischen und/oder morphologischen Corpus der Fall ist.

Vor diesem auch forschungspolitisch relevanten Hintergrund darf ich dem Ausbau unseres lexikalischen Corpus mit besonders großem Interesse entgegensehen. Davon abgesehen wird es bei der Dialektometrisierung des ALD-II in den nächsten Jahren vordringlich um die Erweiterung des bisher nur sehr marginal behandelten morphologischen Corpus[39] sowie um den Aufbau eines völlig neuen syntaktischen Subcorpus gehen. All das sollte es ermöglichen, die Klassifikation der oberitalienischen und der rätoromanischen Mundarten corpusseitig abzurunden, zu komplettieren und somit datenseitig auf noch sicherere Beine zu stellen.

Bibliographie

ALD-I = Goebl, Hans / Bauer, Roland / Haimerl, Edgar (eds.) (1998): *Atlant linguistich dl ladin dolomitich y di dialec vejins, 1a pert / Atlante linguistico del ladino dolomitico e dei dialetti limitrofi. 1a parte / Sprachatlas des Dolomitenladinischen und angrenzender Dialekte. 1. Teil*, Wiesbaden.

ALD-II = Goebl, Hans et al. (eds.) (2012): *Atlant linguistich dl ladin dolomitich y di dialec vejins, 2a pert / Atlante linguistico del ladino dolomitico e dei dialetti limitrofi. 2a parte / Sprachatlas des Dolomitenladinischen und angrenzender Dialekte. 2. Teil*, Straßburg.

Bauer, Roland (2003): *Dialektometrische Analyse des Sprachatlasses des Dolomitenladinischen und angrenzender Dialekte (ALD-I)*, Salzburg, [Habilitationsschrift].

Bauer, Roland (2009): *Dialektometrische Einsichten. Sprachklassifikatorische Oberflächenmuster und Tiefenstrukturen im lombardo-venedischen Dialektraum und in der Rätoromania*, San Martin de Tor.

Bauer, Roland (2012): "Wie ladinisch ist *Ladin Dolomitan*? Zum innerlinguistischen Naheverhältnis zwischen Standardsprache und Talschaftsdialekten", in: *Ladinia* XXXVI, 205-335.

Goebl, Hans: "Dialektometrie", in: *Grazer Linguistische Studien* 1 (1975), 32-38.

Rohlfs, Gerhard (1971): *Romanische Sprachgeographie. Geschichte und Grundlagen, Aspekte und Probleme mit dem Versuch eines Sprachatlas der romanischen Sprachen*, München.

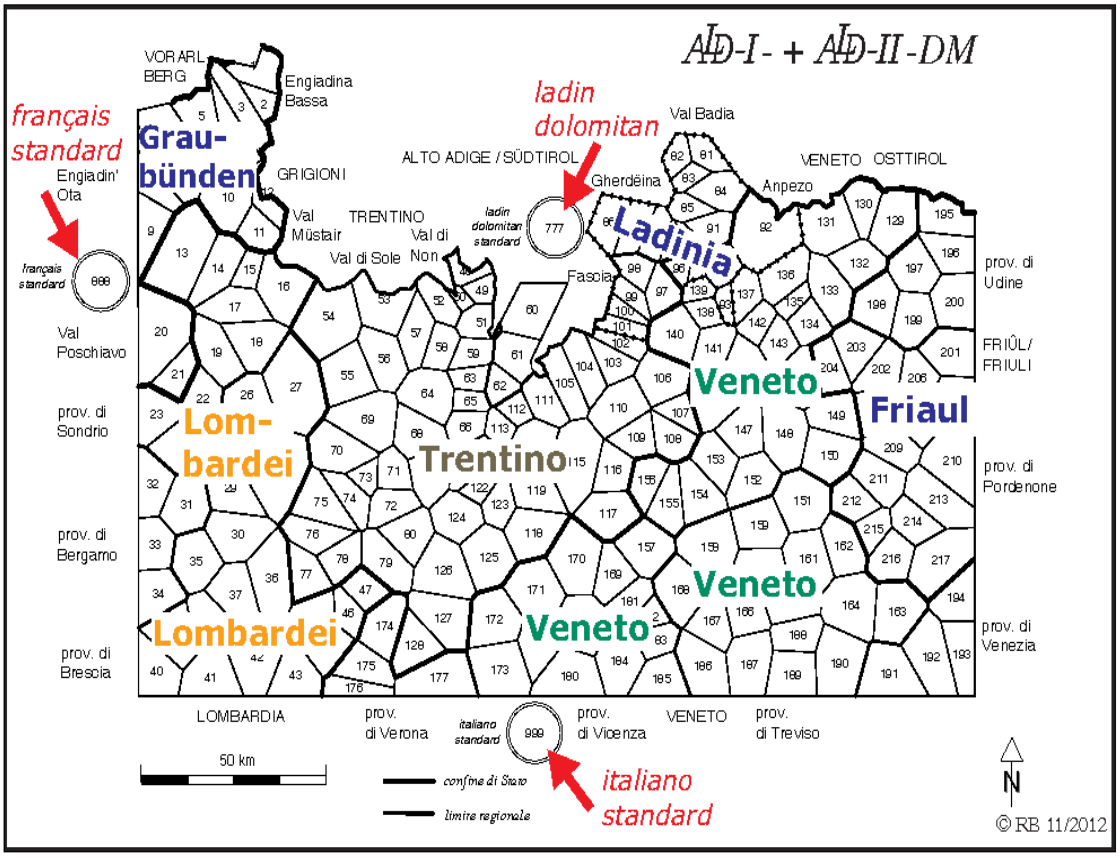
Rührlinger, Brigitte (2012/13): *Fonetica, morfologia e morfosintassi verbale dei dialetti lombardi nord-orientali nel loro contesto geolinguistico*, Salzburg (Fachbereich Romanistik); [unveröffentlichte Dissertation].

Séguy, Jean (1973): "La dialectométrie dans l'Atlas linguistique de la Gascogne", in: *Revue de Linguistique Romane* 37, 1-24.

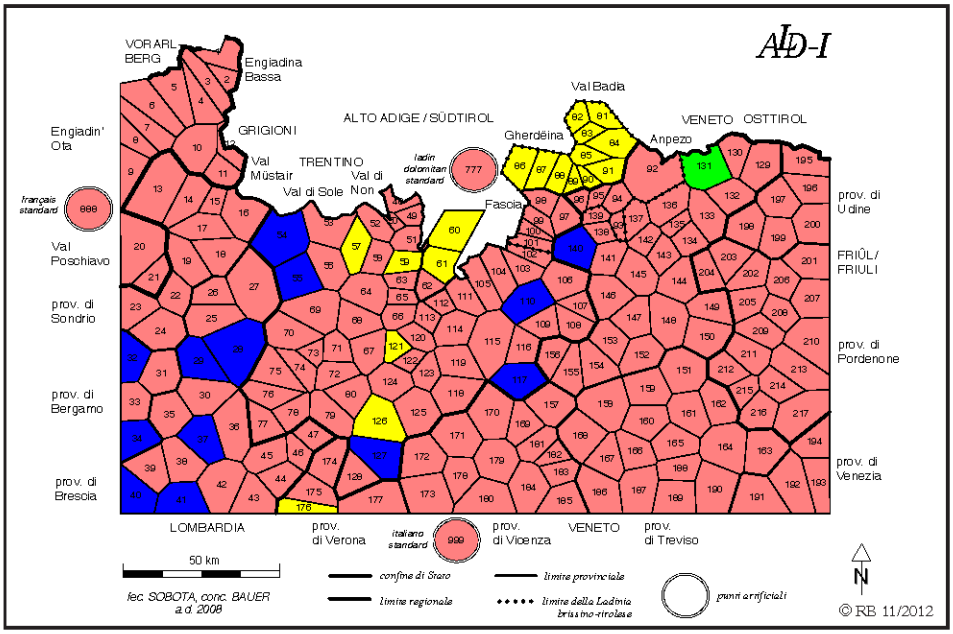
SPELL (Servisc de Planificazion y Elaborazion dl Lingaz Ladin) (2001): *Gramatica dl Ladin Standard*, Vich/San Martin de Tor/Bulsan.

SPELL (Servisc de Planificazion y Elaborazion dl Lingaz Ladin) (2002): *Dizionar dl Ladin Standard*, Vich/San Martin de Tor/Bulsan.

[39] Dazu gibt es bereits umfangreiche Vorarbeiten, die im Rahmen einer an der Universität Salzburg erstellten Dissertation geleistet wurden (cf. Rührlinger 2012/13).



Karte 1: Beobachtungsraum ALD-DM: 217 Messpunkte (24.500 km^2), drei Kunstpunkte



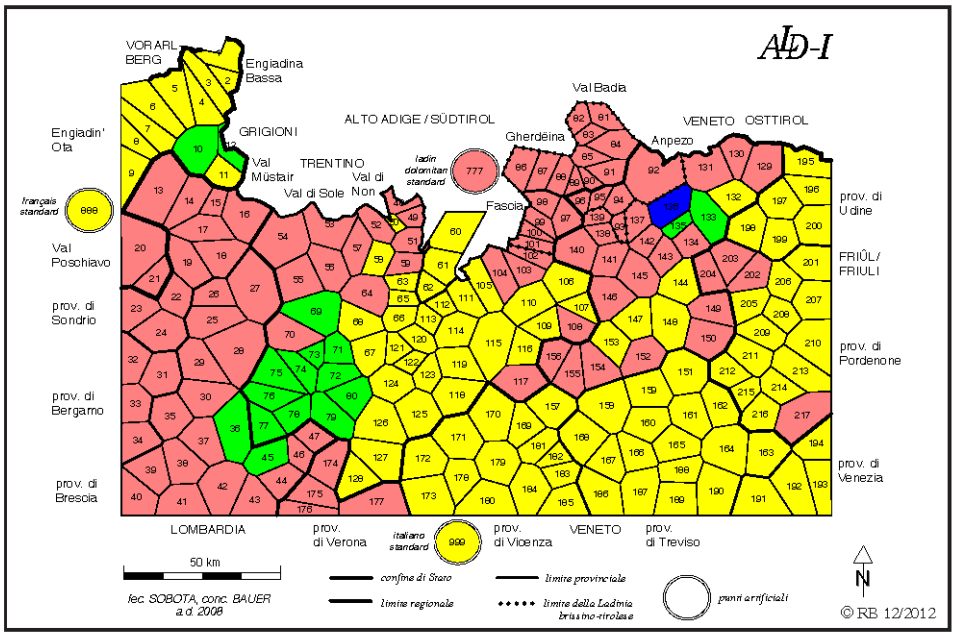
Typ ÁMICU: [*amí*], [*amíco*], [*amís*] (188)

Typ SÓCIU: [*ṣǫ́ći*], [*ṣǫ́ćo*] (13)

Typ COMPÁNIO: [*kǫmpáñ*], [*kompáño*], [*kumpáni*] (18)

Typ COLLÉGA: [*kolę́ga*] (1)

Karte 2: Lexikalische Arbeitskarte zur Originalkarte ALD-I-25 *l'amico*



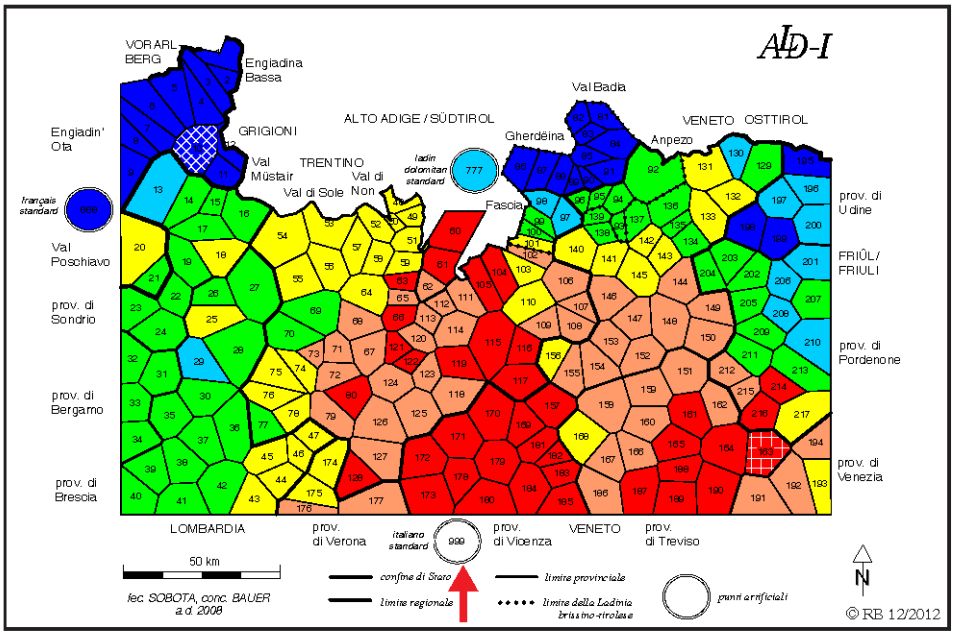
Typ "zéro": [*ūš*], [*us*], [*o̱s*] (99)

Typ *v-*: [*vo̱će*], [*vo̱ze*] (103)

Typ *g-*: [*go̱š*], [*gūš*], [*gu̱o̱s*] (18)

Typ *γ-*: [*γu̱o̱s*] (1)

Karte 3: Phonetische Arbeitskarte zur Karte ALD-I-873 *la voce*, Entwicklung des Anlauts V̲ÓCE



Legende

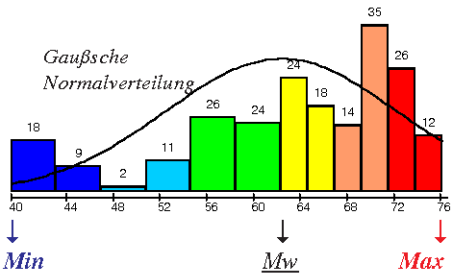
MMinMwMaxX 6-fach, nach $RIW_{999,k}$

[1]	≥ **39,23** – 46,95	n = 27
[2]	> 46,95 – 54,67	n = 13
[3]	> 54,67 – 62,38	n = 50
[4]	> 62,38 – 67,08	n = 42
[5]	> 67,08 – 71,78	n = 49
[6]	> 71,78 – **76,48**	n = 38

Summe: 219

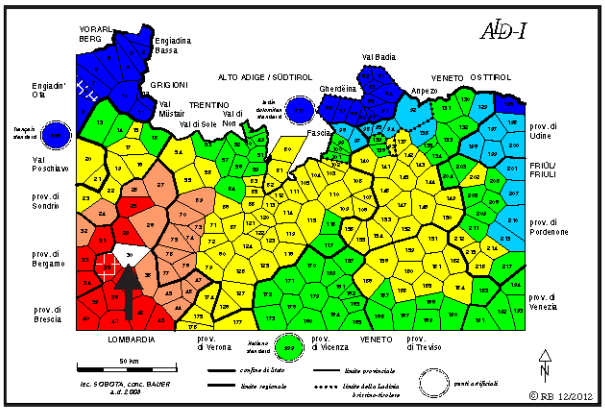
Histogramm der Ähnlichkeitsverteilung

MMinMwMaxX 12-fach, nach $RIW_{999,k}$



Prüfbezugspunkt	Kunstpunkt 999, *italiano standard* (siehe roter Pfeil)
Messmoment	Relativer Identitätswert ($RIW_{999,k}$)
Datenmatrix	N = 220 Orte, p = 4.310 Arbeitskarten
Intervallalgorithmus	*MMinMwMaxX* mit 6 Intervallen (6 Farbstufen)
Kreissymbole	Kunstpunkte

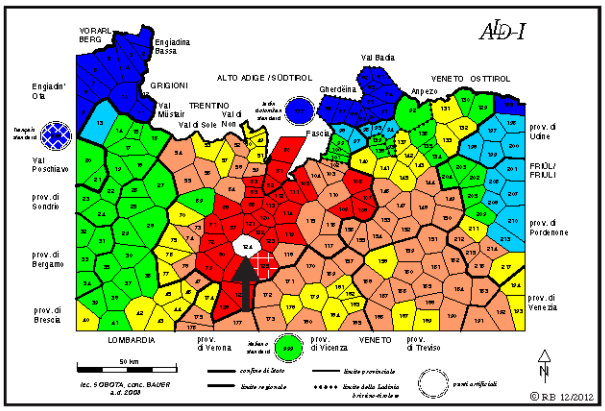
Karte 4: Ähnlichkeitskarte zum Standarditalienischen (Gesamtcorpus, Gesamtnetz)



lombardischer
Prüfbezugspunkt:

AIS-P. 30 Breno

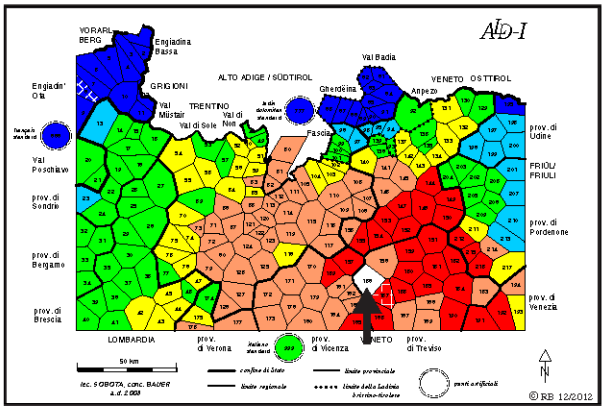
(siehe weißes Polygon und schwarzer Pfeil)



trentinischer
Prüfbezugspunkt:

AIS-P. 124 Aldeno

(siehe weißes Polygon und schwarzer Pfeil)



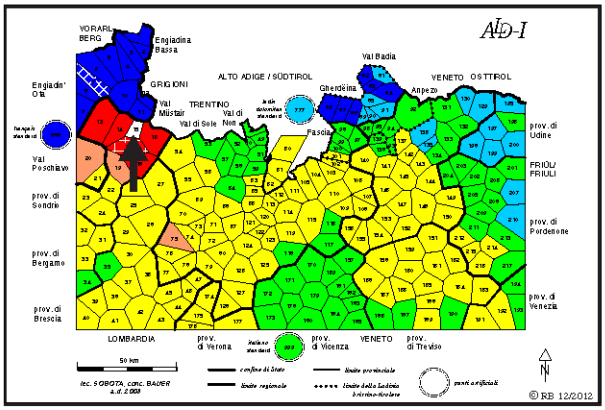
venedischer
Prüfbezugspunkt:

AIS-P. 168 Crespano

(siehe weißes Polygon und schwarzer Pfeil)

drei verschiedene Prüfbezugspunkte, jeweils räumlich weit von der Rätoromania entfernt, jeweils große sprachliche Distanz des **Rätoromanischen** (siehe **blaue Polygone**)

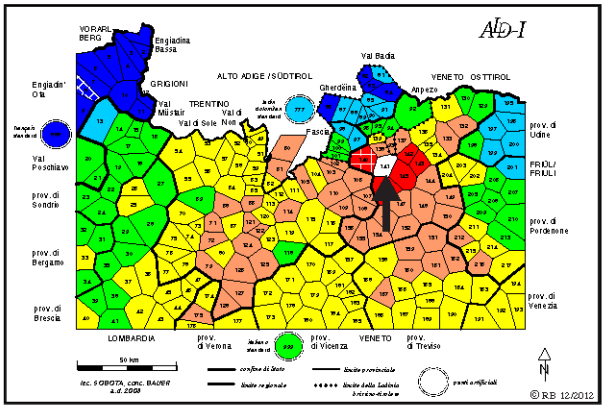
Karte 5: Oberitalienische Ähnlichkeitskarten 1
(räumliche Distanz + sprachliche Distanz)



alpinlombardischer
Prüfbezugspunkt:

ALD-P. 15 Bormio

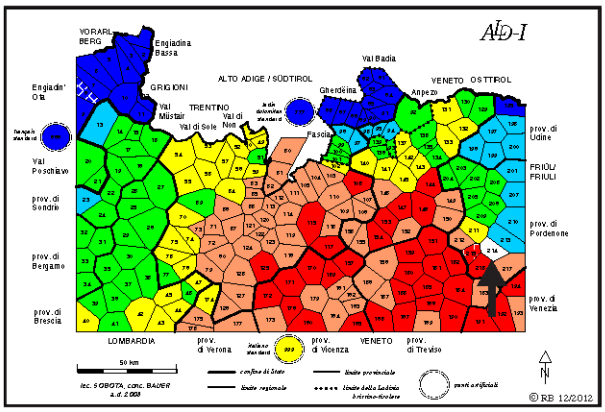
(siehe weißes Polygon und schwarzer Pfeil)



nord-venedischer
Prüfbezugspunkt:

ALD-P. 141 Cencenighe

(siehe weißes Polygon und schwarzer Pfeil)



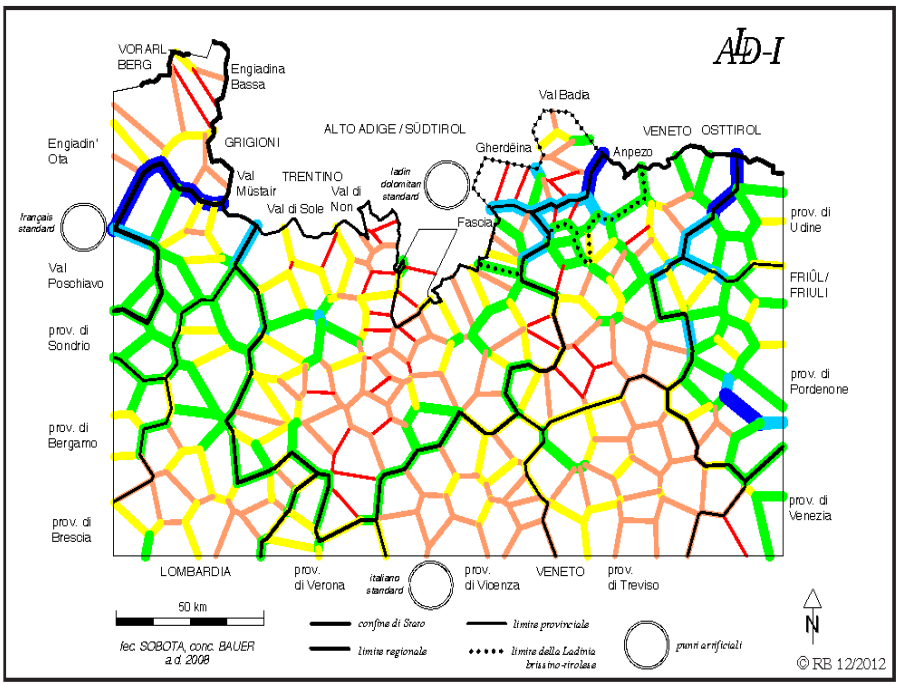
west-friaulischer
Prüfbezugspunkt:

ALD-P. 214 Pordenone

(siehe weißes Polygon und schwarzer Pfeil)

drei verschiedene Prüfbezugspunkte, räumlich jeweils nahe bei der Rätoromania gelegen, jeweils große sprachliche Distanz des **Rätoromanischen** (siehe **blaue Polygone**)

Karte 6: Oberitalienische Ähnlichkeitskarten 2
(räumliche Nähe + sprachliche Distanz)



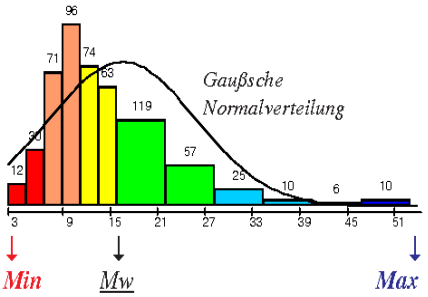
Legende

MinMwMax 6-fach, nach $RDW_{j,k}$

[1]	> 41,71 – **54,27**	n = 16
[2]	> 29,15 – 41,71	n = 35
[3]	> 16,59 – 29,15	n = 176
[4]	> 11,75 – 16,59	n = 137
[5]	> 6,92 – 11,75	n = 167
[6]	≥ **2,09** – 6,92	n = 42
		Summe: 573

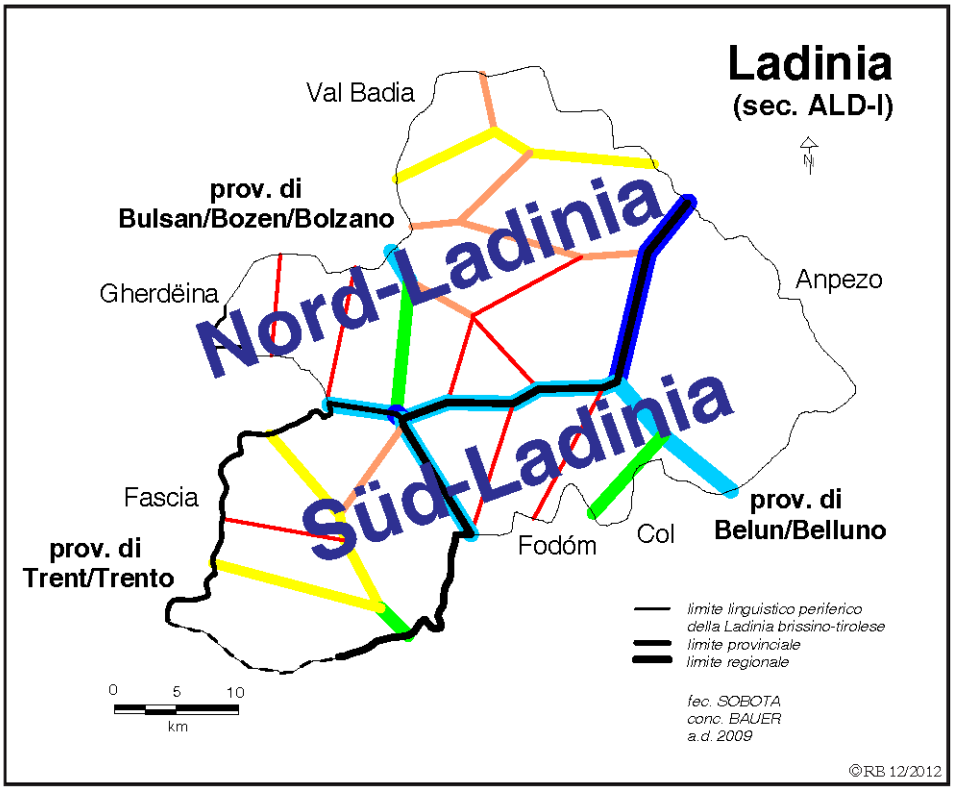
Histogramm der Ähnlichkeitsverteilung

MinMwMax 12-fach, nach $RDW_{j,k}$



Messmoment	Relativer Distanzwert ($RDW_{j,k}$)
Datenmatrix	N = 220 Orte, p = 4.310 Arbeitskarten, n = 573 Schotten
Intervallalgorithmus	*MinMwMax* mit 6 Intervallen (6 Farbstufen)
Strichstärke	diskret variierend, 6 Ausprägungen, maximal 150 Pixel
Kreissymbole	Kunstpunkte (hier unberücksichtigt)

Karte 7: Quantitative Isoglossenkarte / Schottenkarte zum ALD-I (Gesamtcorpus, Gesamtnetz)



Legende

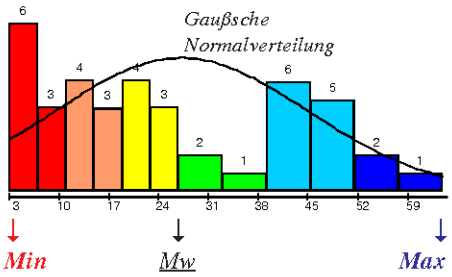
MinMwMax 6-fach, nach RDW_{jk}

[6]	> 52,08 – **64,50**	n =	3
[5]	> 39,66 – 52,08	n =	11
[4]	> 27,24 – 39,66	n =	3
[3]	> 19,16 – 27,24	n =	7
[2]	> 11,08 – 19,16	n =	7
[1]	≥ **3,00** – 11,08	n =	9

Summe: 40

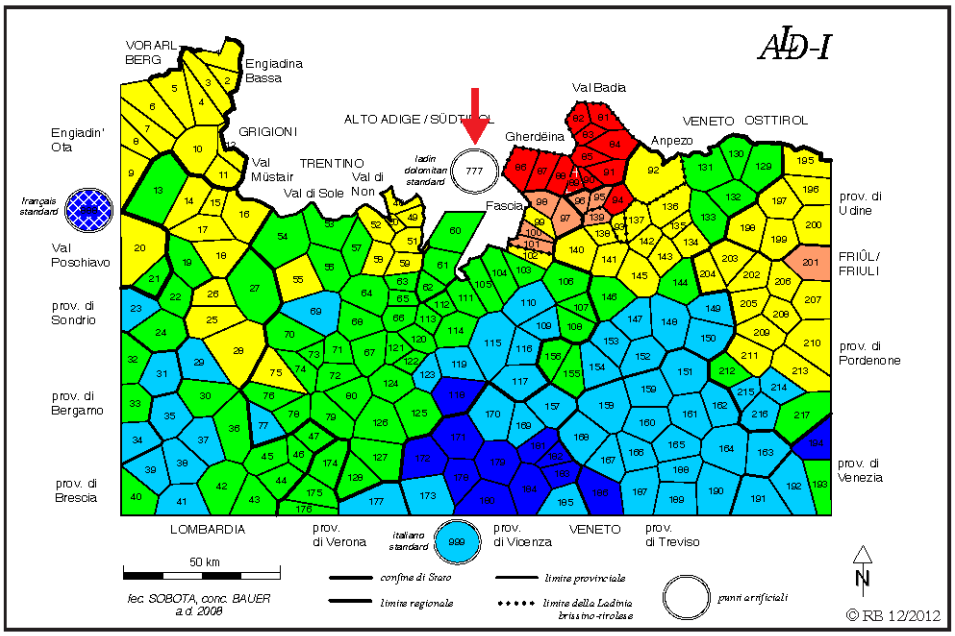
Histogramm der Ähnlichkeitsverteilung

MinMwMax 12-fach, nach RDW_{jk}



Messmoment	Relativer Distanzwert (RDW_{jk})
Datenmatrix	N = 21 Orte
	p = 2.969 Arbeitskarten (Gesamtcorpus)
	n = 40 Schotten
Strichstärke	diskret variierend, 6 Ausprägungen, maximal 170 Pixel
Intervallalgorithmus	*MinMwMax* mit 6 Intervallen (6 Farbstufen)

Karte 8: Quantitative Isoglossenkarte zum ALD-I (Gesamtcorpus, Teilnetz Ladinien)



Legende

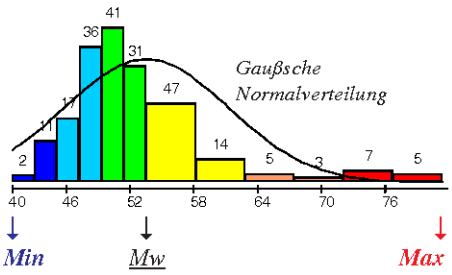
MMinMwMaxX 6-fach, nach $RIW_{777,k}$

[1]	≥ **40,93** – 45,14	n = 13
[2]	> 45,14 – 49,36	n = 53
[3]	> 49,36 – 53,57	n = 72
[4]	> 53,57 – 62,96	n = 61
[5]	> 62,96 – 72,35	n = 8
[6]	> 72,35 – **81,74**	n = 12

Summe: 219

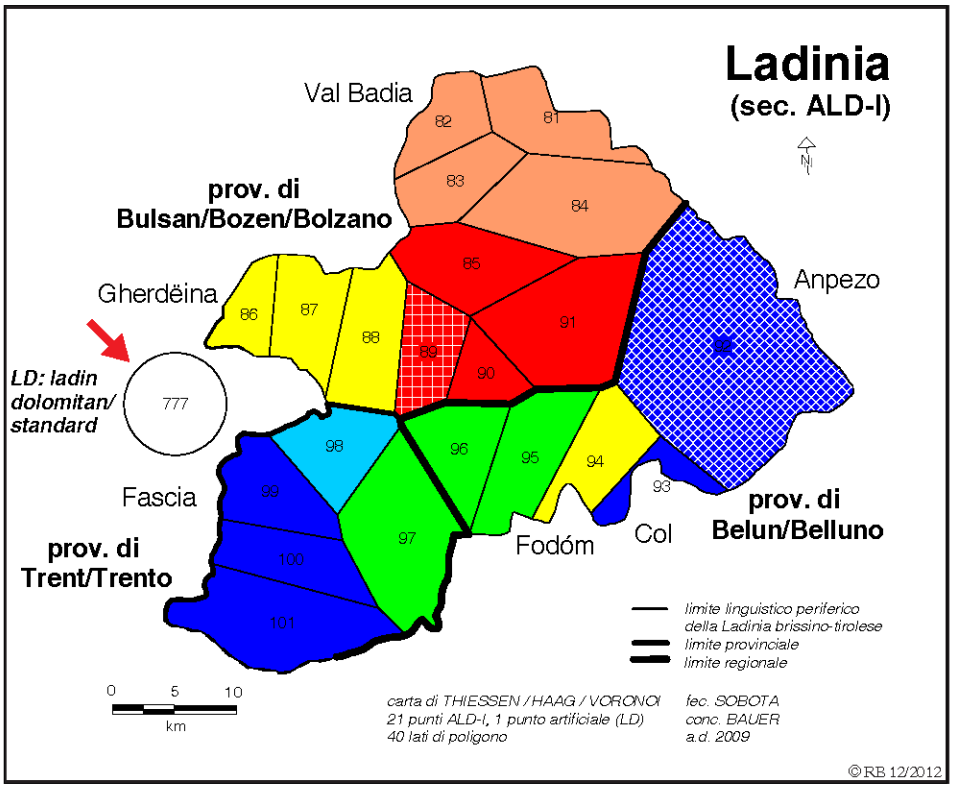
Histogramm der Ähnlichkeitsverteilung

MMinMwMaxX 12-fach, nach $RIW_{777,k}$



Prüfbezugspunkt	Kunstpunkt 777, *Ladin Dolomitan* (siehe roter Pfeil)
Messmoment	Relativer Identitätswert ($RIW_{777,k}$)
Datenmatrix	N = 220 Orte, p = 4.310 Arbeitskarten
Intervallalgorithmus	*MMinMwMaxX* mit 6 Intervallen (6 Farbstufen)
Kreissymbole	Kunstpunkte

Karte 9: Ähnlichkeitskarte zum Standardladinischen (Gesamtcorpus, Gesamtnetz)



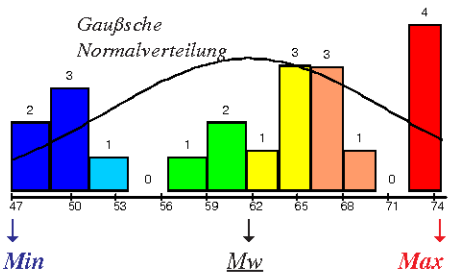
Legende

MMinMwMaxX 6-fach, nach $RIW_{777,k}$

[1]	≥ **46,05** – 51,30	n =	5
[2]	> 51,30 – 56,55	n =	1
[3]	> 56,55 – 61,79	n =	3
[4]	> 61,79 – 66,13	n =	4
[5]	> 66,13 – 70,47	n =	4
[6]	> 70,47 – **74,80**	n =	4
		Summe:	21

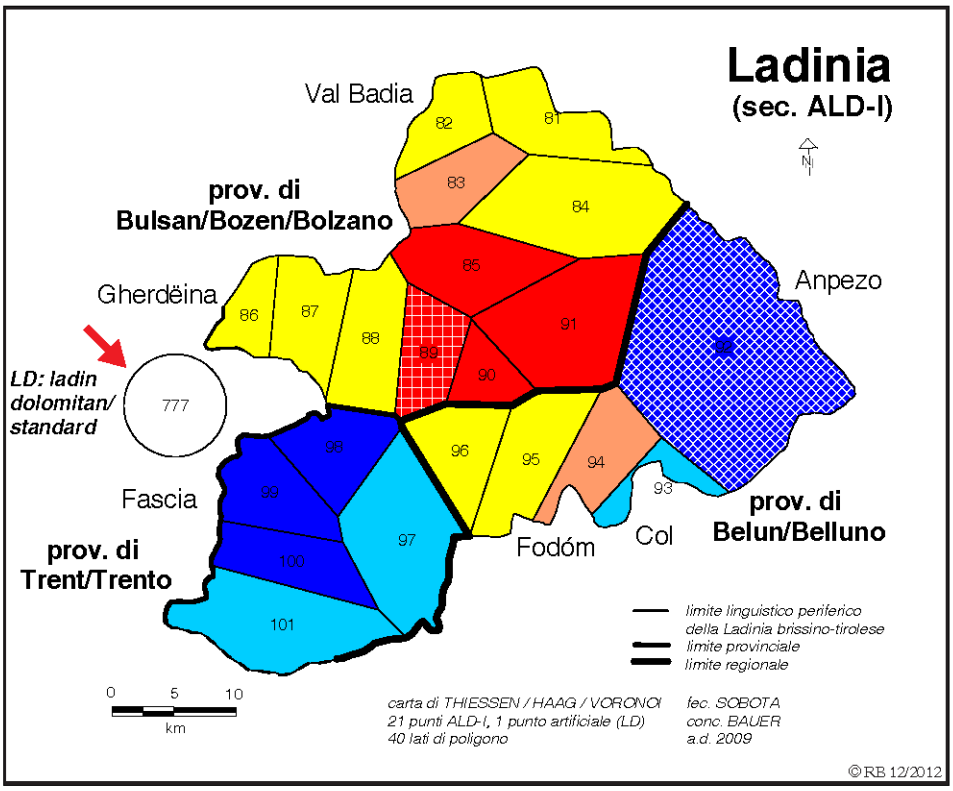
Histogramm der Ähnlichkeitsverteilung

MMinMwMaxX 12-fach, nach $RIW_{777,k}$



Prüfbezugspunkt	Kunstpunkt 777, *Ladin Dolomitan* (siehe roter Pfeil)
Messmoment	Relativer Identitätswert ($RIW_{777,k}$)
Datenmatrix	N = 22 Orte, p = 2.969 Arbeitskarten
Intervallalgorithmus	*MMinMwMaxX* mit 6 Intervallen (6 Farbstufen)

Karte 10: Ähnlichkeitskarte zum Standardladinischen (Gesamtcorpus, Teilnetz Ladinien)



Legende

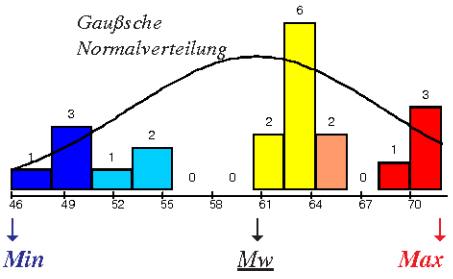
MMinMwMaxX 6-fach, nach $RIW_{777,k}$

[1]	≥ **45,80** – 50,76	*n* =	4
[2]	> 50,76 – 55,73	*n* =	3
[3]	> 55,73 – 60,69	*n* =	0
[4]	> 60,69 – 64,51	*n* =	8
[5]	> 64,51 – 68,32	*n* =	2
[6]	> 68,32 – **72,14**	*n* =	4

Summe: 21

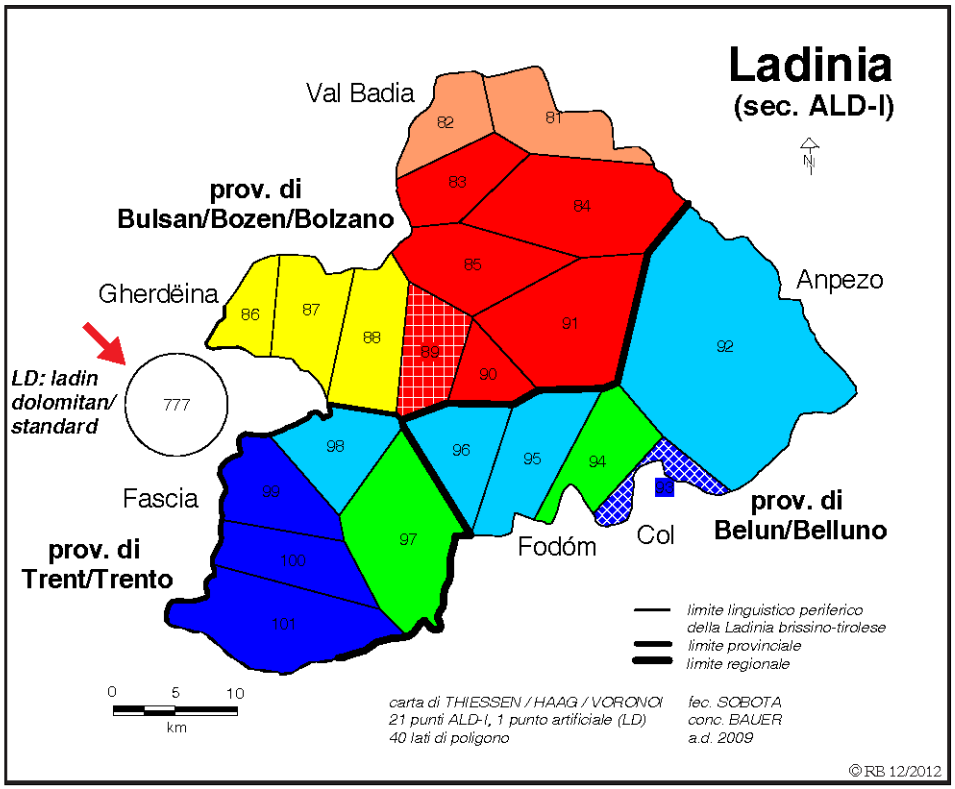
Histogramm der Ähnlichkeitsverteilung

MMinMwMaxX 12-fach, nach $RIW_{777,k}$



Prüfbezugspunkt	Kunstpunkt 777, *Ladin Dolomitan* (siehe roter Pfeil)
Messmoment	Relativer Identitätswert ($RIW_{777,k}$)
Datenmatrix	*N* = 22 Orte, *p* = 2.037 Arbeitskarten (Phonetik)
Intervallalgorithmus	*MMinMwMaxX* mit 6 Intervallen (6 Farbstufen)

Karte 11: Ähnlichkeitskarte zum Standardladinischen (Macro- und Microphonetik, Teilnetz Ladinien)



Legende

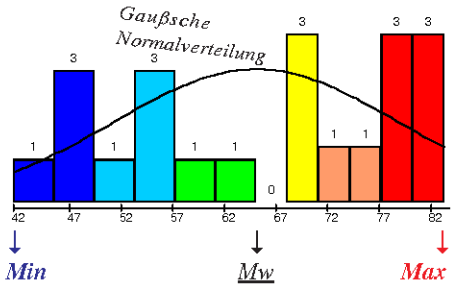
MMinMwMaxX 6-fach, nach $RIW_{777,k}$

[1]	≥ **41,56** – 49,40	*n* =	4
[2]	> 49,40 – 57,24	*n* =	4
[3]	> 57,24 – 65,08	*n* =	2
[4]	> 65,08 – 71,27	*n* =	3
[5]	> 71,27 – 77,45	*n* =	2
[6]	> 77,45 – **83,63**	*n* =	6

Summe: 21

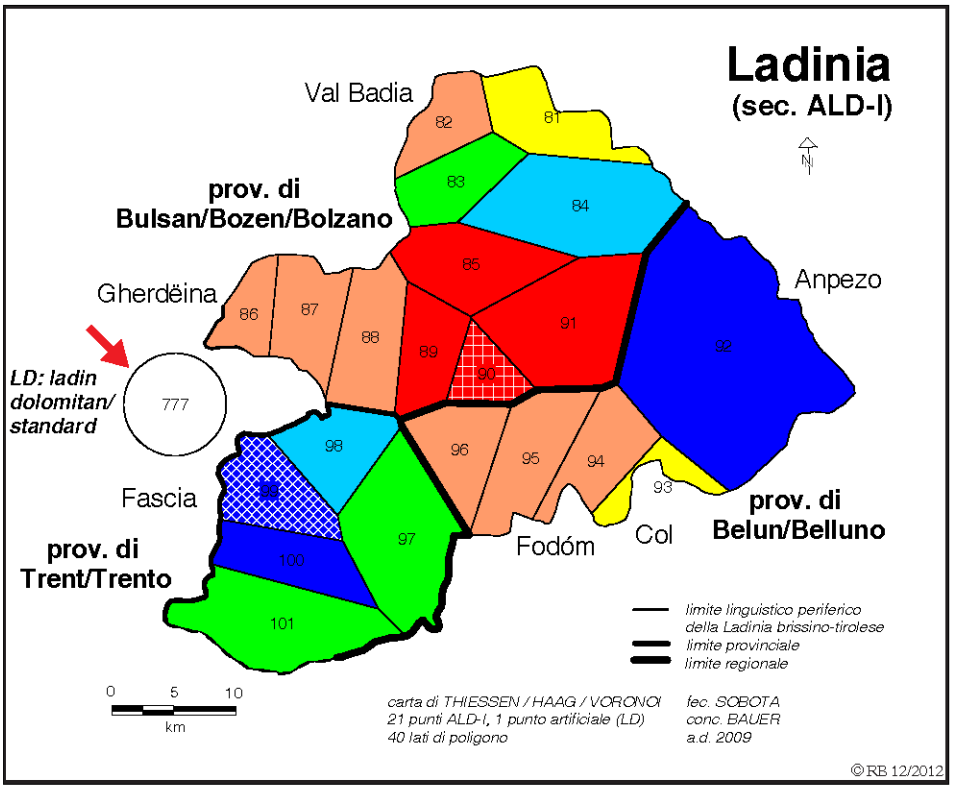
Histogramm der Ähnlichkeitsverteilung

MMinMwMaxX 12-fach, nach $RIW_{777,k}$



Prüfbezugspunkt	Kunstpunkt 777, *Ladin Dolomitan* (siehe roter Pfeil)
Messmoment	Relativer Identitätswert ($RIW_{777,k}$)
Datenmatrix	*N* = 22 Orte, *p* = 789 Arbeitskarten (Konsonantismus)
Intervallalgorithmus	*MMinMwMaxX* mit 6 Intervallen (6 Farbstufen)

Karte 12: Ähnlichkeitskarte zum Standardladinischen (Phonetik/Konsonantismus, Teilnetz Ladinien)



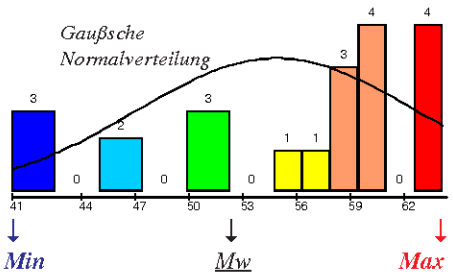
Legende

MMinMwMaxX 6-fach, nach $RIW_{777,k}$

[1]	≥ **40,11** – 45,04	*n* =	3
[2]	> 45,04 – 49,96	*n* =	2
[3]	> 49,96 – 54,88	*n* =	3
[4]	> 54,88 – 58,05	*n* =	2
[5]	> 58,05 – 61,21	*n* =	7
[6]	> 61,21 – **64,37**	*n* =	6
	Summe:		21

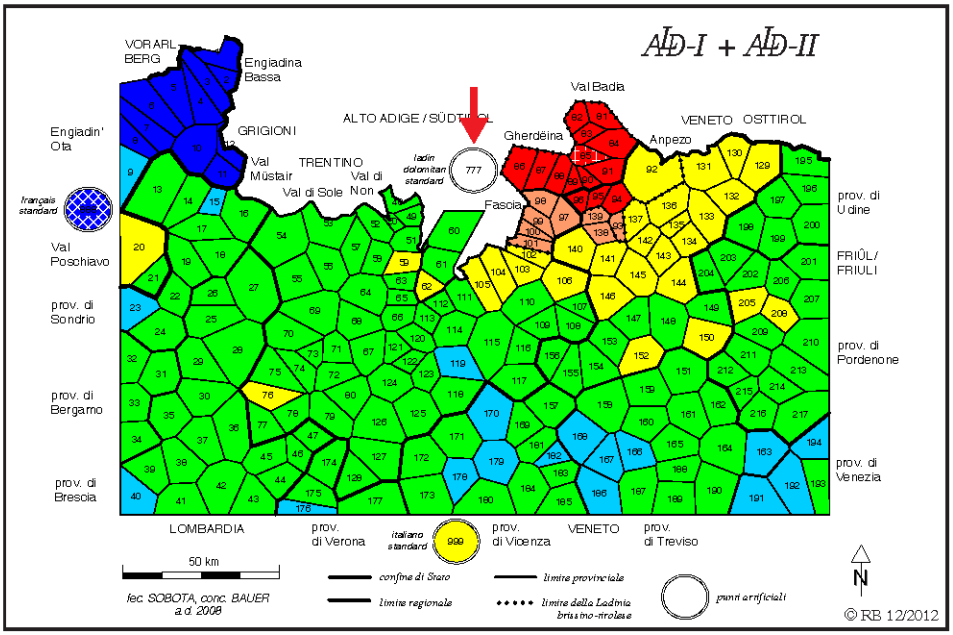
Histogramm der Ähnlichkeitsverteilung

MMinMwMaxX 12-fach, nach $RIW_{777,k}$



Prüfbezugspunkt	Kunstpunkt 777, *Ladin Dolomitan* (siehe roter Pfeil)
Messmoment	Relativer Identitätswert ($RIW_{777,k}$)
Datenmatrix	*N* = 22 Orte, *p* = 896 Arbeitskarten (Vokalismus)
Intervallalgorithmus	*MMinMwMaxX* mit 6 Intervallen (6 Farbstufen)

Karte 13: Ähnlichkeitskarte zum Standardladinischen (Phonetik/Vokalismus, Teilnetz Ladinien)



Legende

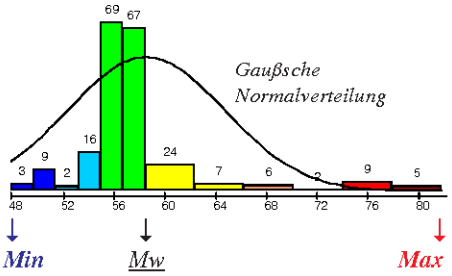
MMinMwMaxX 6-fach, nach $RIW_{777,k}$

[1]		≥ **47,80** – 51,37	n = 12
[2]		> 51,37 – 54,93	n = 18
[3]		> 54,93 – 58,50	n = 136
[4]		> 58,50 – 66,37	n = 31
[5]		> 66,37 – 74,24	n = 8
[6]		> 74,24 – **82,11**	n = 14

Summe: 219

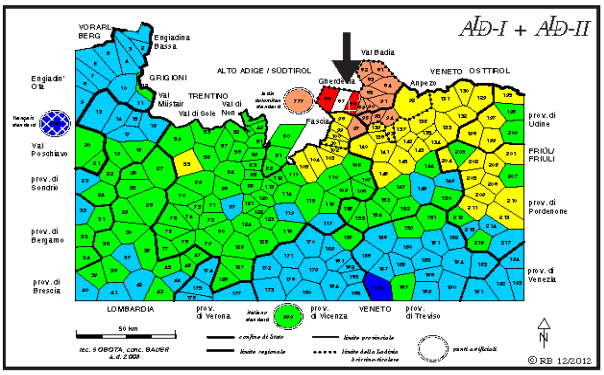
Histogramm der Ähnlichkeitsverteilung

MMinMwMaxX 12-fach, nach $RIW_{777,k}$



Prüfbezugspunkt	Kunstpunkt 777, *Ladin Dolomitan* (siehe roter Pfeil)
Messmoment	Relativer Identitätswert ($RIW_{777,k}$)
Datenmatrix	N = 220 Orte, p = 877 lexikalische Arbeitskarten
Intervallalgorithmus	*MMinMwMaxX* mit 6 Intervallen (6 Farbstufen)
Kreissymbole	Kunstpunkte

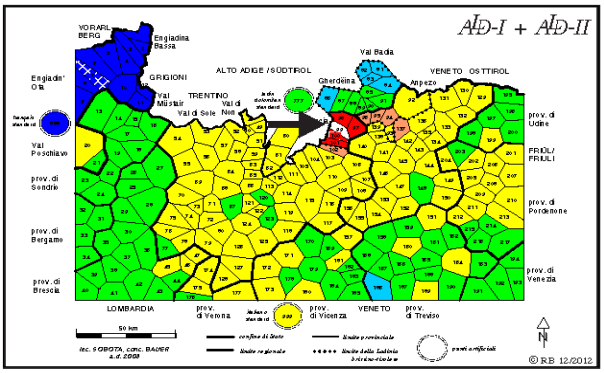
Karte 14: Ähnlichkeitskarte zum Standardladinischen (lexikalisches Subcorpus, Gesamtnetz)



grödnerischer Prüfbezugspunkt:

ÄD-P. 87 St. Christina

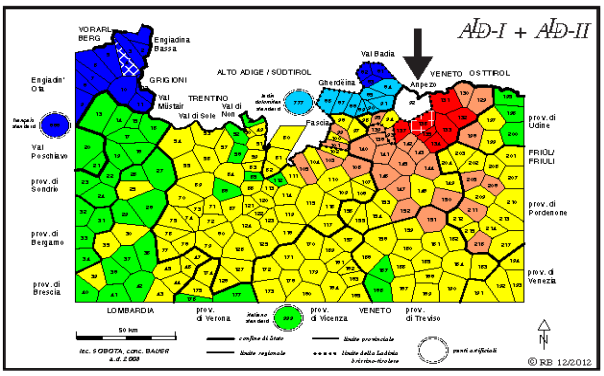
(siehe weißes Polygon und schwarzer Pfeil)



fassanischer Prüfbezugspunkt:

ÄD-P. 99 Moncion

(siehe weißes Polygon und schwarzer Pfeil)



ampezzanischer Prüfbezugspunkt:

ÄD-P. 92 Cortina d'Ampezzo

(siehe weißes Polygon und schwarzer Pfeil)

Karte 15: Dolomitenladinische Ähnlichkeitskarten (lexikalisches Subcorpus, 877 Arbeitskarten)

Zur Erhebung und kartographischen Darstellung von Daten zur deutschen Alltagssprache online: Möglichkeiten und Grenzen

Robert Möller, Université de Liège & Stephan Elspaß, Universität Salzburg

Der „Atlas zur deutschen Alltagssprache“ (AdA) ist ein seit ca. 10 Jahren bestehendes Projekt der digitalen Sprachgeographie, das sich der arealen Variation in der Alltagssprache in den deutschsprachigen Ländern widmet. Zum einen nutzt es traditionelle Verfahren der indirekten Erhebung von Sprachdaten sowie ihrer Darstellung auf Sprachkarten, zum anderen können diese Verfahren aufgrund der Möglichkeiten, die die digitale Datenverarbeitung bietet, erheblich erweitert und weitergeführt werden – und dies auf eine effiziente und kostengünstige Weise. So konnten per Online-Fragebögen bisher bis zu 10.000 und mehr Informanten erreicht werden, per Online-Karten können die Ergebnisse der Erhebungen einer breiten sprachinteressierten Öffentlichkeit bekannt gemacht werden, und da die Daten in digitaler Form vorliegen, ist eine weitere Verarbeitung (etwa für neuere dialektometrische Verfahren) jederzeit möglich. Im vorliegenden Beitrag werden zunächst die Methoden und einige Möglichkeiten dieses Projekts (etwa im Bereich der *perceptual dialectology*) vorgestellt. Darüber hinaus werden an einigen Beispielen auch mögliche Grenzen des Erhebungskonzepts diskutiert.

1 Vorbemerkungen zum Atlas zur deutschen Alltagssprache (AdA)

Seit jeher ist das ‚Vor-Augen-Führen‘ von arealen Sprachverhältnissen in Karten Darstellungsmittel und die linguistische Untersuchung solcher Karten Gegenstand sprachgeographischer Forschung. Seit Beginn der ersten großen Sprachatlasprojekte zum Deutschen haben sich freilich die Varietätenverhältnisse in den meisten deutschsprachigen Gebieten vollkommen gewandelt – und damit auch die zu erhebenden und darzustellenden Gegenstände. Auch wenn sich die Methodik im Grundsatz relativ wenig geändert hat, so eröffnen die Weiterentwicklung und Verbreitung der ‚neuen Medien‘ doch für die Sprachgeographie völlig neue Möglichkeiten der indirekten Erhebung (per Online-Fragebogen), der technischen bzw. rechnerischen Verarbeitung (z. B. mit dialektometrischen Verfahren) sowie der Präsentation (Online-Karten) areal verbreiteter Sprachdaten. In diesem Beitrag werden Verfahren und Teilergebnisse des Projekts *Atlas zur deutschen Alltagssprache* (AdA) vorgestellt, das seit etwa zehn Jahren das Internet zum einen als Befragungsinstrument nutzt und es zum anderen als Raum für die kostenlose und frei verfügbare Präsentation von über 350 Sprachkarten zur Verbreitung arealer Varianten in nähesprachlichen Registern des Deutschen verwendet. Nach einer einführenden Erörterung des zugrundegelegten Begriffs von regional gefärbter „Alltagssprache“ wollen wir einen Überblick über das grundsätzliche Verfahren zur Erhebung des wahrgenommenen Sprachgebrauchs wie auch über ein einmalig erprobtes Verfahren zur Elizitierung wahrgenommener sprachlicher Ähnlichkeiten zwischen verschiedenen Regionen geben. Im Weiteren diskutieren wir eventuelle Grenzen des Erhebungsverfahrens wie auch der kartographischen Darstellung. Zum Abschluss werden wir kurz auf weitere Auswertungs- und Nutzungsmöglichkeiten des AdA eingehen.

2 Regional gefärbte „Alltagssprache“

Der *Atlas zur deutschen Alltagssprache* (AdA, http://www.atlas-alltagssprache.de) will die Sprachformen erfassen, die Sprecherinnen und Sprecher des Deutschen in der Alltagskommunikation verwenden, also „im sozialen und funktionalen (‚Nähe‘-)Bereich des Privaten, des spontanen Gesprächs unter Freunden, Verwandten oder Bekannten oder auch im informellen Austausch unter nicht näher Bekannten aus demselben Ort, etwa im örtlichen Lebensmittelgeschäft“. In einem Großteil des deutschen Sprachraums ist dies nicht mehr der Basisdialekt. So ist die Fragestellung des AdA nicht auf eine dialektale Norm (oder auf eine standardnahe Varietät) gerichtet, sondern der Ansatz ist pragmatisch: „Bitte geben Sie bei den folgenden Fragen jeweils an, welches Wort man in Ihrer Stadt normalerweise hören würde – egal, ob es mehr Mundart oder Hochdeutsch ist.“

Dies ist dieselbe Fragestellung, die schon Jürgen Eichhoff seinem *Wortatlas der deutschen Umgangssprachen* (WDU) zugrundegelegt hat (vgl. den Fragebogen in WDU II 1978, Anhang), an den sich der AdA in seiner Grundkonzeption anschließt. Mit „Umgangssprache“ meint Eichhoff dasselbe wie wir mit „Alltagssprache“. Wir bevorzugen es, von „Alltagssprache“ zu sprechen, da der Terminus „Umgangssprache“ in der deutschsprachigen Soziolinguistik und Variationslinguistik zumeist eher auf den Zwischenbereich zwischen Dialekt und Standard – unter Ausschluss dieser beiden Pole – bezogen ist, wobei oft unklar ist, ob dabei an eine spezifische Varietät gedacht ist oder nicht (vgl. etwa Löffler 2005). Dies ist sicherlich in weiten Teilen des deutschen Sprachgebiets der Bereich, in dem die Formen der Alltagskommunikation angesiedelt sind, mit geringeren oder größeren Anteilen dialektaler Formen, aber es gibt durchaus auch Gebiete, in denen in den eingangs evozierten Situationen der Ortsdialekt verwendet wird.

Angesichts der typischen Varianz im alltagssprachlichen Gebrauch (vgl. z. B. Möller 2013, Kap. 7) stellt sich natürlich die Frage, was man bekommt, wenn man nach dem "ortsüblichen Gebrauch" fragt. Die Antwort ist jedoch eigentlich einfach: Falls keine einheitliche Vorstellung vom "ortsüblichen Gebrauch" besteht, bekommt man ein Abbild der üblichen Varianz.

Dies setzt natürlich eine große Datenmenge voraus, aber ein großer Vorteil des für den AdA verwendeten Erhebungsverfahrens liegt genau darin, dass man eine solche bekommt.

3 Zur Erhebung

3.1 Online-Fragebogen zum wahrgenommenen Sprachgebrauch

Die Erhebung für den *Atlas zur deutschen Alltagssprache* läuft über einen Online-Fragebogen (vgl. z. B. schon Varilex, Ueda 1995ff.). Der AdA-Fragebogen, der seit 2003 in jährlich neuen Erhebungsrunden ins Netz gestellt wird, ist dabei kombiniert mit der Präsentation der Ergebniskarten der früheren Erhebungsrunden. In populär gehaltener Kartierung und Kommentierung können die Informanten sich also das Ergebnis ihres Mitwirkens ansehen; wer seine Mailadresse angegeben hat, wird angeschrieben, wenn die Karten ins Netz gestellt sind – und dabei um Mitwirkung beim neuen Fragebogen gebeten. Gleichzeitig werden die Angeschriebenen gebeten, den Fragebogen-Link an potentiell Interessierte weiterzuleiten. Über diese Mischung aus Anreiz- und Schneeballsystem ist über die Jahre ein Stamm von mehreren tausend Informanten entstanden.

Im Zentrum der AdA-Erhebungen steht in erster Linie lexikalische Variation, die im deutschen Sprachraum bis in die Standardsprache noch eine große Rolle spielt (vgl. auch VWB 2004), nicht nur in Fällen wie *Samstag* vs. *Sonnabend*, sondern oft auch bei Wörtern bzw. Begriffen, bei denen kaum oder gar nicht bekannt ist, dass regionale Unterschiede bestehen (s. u. 3.). Oft kann bei lexikalischen Phänomenen im Fragebogen mit Fotos gearbeitet werden, was vermutlich günstig für die Motivation der Informanten ist. Das Interesse an solchen Fragen ist aber auch verbreitet, wie man z. B. in Internet-Foren immer wieder feststellen kann (vgl. etwa nur http://www.gutefrage.net/frage/samstag-sonnabend-wieso-2-bezeichnungen, 28.3.2013). Erhoben werden aber auch Unterschiede in der Lautung, soweit dies mit literarischer Umschrift möglich ist (Tonbeispiele könnten aber ohne weiteres einbezogen werden), und auch Fragen zur morphologischen und syntaktischen Variation werden regelmäßig in den Fragebogen einbezogen, gelegentlich auch eher volkskundliche Fragen.

3.2 Erhebungsrunde 6: Ein Online-Fragebogen zur wahrgenommenen sprachlichen Ähnlichkeit

Die 6. Erhebungsrunde 2008/9 hatte eine andere Ausrichtung: Hier ging es um subjektive Raumbildung, d. h. die Frage nach wahrgenommener und/oder angenommener Ähnlichkeit zwischen der Alltagssprache am eigenen Ort und der an anderen Orten. Während solche Wahrnehmungsräume meistens erhoben werden, indem Informanten sie auf Karten markieren (vgl. Anders 2008, 2010; Lameli et al. 2008), wurde in der 6. AdA-Runde eine tendenziell nach geographischer Lage angeordnete Ortsliste präsentiert und folgende Arbeitsanweisung gegeben:

> Bitte klicken Sie in der folgenden Liste die Orte an, in denen die Leute im Alltag ungefähr so ähnlich sprechen wie in Ihrem Ort. Natürlich kann man fast immer kleine Unterschiede feststellen, aber trotzdem empfinden Sie wahrscheinlich nicht nur die Sprache Ihres eigenen Orts, sondern die einer ganzen Gegend als vertraut. Welche Orte gehören dazu?

Die aufgelisteten Orte waren die des WDU-/AdA-Ortsnetzes; die Raster-Gliederung richtete sich nach dem Raster des WDU, die Position innerhalb der Rasterquadrate tendenziell nach geographischer Breite. Dies hatte nicht nur technische Gründe: So zeigte die Untersuchung von Lameli et al. (2008) nicht nur die Probleme von Kartenzeichnungsaufgaben auf, sondern machte auch die besondere Rolle von Städtenamen für die geographische Orientierung von Gewährsleuten deutlich.

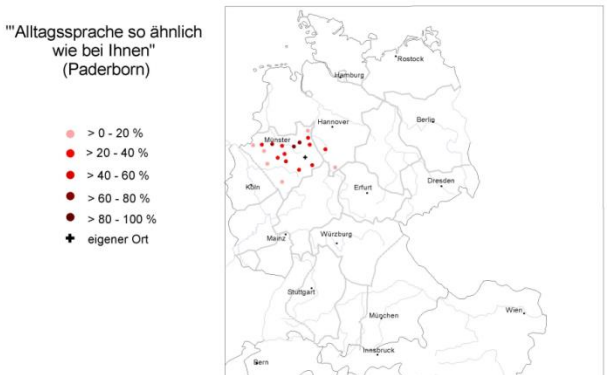


Abbildung 1: Wahrgenommene Ähnlichkeit der Alltagssprache umliegender Orte zu der von Paderborn (als eigenem Ort) (Erhebungsrunde 6)

Das Ergebnis lässt sich so darstellen wie in der Karte "Paderborn" (Abb. 1): Die Färbung der Orte richtet sich nach dem Anteil der Informanten, die den jeweiligen Ort als sprachlich ähnlich zu ihrem Heimatort angeklickt haben – im Fall von Paderborn sieht man deutlich, dass hier ein Konzept "Westfalen" erkennbar wird statt einer konzentrischen Formation um Paderborn herum.

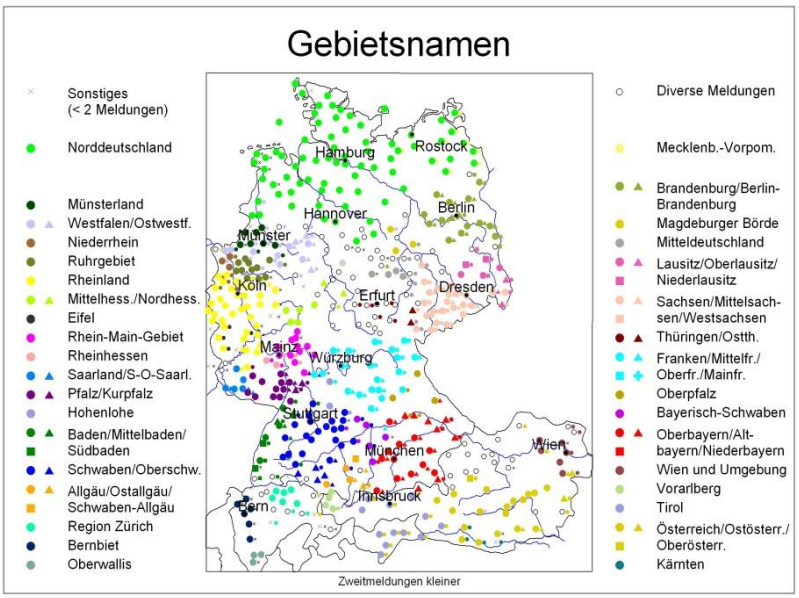


Abbildung 2: Zusammengefasste Antworten auf die Frage: "Wie würden Sie das Gebiet nennen, in dem die Leute ungefähr so sprechen wie bei Ihnen?" (Erhebungsrunde 6)

Es sollte auch angegeben werden, wie die Bezeichnung dieses sprachlichen "Heimatgebiets" ist, sofern eine existiert; die Bezeichnungen, die am häufigsten genannt wurden, zeigt die Karte "Gebietsnamen" (Abb. 2). Dies ist unseres Wissens die erste Karte, die eine Einteilung des deutschsprachigen Raums in größere Gebiete nach rein laienlinguistischer Benennung wiedergibt.

Neben der Ähnlichkeit zur näheren Umgebung wurde auch die wahrgenommene Ähnlichkeit zwischen der Sprache am eigenen Ort und der Sprache in einer Reihe von größeren Zentren erfragt. Anzuklicken war hierbei ein Wert auf einer Skala von 1 bis 6. In der Karte für Berlin (Abb. 3) sieht man z. B., dass die Nord-Süd-Teilung, die sich in vielen AdA-Karten findet, auch im Bewusstsein der Sprecher eine Rolle spielt: Orte in Franken geben hier eine geringere Ähnlichkeit an als deutlich weiter entfernte Orte in Nordrhein-Westfalen oder Niedersachsen.

3.3 Gewährspersonen

Diese 6. Runde hat interessante Ergebnisse erbracht, war allerdings für die Informanten offenbar weniger attraktiv als die anderen: Sie zeigt einen deutlichen Rückgang der Teilnehmerzahl, während in den anderen Runden zwischen 2003 und heute ein fast kontinuierlicher Anstieg der Teilnehmerzahlen zu verzeichnen ist, von anfangs knapp 1.800 brauchbaren Antworten auf über 10.000 in der 10. Runde. Es hat sich, wie erwähnt, im Lauf der Jahre ein Informantenstamm herausgebildet, der bei den neuen Runden immer wieder angeschrieben wird und dafür sorgt, dass schon innerhalb der ersten Tage mehre-

re Tausend Antworten eingehen (8. Runde: über 4.000 an den ersten 4 Tagen). Das Alter der Informanten ist dementsprechend im Lauf der Zeit leicht gestiegen, entspricht aber immer noch dem "jüngeren und mittleren" Alter, das auch Eichhoff anvisierte: Nach wie vor sind über 70 % der Informanten unter 40 Jahren alt

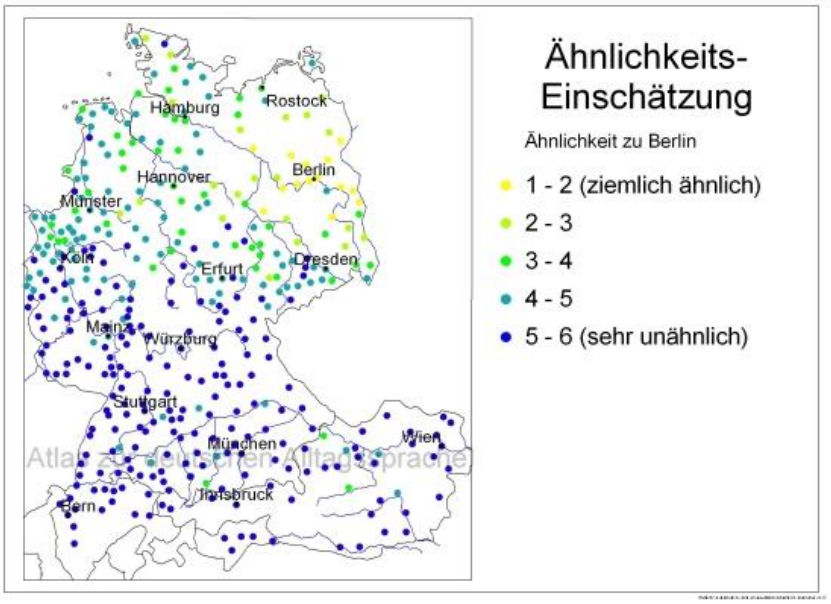


Abbildung 3: Einschätzung der Ähnlichkeit der Alltagssprache in Berlin zur Frage: "Wie würden Sie das Gebiet nennen, in dem die Leute ungefähr so sprechen wie bei Ihnen?" (Erhebungsrunde 6)

4 Grenzen dieses Erhebungskonzepts?

Die Möglichkeiten, die eine Online-Erhebung bietet, werden in diesen Zahlen deutlich sichtbar. Es stellt sich natürlich auch die Frage nach den Grenzen dieses Konzepts: Wenn die Auswahl der Informanten keinerlei Kontrolle unterliegt (es werden zwar Sozialdaten erfragt, aber die Richtigkeit dieser Angaben kann nicht überprüft werden), dann stellt

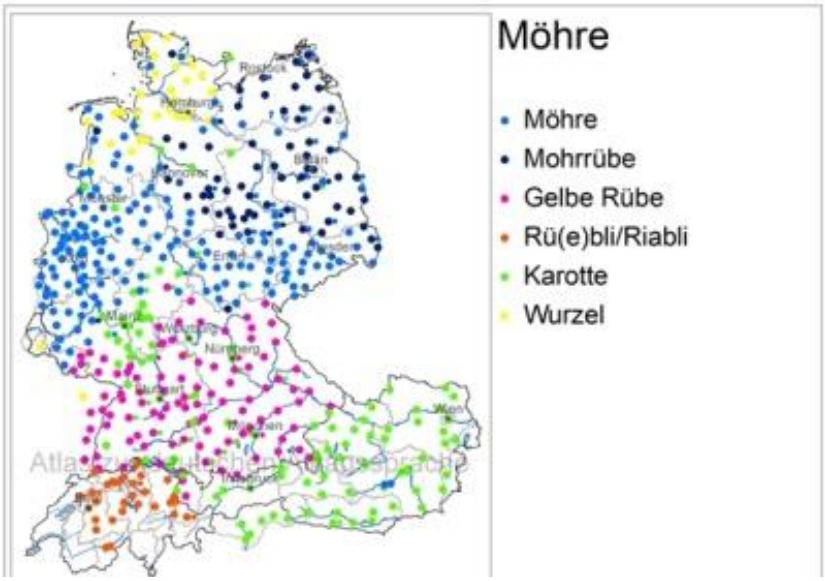


Abbildung 4: AdA-Karte *Möhre/Karotte/Gelbe Rübe/*... (Erhebungsrunde 9)

sich erstens die Frage, wie brauchbar die erhobenen Daten sind.

Das stärkste Argument für die Tauglichkeit der Erhebungsmethode sind hier die Kar-

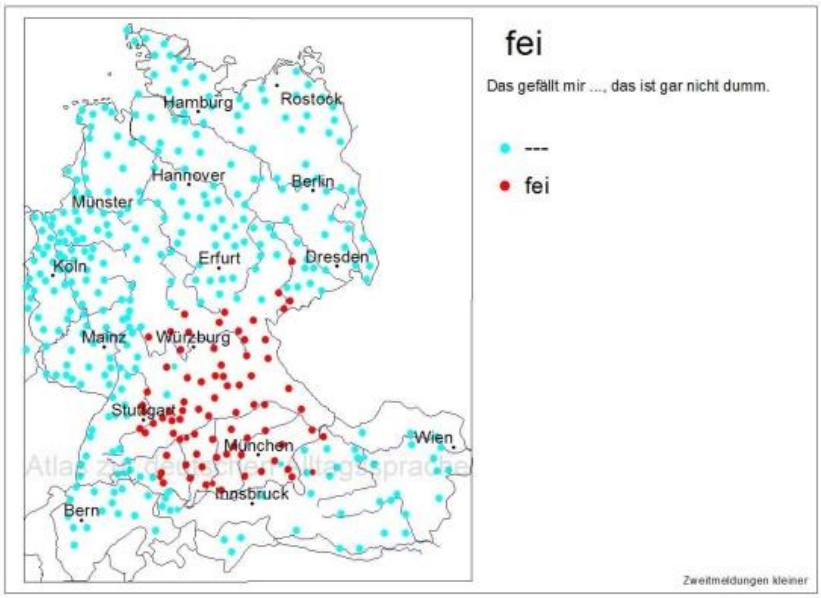


Abbildung 5: AdA-Karte Verbreitung der Partikel *fei* (Erhebungsrunde 3)

tenbilder, die oft erstaunlich klare Arealbildungen zeigen. Diese könnten nicht zustande kommen, wenn ein erheblicher Teil der Informanten keine verlässlichen Antworten geben würde. Solche klaren Arealbildungen zeigen etwa Karten wie die in den Abbildungen 4 (zu lexikalischen Varianten) sowie 5 und 6 (zur Verbreitung einer Partikel bzw. grammatischer Varianten).

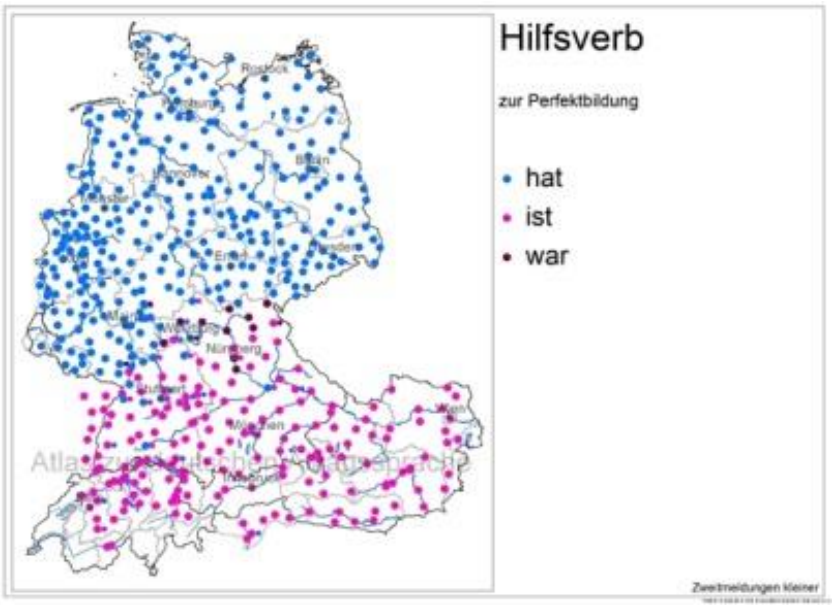


Abbildung 6: AdA-Karte *ist/hat gestanden* (Erhebungsrunde 9)

Eine Grenze des Erhebungskonzepts könnte man zweitens darin sehen, dass die Frage "Wie sagt man in Ihrem Ort?" zur Reproduktion von Stereotypen führt (vgl. auch Möller 2012, 98-102). Bei Kartenthemen, bei denen die regionalen Varianten allgemein bekannt sind (etwa *Brötchen – Semmel – Weck(le)*, vgl. Abb. 7), ist mit diesem Risiko durchaus zu rechnen.

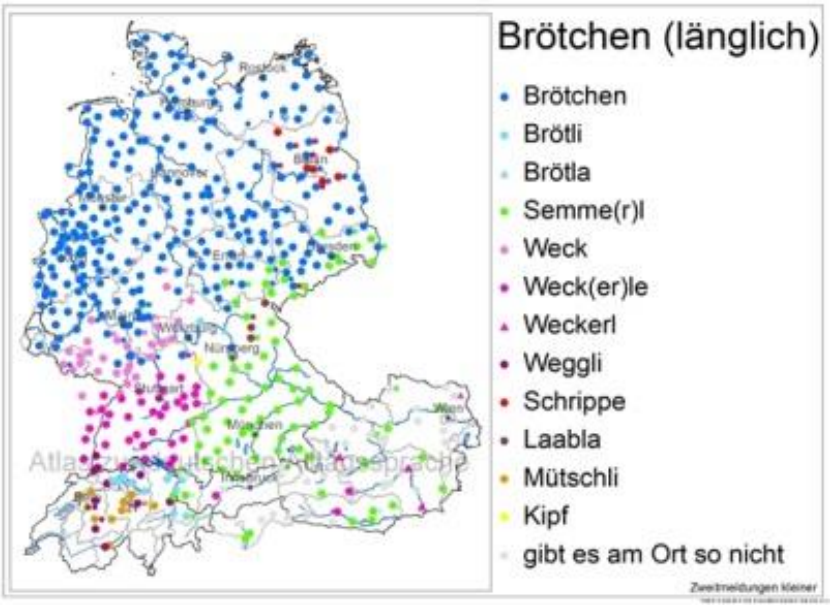


Abbildung 7: AdA-Karte *Brötchen/Semmel/Weckle* … (Erhebungsrunde 9)

Dagegen kann es sich in solchen Fällen nicht um Stereotypen-Reproduktion handeln, in denen die Existenz von regionalen Varianten und schon gar deren Verteilung vorher weitgehend unbekannt war. Wenn sich auch in solchen Fällen ein klares Kartenbild ergibt (vgl. Abb. 8), dann muss davon ausgegangen werden, dass hier auf den Fragebögen wirklich jeweils die ortsüblichen Varianten angeklickt bzw. genannt wurden.

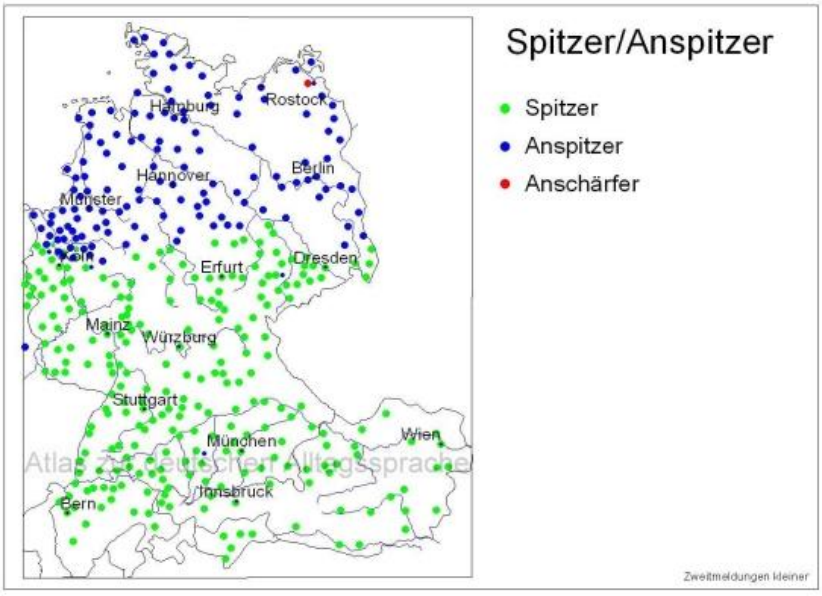


Abbildung 8: AdA-Karte *Spitzer/Anspitzer* (Erhebungsrunde 5)

In vielen Fällen existiert auch Variation an einzelnen Orten. In Abb. 9, die den Ausschnitt der Karte zu den Varianten des Demonstrativpronomens *das* für Nordrhein-Westfalen zeigt, sind in der Darstellung links die Anteile der Antwortvarianten pro Ort wiedergegeben. Aus Raumgründen muss in der Online-Präsentation des AdA auf eine solch differenzierte Darstellung verzichtet werden. Der Ausschnitt rechts gibt – zum Vergleich – die für die Kartenpräsentation des Online-AdA gewählte Form wieder. Durch die Dichte der Ortspunkte und die Kartierung von gegebenenfalls zwei Meldungen pro Ort – der größere Punkt steht für die mehrheitlich von den Informanten am Ort genannte Variante, der kleinere Punkt für die Minderheitsvariante, sofern sie über einem Schwellenwert von 33% liegt – wird die Variation aber auch so recht gut sichtbar. Grundsätzlich bleibt es jedoch möglich, bei Bedarf in jedes Areal ‚hineinzuzoomen' und die genaueren Anteile der Varianten in Tortendiagrammen oder anderen Symbolen zu veranschaulichen.

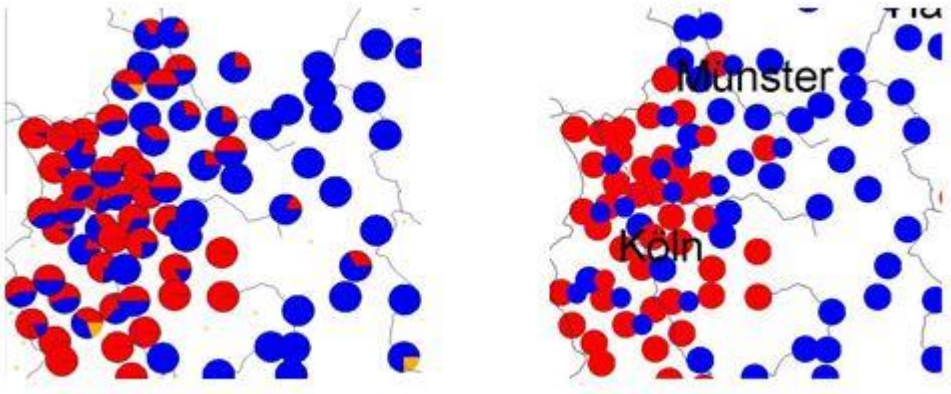


Abbildung 9: Ausschnitt AdA-Karte *das* (blau) / *dat* (rot) (Erhebungsrunde 1) – verschiedene Darstellungen der Variation an einzelnen Belegorten

Darüber hinaus ist auch eine Differenzierung der Kartendarstellungen nach Alter oder Geschlecht der Gewährspersonen inzwischen – angesichts der vorliegenden Datenmenge – ohne weiteres möglich. Erste Untersuchungen haben gezeigt, dass es kaum Unterschiede gibt zwischen den Kartenbildern, die ausschließlich auf den Antworten der männlichen Gewährspersonen beruhen, und denen, die ausschließlich die Antworten von weiblichen Gewährspersonen wiedergeben – wohingegen das Alter der Informanten durchaus einen Einfluss auf die Kartenbilder haben kann (Lang 2008).

5 Ausblick: Weitere Ergebnisse und Möglichkeiten in Auswahl

Abschließend wollen wir kurz anhand ausgewählter Karten auf zwei Möglichkeiten der Nutzung und Auswertung des AdA eingehen.

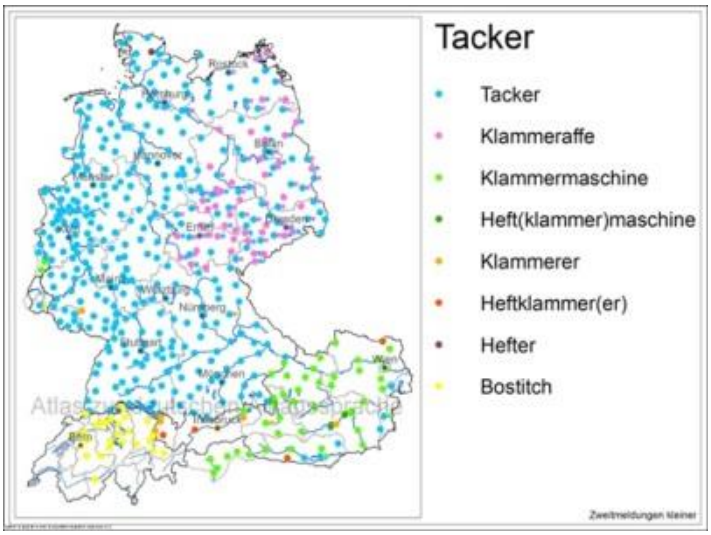


Abbildung 10: AdA-Karte *Tacker/Klammermaschine/Bostitch/...* (Erhebungsrunde 5)

Was die Sprachgeographie bisher vor allem an dialektalen Daten untersucht und getestet hat, lässt sich mit den AdA-Daten und -Karten nun auch an Daten gegenwärtiger regional gefärbter Alltagssprache durchführen. Das betrifft vor allem Analysen zum Zusammenhang zwischen außersprachlichen Faktoren und der räumlichen Verbreitung von Varianten. Die Arealbildung in den bisher erstellten und veröffentlichten Karten des AdA zeigt vielfach bekannte Muster. So stimmen auf vielen Karten die Staatsgrenzen oder andere aktuelle oder historische Grenzen mit sprachlichen Grenzen überein; z. B. wird das Gerät zum Zusammenheften von kleineren Stapeln Papier in Österreich *Klammermaschine* genannt, in der Schweiz (nach einem Hersteller) *Bostitch*, in Deutschland vorwiegend *Tacker*, (nur) im Gebiet der neuen Bundesländer daneben aber auch *Klammeraffe* (Abb. 10)

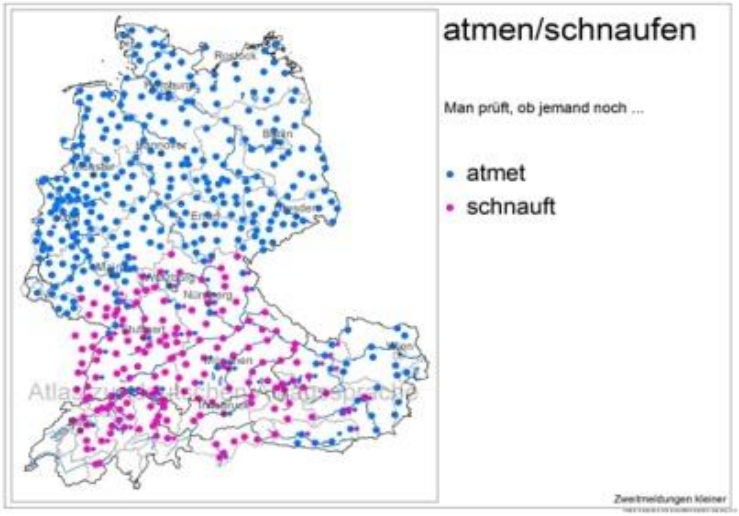


Abbildung 11: AdA-Karte *atmen/schnaufen* (Erhebungsrunde 9)

.

Ein weiteres typisches Muster ist die schon erwähnte Nord-Süd-Teilung entlang der Mainlinie, die schon in den WDU-Karten dominiert (s. Durrell 1989; Möller 2003), und ebenso in neuen Karten, wie die zur Distribution von *atmen* und *schnaufen* (i. S. von '(normales) Einziehen und Ausströmen-Lassen der Luft', s. Abb. 11).

In einer Reihe von Karten wurden schließlich auch bewusst Kartenthemen des WDU wieder aufgenommen, um einen Vergleich über die 20 bis 40 Jahre Distanz zwischen den beiden Erhebungen zu ermöglichen. In manchen Fällen zeigt sich dabei große Konstanz, wie etwa in der Karte „*ist/hat gestanden*" (s. o. Abb. 6 im Vergleich mit Eichhoff 1978, Bd. 2, 125), in anderen dagegen erhebliche Veränderungen, wie in der Karte „*Möhre/Karotte/Gelbe Rübe/...*" (s. o. Abb. 4 im Vergleich mit Eichhoff 1978, Bd. 2, 89). Auf diese Weise lässt sich die Darstellung der aktuellen sprachgeographischen Variation mit der Untersuchung gegenwärtigen Wandels verknüpfen (vgl. Elspaß 2005).

Bibliographie

Anders, Christina Ada (2008): Mental Maps linguistischer Laien zum Obersächsischen. In: Christen, Helen/Ziegler, Evelyn (Hrsg.): *Sprechen, Schreiben, Hören. Zur Produktion und Perzeption von Dialekt und Standardsprache zu Beginn des 21. Jahrhunderts*. Wien, 203-229.

Anders, Christina Ada (2010): Wahrnehmungsdialektologie. Das Obersächsische im Alltagsverständnis von Laien (Linguistik – Impulse & Tendenzen, 36). Berlin, New York.

Durrell, Martin (1989): Die „Mainlinie" als sprachliche Grenze. In: Putschke, Wolfgang/Veith, Werner/Wiesinger, Peter (Hrsg.): *Dialektgeographie und Dialektologie*. [Festschrift Günter Bellmann] (Deutsche Dialektgeographie, 90). Marburg, 89-109.

Eichhoff, Jürgen (1977ff.): *Wortatlas der deutschen Umgangssprachen*. Bd. I/II: Bern: Francke [1977/78]; Bd. III: München u. a. [1993]; Bd. IV: Bern, München [2000].

Elspaß, Stephan (2005): Zum Wandel im Gebrauch regionalsprachlicher Lexik. Ergebnisse einer Neuerhebung. *Zeitschrift für Dialektologie und Linguistik* 72, 1-51.

Elspaß, Stephan/Möller, Robert (2003ff.): *Atlas zur deutschen Alltagssprache (AdA)*, http://www.atlas-alltagssprache.de.

Lameli, Alfred/Purschke, Christoph/Kehrein, Roland (2008): Stimulus und Kognition. Zur Aktivierung mentaler Raumbilder. *Linguistik online* 35/3, 55-86, http://www.linguistik-online.de/35_08/lameliEtAl.pdf.

Lang, Benjamin (2008): *Zur dialektometrischen Analyse des „Atlas zur deutschen Alltagssprache"*. Unveröff. Zulassungsarbeit, Universität Augsburg.

Löffler, Heinrich (2005): Wieviel Variation verträgt die deutsche Standardsprache? Begriffsklärung: Standard und Gegenbegriffe. In: Eichinger, Ludwig M./Kallmeyer, Werner (Hrsg.): *Standardvariation. Wie viel Variation verträgt die deutsche Sprache?* (Jahrbuch des Instituts für Deutsche Sprache 2004). Berlin, New York, 7-27.

Möller, Robert (2003): Zur diatopischen Gliederung des alltagssprachlichen Wortgebrauchs. Eine dialektometrische Auswertung von Jürgen Eichhoff, Wortatlas der deutschen Umgangssprachen (Bd. 1-4; 1977, 1978, 1993, 2000). *Zeitschrift für Dialektologie und Linguistik* 70, 259-297.

Möller, Robert (2012): Der Sprachgebrauch „bei uns" – Arealbildung in Karten des *Atlas zur deutschen Alltagssprache*, objektive Grenzen und subjektive Räume. In: Hansen, Sandra et al. (Hrsg.): *Dialectological and Folk Dialectological Conceps of Space. Current Methods and Perspectives in Sociolinguistic Research on Dialect Change*. Berlin, Boston, 96-118.

Möller, Robert (2013): Erscheinungsformen rheinischer Alltagssprache. Untersuchungen zu Variation und Kookkurrenzregularitäten im ‚mittleren Bereich' zwischen Dialekt

und Standardsprache (Zeitschrift für Dialektologie und Linguistik – Beihefte, 153), Stuttgart.

Möller, Robert/Elspaß, Stephan (2008): Erhebung dialektgeographischer Daten per Internet: ein Atlasprojekt zur deutschen Alltagssprache. In: Elspaß, Stephan /König, Werner (Hrsg.): *Sprachgeographie digital. Die neue Generation der Sprachatlanten.* Hildesheim, Zürich, New York, (Germanistische Linguistik 190-191), 115-132.

Preston, Dennis R. (2011): Methods in (applied) folk linguistics: Getting into the minds of the folk. *AILA Review* 24, 15-39.

Ueda, Hiroto (1995ff.): *VARILEX, Variación Léxica del Español en el Mundo*. Universität Tokio, http://lecture.ecc.u-tokyo.ac.jp/~cueda/varilex/.

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.

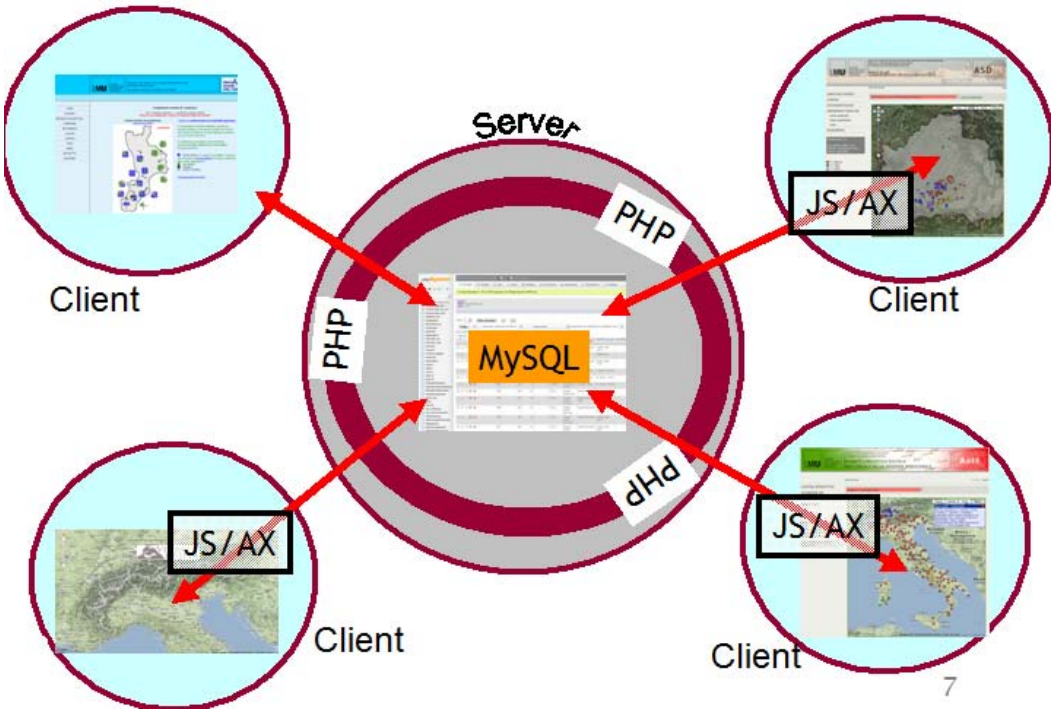


Abbildung 2: Client-Server-Prinzip

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:

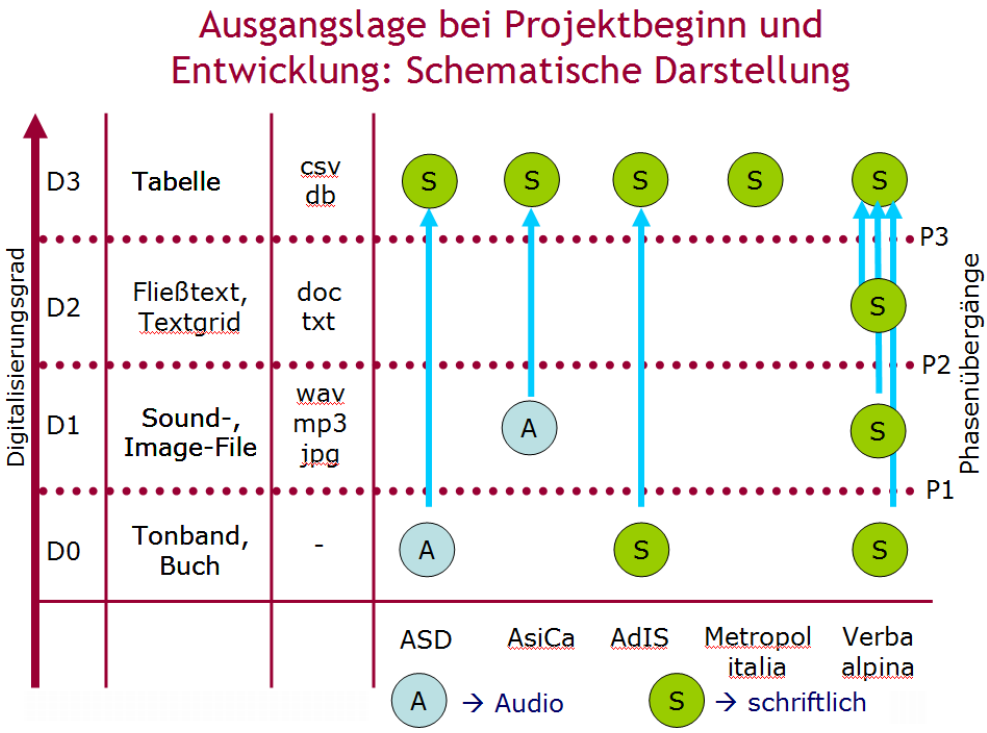


Abbildung 3: Digitalisierungsgrade und -phasen

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
Acc1wIQ1	31	1	informante	'nannuma
Acc1wIQ1	31	2	informante	jia
Acc1wIQ1	31	3	informante	mu
Acc1wIQ1	31	4	informante	piʃka
Acc1wIQ1	31	5	informante	sempre
Acc1wIQ1	31	6	informante	i
Acc1wIQ1	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
Acc1wIQ1	31	1	informante	'nannuma	Nome di parentela + poss. enclit.
Acc1wIQ1	31	2	informante	jia	verbo
Acc1wIQ1	31	3	informante	mu	Congiunzione finale
Acc1wIQ1	31	4	informante	piʃka	verbo
Acc1wIQ1	31	5	informante	sempre	Awerbio temporale
Acc1wIQ1	31	6	informante	i	Preposizione
Acc1wIQ1	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 ˈvai̯ŋtər flɛjən də gəˈdrɛçt ˈbladər ɛn dər ˈlɔft ˈəram
A	Donnersmarkt (1032a)	46.14	23.97	gəˈdrɛçt	æm ˈvæŋt͡car flæjan də gəˈdrɛçt ˈblædar æn dar 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 ˈvan͡jtçər 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 ˈvæŋtar ˈflæjan də ˈdrɛç ˈblædar æn dar 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.

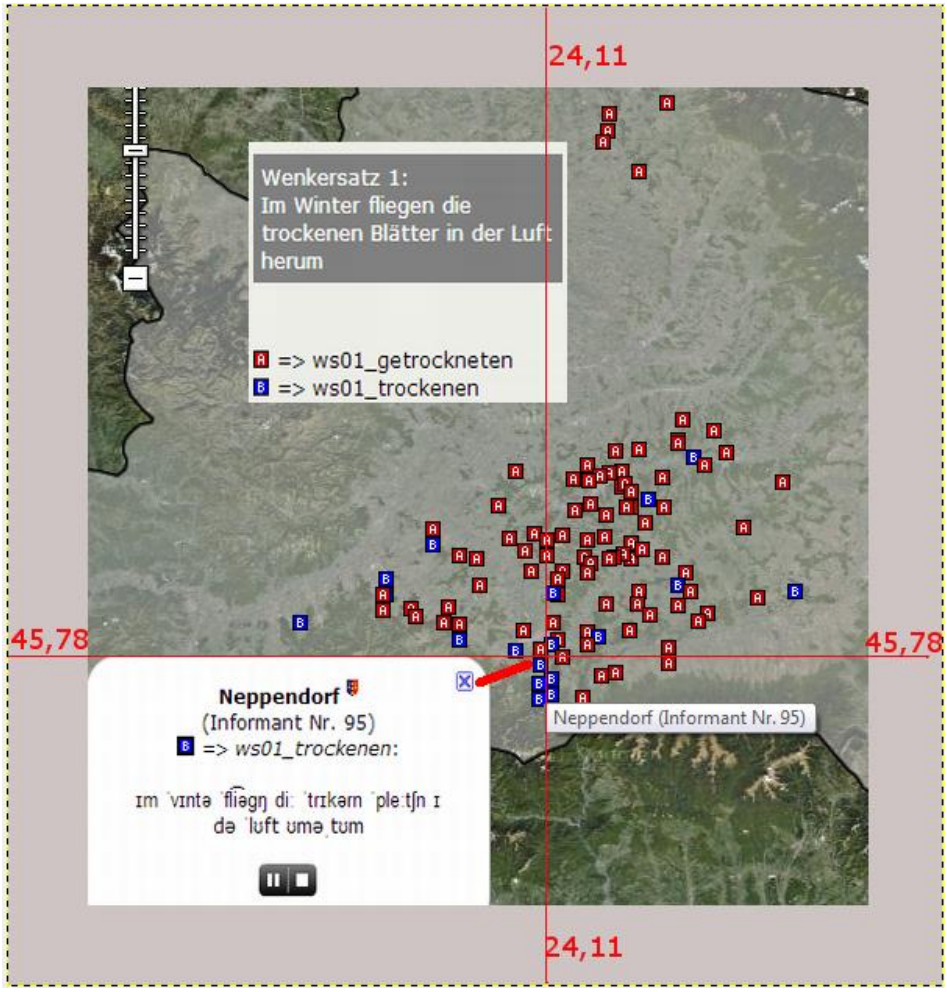


Abbildung 12: Abbildung georeferenzierter Korpusdaten auf einer Google-Karte

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.

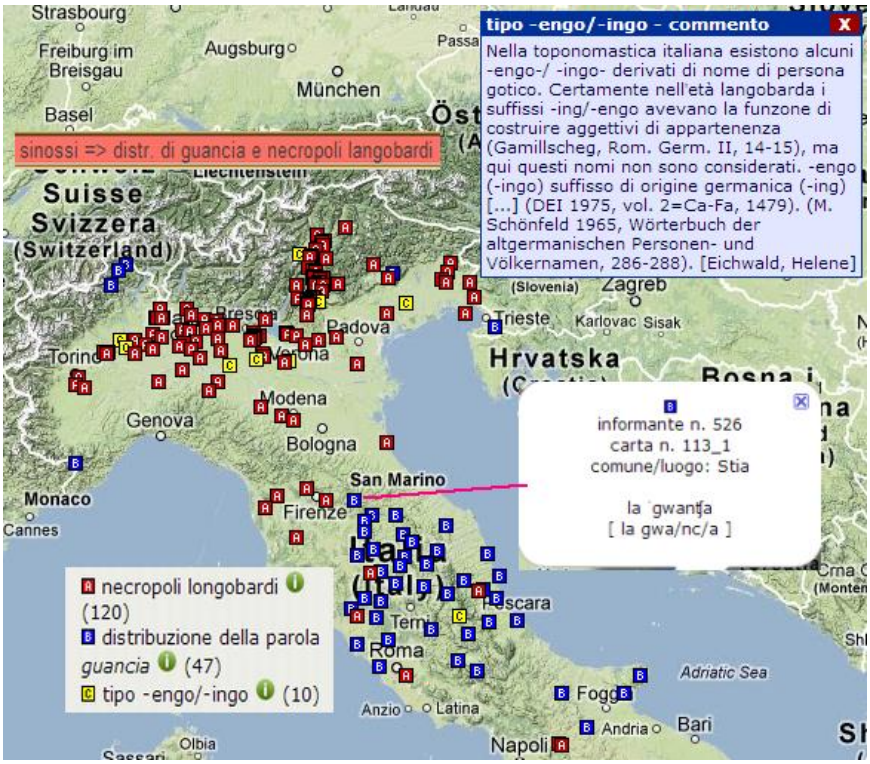


Abbildung 13: Synoptische Darstellung von Daten unterschiedlicher Art (Sprache, Archäologie)

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.

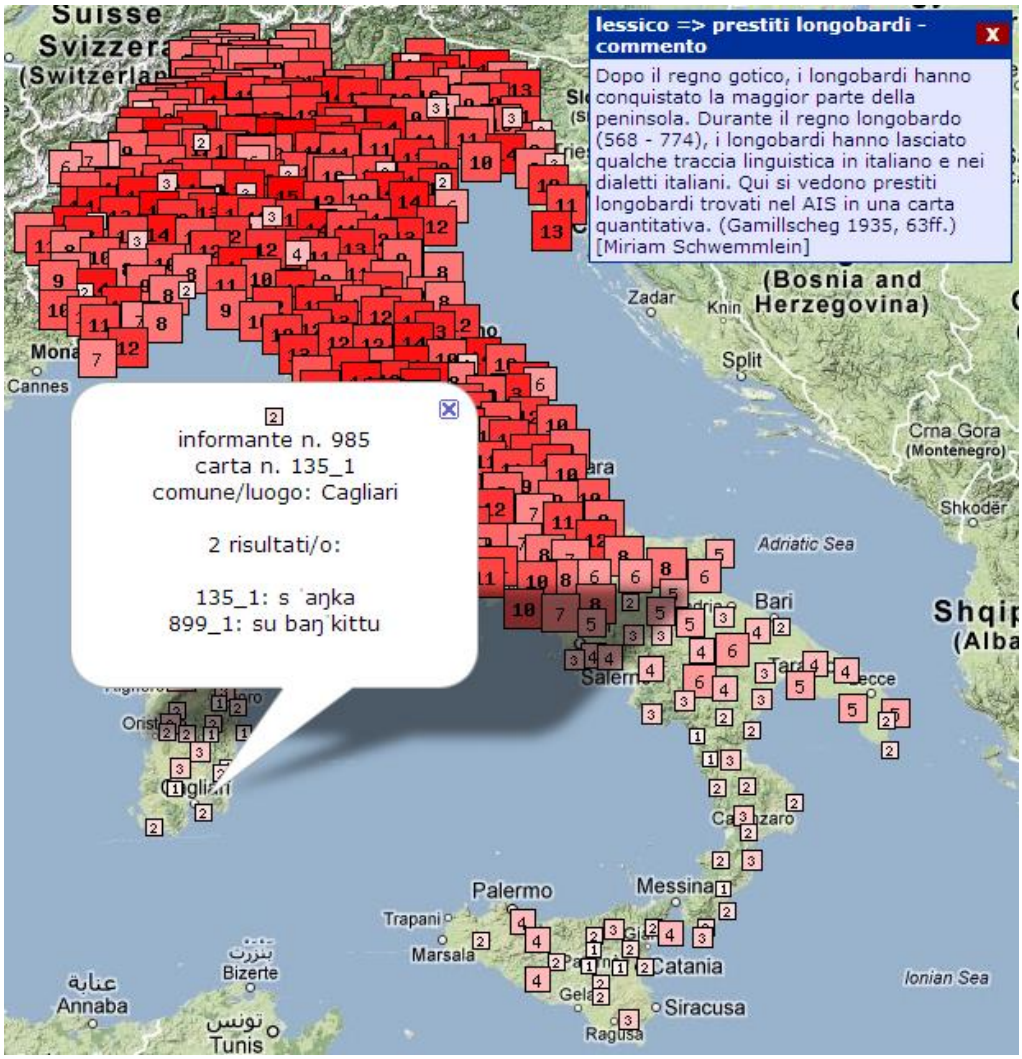


Abbildung 14: Kumulierende Darstellung von Analyseergebnissen auf 'quantitativen' Karten (Beispiel: Anzahl von Langobardismen in den Aufnahmeorten des AIS)

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.

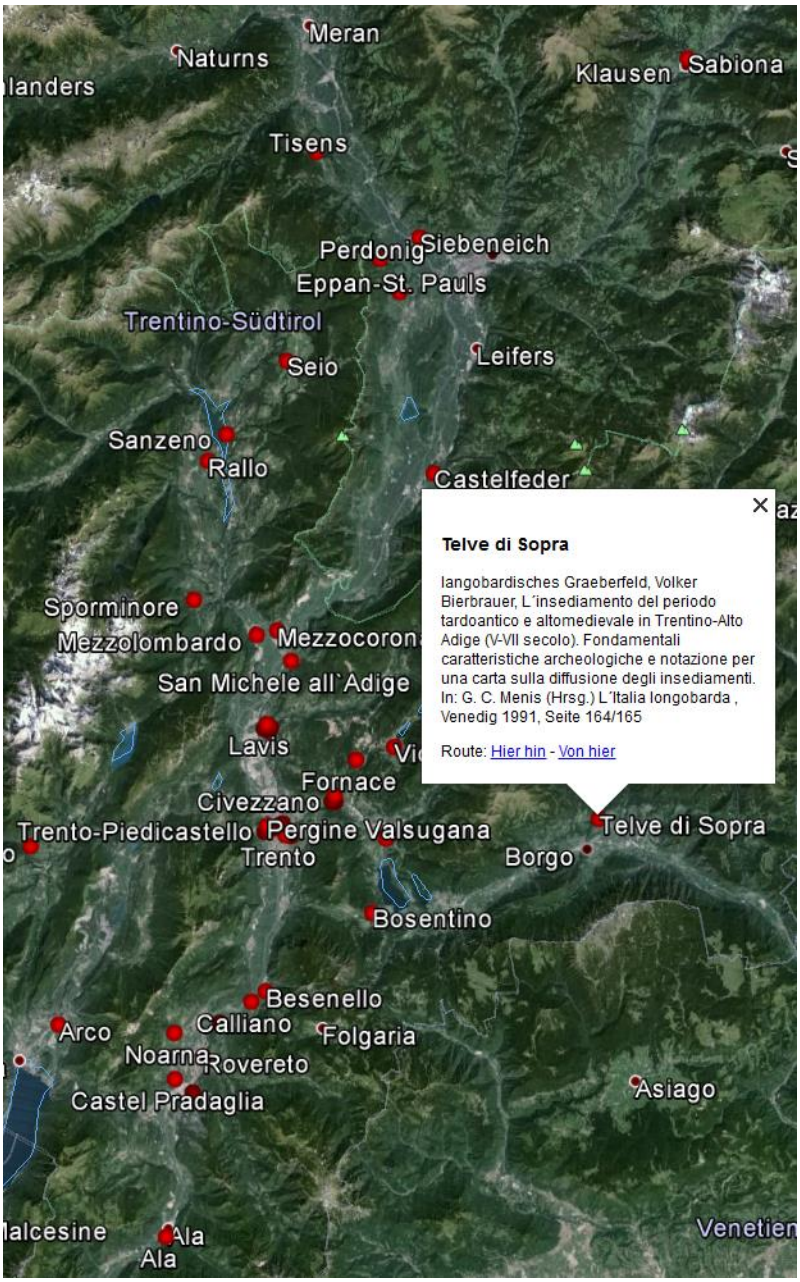


Abbildung 16: Visualisierung des Abfrageergebnisses (Abbildung 15) auf einer Google-Earth-Karte

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).

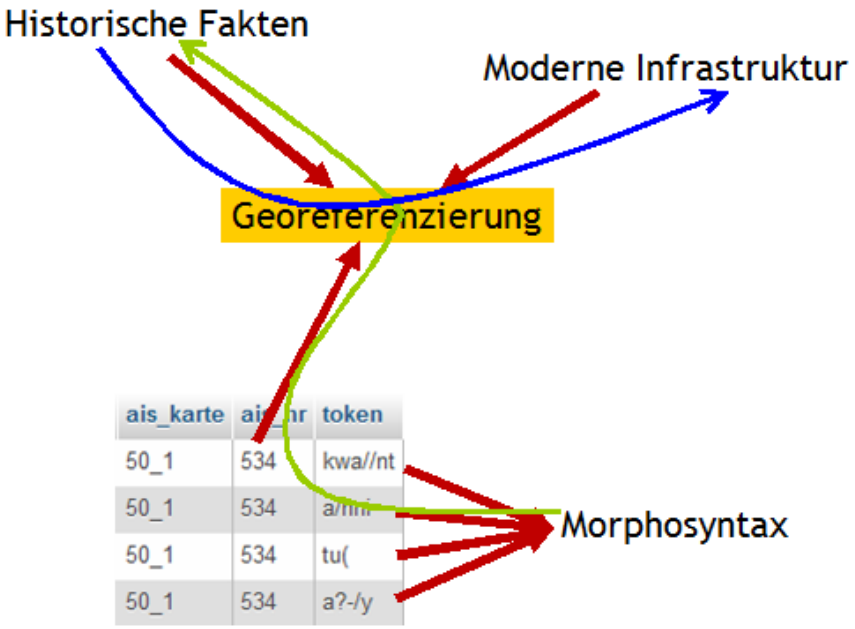


Abbildung 17: Entstehung eines Datennetzes/-gitters und die zentrale Bedeutung der Georeferenzierung

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.

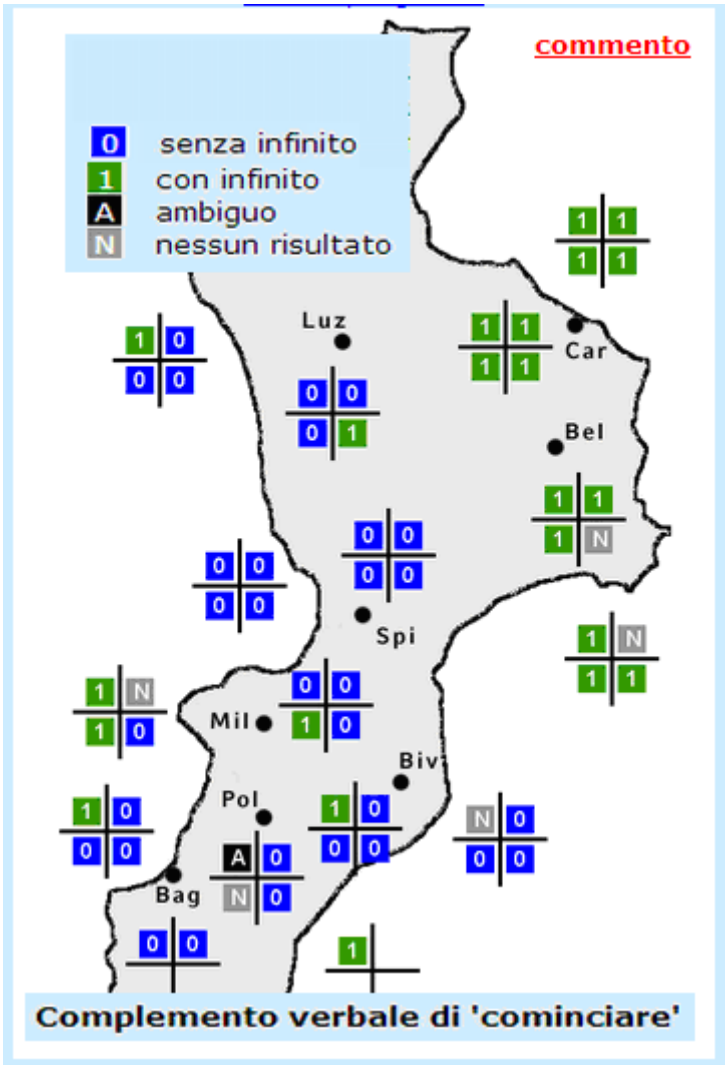


Abbildung 18: Abbildung teil-georeferenzierbarer Daten (Beispiel AsiCa)

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.

THE ROMANIAN MULTIMEDIA PROSODIC ATLAS (AMPROM)

Anca-Diana BIBIRI & Oana PANAITE & Prof. Adrian TURCULEȚ, Social-Human Interdisciplinary Research Department & Alexandru Ioan Cuza, University of Iaşi

The Multimedia Atlas of Romania (AMPRom) is a last generation atlas which combines principles of geolinguistics with techniques of instrumental phonetics and those of informatics, followed the scientific-methodological approach used by AMPER (*L'Atlas Multimédia Prosodique de l'Espace Roman*). AMPRom is conceived as an interactive database bringing together data collection and acoustic analysis concerning prosodic features of linguistic varieties specific to the Romanian language. The objective of the atlas is the documentation of geo-prosodic variation of Romanian language both at the level of dialects and idioms and at the level of colloquial literary language, which will enable the achievement of an intonative typology of the varieties of Romanian language.

1 Introduction

Romanian multimedia prosodic atlas - AMPRom - is the first prosodic atlas which aims to present the main prosodic patterns (intonation patterns) of the Romanian language varieties identified both at the level of the diatopic variants of the standard language and at the level of the dialect variants. AMPRom will be a completion of the Romanian linguistic atlases – The Romanian Linguistic Atlas (ALR) and NALR (The New Romanian Linguistic Atlas) by region (NALR and ALRR) – which have not recorded the intonation. This paper aims a broad presentation of the AMPRom project with current state of research, results and prospects for the Romanian language.

Starting point and pattern for AMPRom is *L'Atlas Multimédia Prosodique de l'Espace Roman* (AMPER), both atlases are in the stage of ongoing projects. AMPRom applies, with some own development, the design, methodology and tools developed within the Romanic project. A number of points in the AMPRom network will be found in AMPER (the part of the Romanic atlas dedicated to the Romanian language: AMPER-ROM[ÂNIA]).

AMPRom is designed in a binary structure, in terms of networks of points: the first deals with Romanian dialects and dialectal varieties (about 50 points of survey in rural areas, where the subjects have elementary education), while the second deals with the standard varieties of the diatopic language (about 20 points of survey made in cultural centers, aimed at subjects with higher education).

2 AMPRom methodology

As a computerized atlas, electronically (by the means of recording, storing, processing and audio-graphical representation of the data), AMPRom is a database, the multimedia atlas contains recorded utterances and sequences (with the informants' own speech), orthographically and phonetically transcribed, acoustically processed in text files (1, 2, 3. txt for the three repetitions of a sequence and 0. txt for the average of three repetitions), including the main physical features of each vowel from the utterance: the duration, maximum intensity (sound energy) and fundamental tone (F0) – the last one is measured in three points of vowel duration (at the beginning, in the middle and at the end). Based on these texts, graphics of intensity and duration are generated (in the form of histograms), melodic profiles of each statement, as well as average profiles (results based

on the three repetitions of a sentence). For the average profiles, tone .wav files are created through synthesis, containing ringtones of the content of segmentation; these synthetic tones can be used in tests of perception and recognition of intonation patterns, to assess and auto-evaluate.

3 Corpora

During the prosodic dialect investigations for AMPRom two questionnaires are used. The first questionnaire consists of a series of statements used for AMPER-ROM. The sets of statements that make up the questionnaire - established by morpho-syntactic and phonetic criteria - are formed by: *declarative* sentences (affirmative and negative) and total *interogative* sentences (affirmative and negative), having the syntactic structure of SVO (subject - verb - object) where S and O receive, in turns, adjective and / or prepositional determinats; the nouns and adjectives that are used in the utterances are trisyllabic oxitones, paroxitones and proparoxitones. Since in the Romanian language the negation receives usually the stress of the phrase, the negative declarative and interrogative-negative were also introduced in the questionnaire.

AMPER-ROM questionnaire (sequence):

twk Nevasta vede un căpitan/ The wife sees a captain

kwt Un căpitan vede nevasta/ A captain sees the wife

dwk Nevasta tinerea vede un căpitan/ The young wife sees a captain

gwt Un căpitan elegant vede nevasta/ An elegant captain sees the wife

swk Nevasta frumoasă vede un căpitan/ The beautiful wife sees a captain

pwt Pasărea vede nevasta/ The bird sees the wife

zwk Nevasta harnică vede un căpitan/ The hardworking wife sees a captain

bwt Pasărea papagal vede nevasta/ The parrot bird sees the wife

twg Nevasta vede un căpitan elegant/ The wife sees an elegant captain

fwt Pasărea frumoasă vede nevasta/ The beautiful bird sees the wife

Syntactic and phonetic restraints to which the 'fix' minimal corpus has to respond that is set for AMPER will be found, for the same reasons of contrastive analysis conditions, also in the questionnaire designed for AMPRom. However, to capture a larger number of Romanian intonation patterns in their territorial distribution, a second questionnaire includes other statements, simpler (without many formal constraints) to facilitate the contact with the subjects and to prepare them for the fixed questionnaire, the AMPER-ROM. That includes about 100 sentences and has two variants: low version (compulsory) and extended version (optional), the latter is done only in some points of inquiry being applied once, with the best informants.

There are presented the types of syntactic structures that make up the AMPRom questionnaire:

- VO structures (with inclusive subject): (1) 1a: L-ai văzut pe Ion?/ Have you seen John? (2) 2a: L-ai văzut pe Vasile?/ Have you seen the Basil? (3) 3a: Ai văzut fetele?/ Have you seen the girls?
- Structures pursuing the ratio of the word order and prosody: (1) 1b: Pe Ion l-ai văzut?/ John was that you have seen? (2) 2b Pe Vasile l-ai văzut?/ Basil was that you have seen? (3) 3b: Fetele le-ai văzut?/ Girls were that you have seen?
- VS/SV Structures: (25) 25a Vine Ion./ There comes John 25b Ion vine./John is coming (28) 28a Cine vine?/ Who is coming? 28b Ion vine./John is coming.
- Structures with double negation elements both in the question and the answer: (26): Nu vine nime(ni) la noi?/ There comes nobody(none)to us? (30): N-a venit nime(ni) la noi./Nobody(none)came to us.
- Structures in which modulators are used (adverbs of manner and semi-adverbs – sure, precisely, certainly, immediately, surely, maybe, whether, really or even modal verbs – I think, it might): 20b: Chiar vine Ion?/ Really, is John coming? 21a: Sigur/Precis (că) vine/Sure/precisely he is coming. (23) 23b S-ar putea să nu vină./It might be that he is not coming. 23c. Cred că vine./ I think he is coming.
- Structures containing different types of questions: partial, alternative, confirmation: (56) 56a: Cât e ceasul?/What time is it? 56b: Cât e ora? / What time is it? (41) Vii ori nu vii?/ Are you coming or not? (55) 55b: Pleci mâine la Iaşi, nu-i aşa?/ You are going tomorrow to Iaşi, aren't you?
- Structures containing vocative addressing and calling: (40): *Ion (Ioane), dă-mi un măr (te rog)!* / Ion (John), give me an apple (please)! (35): 35a: *Ana!/ Ann!,* 35b: *Maria!/ Mary!,* 35c: *Ileana!/Helen!*
- Structures that require intonation suspension (to express the continuity): (49) *– Apucă-te/Ia şi-nvaţă, că de nu.../ Start/Let's learn, or else...*
- Structures that are intended to stress the prosodic expression that emphasizes the subordination relationship: (72) 72a: *Când am ajuns la piaţă, ploua cu găleata/ When I arrived at the market, it was pouring*; 72b: *Ploua cu găleata, când am ajuns la piaţă./ It was pouring rain when we arrived at the market.*
- Exclamatory structures: (84): Ce batic frumos ai!/ That's a beautiful scarf! (85): Ce miroase a pâine caldă!/That's a good smell of hot bread!
- Structures on intercalation prosody: (74) 74a: Tata mi-a zis: Du-te repede şi cheam-o pe soră-ta! / My father said, 'Go quickly and call your sister'!74b: Du-te repede şi cheam-o pe soră-ta! mi-a zis tata. / Go quickly and call your sister! my father said.74c: Du-te repede – mi-a zis tata – şi cheam-o pe soră-ta!/ 'Go quickly' my father said.'and call your sister!'
- Structures containing enumerations: (66): Am fost la piaţă/târg şi am cumpărat: roşii, ceapă, morcov şi ardei./ I was at the market / fair and bought tomatoes, onions, carrots and peppers. (84): Luni, marţi, miercuri...(şi) duminică./ Monday, Tuesday, Wednesday ... (and) Sunday.

- Structures containing a sequence of short sentences: (79): De dimineaţă m-am trezit, am pregătit micul dejun şi apoi am plecat la serviciu./ This morning I woke up, I made breakfast and then went to work.
- Sentences with the same structure (V) for the affirmative, interrogative and imperative mood: (80): *Aşteaptă/Wait.*. (81): *Aşteaptă?/Wait?* (82): *Aşteaptă!/Wait!/ Aşteaptă-mă!/Wait for me!*
- Structures with a focus on constituents of the statement: (4): 4a Pe **Vasile** l-ai văzut ?/Was **Basil** that you saw? 4b L-ai văzut pe **Vasile**?/ Did you see **Basil**?; (58): Bei **vin**?/Are you drinking **wine**?
- Structures with a successive focus on constituents of the statement (64): ***Mănânci*** *peşte?/* ***Are*** *you* ***eating*** *fish?* 65a: *Mănânci* ***peşte****?/ Are you eating* ***fish****?*
- Affective structures: (56f): E/îi amiază? / Is it/ It's noon? It's already noon? (59): **Bei vin?/Are you drinking wine?**

The extended form of the questionnaire contains other type of syntactic structures:

- Structures pursuing the prosody of idioms and phrases: (89 a, b, c...): da de unde/what, no way; nu mai spune/ yah, do not say; ce folos;/ so, what; nici vorba/pomeneala/no way/not at all; cum/unde să facă ea aşa ceva/what/how did she do that; da mai ştii?/that could be?, ei şi?/so, what?, iote/iete/there/;, măi/uite/inga/ni la el!/Hey/look/you/you there.
- Structures containing greetings and politeness: (91): Bună ziua!/ Good afternoon!; (97): Poftim/There you go!/ Na!/Here! – Mulţumesc/mulţam!/ Thank you/Thanks – Poftim, pentru puţin, cu plăcere, să creşti mare (la copii)./ There, Don't mention it, you are welcome, May you grow strong! (for children).
- Structures that use adverbs and adverbial phrases to strengthen the assertion and negation: (104): Da,/Yes Sigur,/Sure Fireşte/Surely, Negreşit!/No doubt! (105): Nu,/No Nicidecum,/No way Niciodată,/Never Nici în ruptul capului!/On no account!
- Imprecations: (107 a, b, c...): Arde-l-ar focu să-l ardă!/ May he burn in hell! Lua-l-ar naiba/dracu să-l ia!/ The hell/the devil with him! Fir-ar/fi-o-ar a dracului!/ Damn it/Damn with it! – Du-te dracului/la dracu/la satana!/ Go to the devil/to Satan!

In some sentences with neutral intonation required by the AMPER-ROM, questionnaire, during the prosodic dialects surveys we asked for limited focus (especially by contrast) of some constituents of the statement. Through the two questionnaires, AMPRom exceeds the fixed corpus of AMPER-ROM restrictions, both in terms of morpho-syntactic structures that are investigated and addressing other aspects of intonation than the neutral one. Statements are recorded at least three times and obtained through indirect questions (for attaching the involved words) and by verbal and non-verbal implications (facial expressions, gestures) to the context and / or forming some speech situations during the continuous dialogue between the investigator and informant, and, in some ultimate circumstances the investigator is saying the statement (with a monotonous intona-

tion, mechanical that does not suggest the actual intonation), the informant statements should not read the statements to avoid the specific "reading" intonation."

The investigation usually begins with discussions between the investigator and locals (free corpus), while the investigator has the opportunity to observe the intonation patterns of local speakers and to choose the best informants to achieve the prosodic dialectal investigations. Than it follows the AMPRom questionnaire, considering that it contains statements similar to usual speech, achieving a favorable atmosphere for the investigation. Then the questionnaire AMPER-ROM is accomplished, demanding that in the end is recorded the focused sentences which the accent placed on constituents from different sentences. The two surveys are repeated three times in different sessions on the same day or / and in the following days.

4 AMPRom network of points

The Romanian language field – as it aprears in ALR (*Romanian Linguistic Atlas*) – contains the follwing dialects: D a c o- r o m a n i a n , A r o m a n i a n (M a c e d o r o m a n i a n), M e g l e n o r o m a n i a n and I s t r o r o m a n i a n with their subdialects and patois; that is the research area for both atlases, which have addionally included the diatopic varieties of standard language/culture. If in the project of AMPER 10 points of survey are included (with codes 90-99) for Romania and The Republic of Moldovia (most of them being rural, but there are also cities where we want to record for the diatopic variants of the literary language), the network of points for the AMPRom starts from the historical provinces and it covers the whole teritory of the Dacoromanian dialect: Moldovia, Greater Wallachia, Transylvania, Maramureş (part of the historic Partium region), Crişana (part of the historic Partium region), Banat (part of the historic Partium region), Lesser Wallachia, Dobruja, Bessarabia, but also the teritories of the other three dialects. For the coding areas and survey points, the Romanian language was divided into 21 areas. Daco-Romanian dialect includes the provinces of Romania, labeled as follows: A = Moldovia, B = south of Bucovinei (Romania), D = Transylvania, E = Maramureş, F = Crişana, G = Banatul, H = Greater Wallachia, J = Oltenia, K = Dobruja, to which there have been added the territories outside the Romanian borders where Romanian speakers are found in compact area: C = Republic of Moldova (Bessarabia), L = Chernivtsi and Transcarpathian, N = region of Ukraine, Odessa Ukraine = M, Vojvodina Timoc Valley (Serbia, Bulgaria) = N, Hungary = O. Aromanian dialect areas are: P = Greece, R = Albania, Macedonia = S Bulgaria = T Megleno dialect is found in: Republic of Macedonia = U, Greece = V; Istro-Romanian dialect is in Croatia = Z. Regarding Aromanian and Meglenoromanian dialects the record/ surveys are made/ will be conducted both in the Balkan Peninsula, and with the Aromanians and Meglenoromanians in Dobruja, where speakers of these dialects were settled in the third decade of the twentieth century.

The density of survey points for AMPRom will depend on their representativeness. We appreciate that it may be sufficient dozens (approximately 70) of points of prosodic dialectal investigation for the intonation patterns.

5 Informants

To achieve the speech prosody documentation, for both the local and standard language there may not be used the same informants; in the first case there will be surveyed rural subjects (selected from ALR and NALR network) and in the second case, the survey will

take place in the cities, in the most important cultural centers for that province. Among the classical criteria required in a dialectical survey that the informants must meet include: communicative availability, average age, spontaneity, minimal school education, good diction and voice, regular voice (not to lose his voice during the pronunciation of the utterance).

In rural areas two informants are used (with the code 1 (odd) for women and 2 (even) for male) indigenous, representative for the local speech, with elementary education (up to high school), middle-aged - 30-50 years (if necessary psychophysical conditions are met, they may be older) who speak natural under the conditions of the investigation. In urban areas the surveys twofold: besides informants 1 and 2 (belonging to lower social class / low and / or middle / middle, with influences of the local dialect), there are used informants 5 (female) and 6 (man) with higher education (belonging to the upper social class / high), speaking cultural language, but which are normal people (i.e. not the "professional speakers", more precisely those who work in the media, teachers, especially those teaching Romanian language and other languages). If there are more than four informants interviewed according to their socio-cultural status, they will receive tokens 3, 4 ... and 7, 8.

6 Acoustic processing of the recorded data

Acoustic analysis tools that are used in processing the prosodic dialectal material that was recorded during investigations are PRAAT / SCRIPT PRAAT for AMPER (Antonio Romano, Albert Rilliard), Matlab, AMPER 2006, Computer interface of prosodic curve. Statements are recorded in digital format (files with .wav extension - Waveform Audio File Format) and acoustic analysis using software tools. The sequence analysis goes through several stages: changing the sampling frequency sound wave of 48 kHz to 16 kHz (GoldWave) delineation and labeling according to the statements used in the questionnaire: numbering / coding of the statements is made for AMPER-ROM, keeping for AMPRom the first four indices of code. The numbering of the points includes number 9 – the code for the Romanian language, one letter - which indicates the area of the survey point (of the 21 mentioned above) and another figure - from 0-9 – from the number of (up to) 70 points of investigation. Encoding a sentence consists of 6 symbols: for the investigated point, the sign of the informant (by gender and training), encoding the statement - with a symbol consisting of numbers and letters (because of the various structure there is not permitted the coding type AMPER) - and its registration number (1, 2, 3 ...). For example, *9A011a1* is the first record of the first utterance of AMPRom questionnaire, formulated by the first female informant in the locality of Iaşi, Moldavia, in the Romanian language.

Assisted by the software - PRAAT – there follows the segmentation and labeling vowel elements (in the case of diphthongs the two vowel marks go toghether), based on oscillograms, spectrograms and by hearing, thus there occur, for each analyzed utterance, texts in which are found physical correlates of vowels: duration, intensity and fundamental frequency (F0 – for the three points of the vowel). Based on these texts, there are obtained, using Matlab routines, average values (0.txt), duration graphs, intensity and individual melodic outlines etc.

7 Achievements and prospects.

The investigations forAMPRom began with a CNCSIS (Project Director Adrian Turculet) grant was obtained and conducted during 2007-2008. During the preparation of surveys for AMPRom, for the Seminar of Phonetics and Dialectology, involving Bachelor and Master students of the Faculty of Letters, there were conducted sample surveys in some localities of Moldovia, Bucovina, Wallachia, Transylvania, Maramures and The Republic of Moldovia. A first database was established for the geoprosodic research of the literary language variation and its regional variants. Data and their processing have been exploited in various communications and papers. We are currently in the stage of the processing of acoustic data collected from surveys conducted in dialectal areas of Moldovia, Bucovina, Maramures, Crisana, a stage in which there attend a team of PhD students and researchers from the *Department of Interdisciplinary Research in Humanities* at the "Alexandru Ioan Cuza" University of Iasi, the last two areas corresponding to *AMPRom. Maramureş* and *AMPRom. Crişana*, ongoing projects financed from the European Social Fund through the Sectoral Operational Program for Human Resources Development in the Project *Innovative Development and Research Impact through Postdoctoral Programs POSDRU/89/1.5/S/49944*); some of the points of the surveyed are part of the international project AMPER-ROM.

AMPRom website was made in spring 2011 and it contains the first database available for the specialists, as well as to anyone interested in issues of prosody. The site has been loaded with processed data analyzed from 5 cities in the area A, represented by Moldovia on the right bank of Prut river: 9A0 – Iaşi, 9A1 - Dolhasca, 9A4 - Pufeşti, 9A5 - Muntenii de Sus and 9A9- Liesti.

For the database – BD in Grenoble were sent eight points of the survey for the Romanian language (AMPER-ROM), which are also included on the DVD accompanying the paper *Intonations romanes*, done in 2011 in Grenoble. In addition, there were still processed 11 cities and 12 villages of Dacoromanian dialect and in the locality Gevgelja in the (Republic of Macedonia) (9U0) for the Meglenoromanian dialect..

8 Conclusions

Acoustic-auditory research on Dacoromanians idioms done so far have shown that areas intonation do not coincide with dialectal areas that can be drawn using segmental data that are present in the Romanian linguistic atlases (RLA, NRLA). AMPRom aims the research of the dacoromanian dialects from a prosodic perspective, to demonstrate the overlap and differences between prosodic and dialectal areas. The Romanian Multimedia Prosodic Atlas (AMPRom) brings new data to the research of diatopice variation, but also to some diastratum aspects and diaphasic prosody for the Romance languages.

The Romanian language geography was incorporated into the Roman geolinguistic by traditional accomplishments found in language atlases: *Linguisticher Atlas für dacorumänischen Sprachgebietes* of Gustav Weigand, *Atlasul lingvistic român* (ALR), *Noul Atlas lingvistic român, pe regiuni* (NALR/ALRR) and subsequently through the atlases of the second generation, interpretative: *Atlasul limbilor romanice* (ALiR) and *Atlas linguarum Europae* (ALE). Through the accompliment of proposed project there can be possible to integrate the Romanian language in the first prosodic atlas (AMPER) and to accomplish the first multimedia romanian prosodic atlas (AMPRom).

9 Bibliography

Intonations romanes, 2011. Intonations romanes, coordonné par Paolo Mairano. În „Géolinguistique“, Hors-série nº 4, Ellug, Université de Grenoble.

Projet AMPER: Atlas Multimédia Prosodique de l'Éspace Roman, 2005. În „Géolinguistique“, hors série 3 (ed. J.P. Lai), Ellug.

Michel Contini. 2e Séminaire international du projet AMPER, 2005. În Projet AMPER, pp. I- XI.

Michel Contini. *Le projet AMPER: passé, present, et avenir,* 2007. În L. de Castro Moutinho, R. L. Coimbra (ed.), *I Jornadas Científicas AMPER-POR. Actas* (Aveiro, 29-30 octobre 2007), pp. 9- 19.

Laurenţia Dascălu-Jinga, 2001, Melodia vorbirii în limba română, Editura Univers Enciclopedic, Bucureşti.

Adrian Turculeţ, L. Botoşineanu, Ana-Maria Minuţ, *Atlasul multimedia prosodic roman (AMPRom). Chestionarul şi reţeaua de puncte*, 2006. În „Limba şi literatura română. Regional-Naţional-Universal“ (Simpozion internaţional Iaşi-Chişinău, 24-27 noiembrie 2005), Iaşi, Casa Editorială „Demiurg“, p. 283-293.

Adrian Turculeţ, Ana-Maria Minuţ, *De la AMPER la AMPRom,* 2007. În Luminiţa Hoarţă Cărăuşu (coord.) *Rezultate şi perspective actuale ale lingvisticii româneşti şi străine*, Editura Universităţii „Al. I. Cuza“, Iaşi, pp. 349-361.

Adrian Turculeţ, 2007, *Un nou atlas lingvistic romanic: AMPER,*. În „Studii şi cercetări lingvistice“, ianuarie-iunie, 2007, Bucureşti, pp. 203-214.

Adrian Turculeţ (ed.), 2008, *La variation diatopique de l'intonation dans le domain roumain et roman,* Editura Universităţii „Al. I. Cuza“, Iaşi.

Adrian Turculeţ, Oana Beldianu, Anca-Diana Bibiri, 2012, *Pentru un Atlas multimedia prozodic român (AMPRom)*. În Colocviul Internaţional „Filologia modernă: Realizări şi perspective în context european. Abordări interdisciplinare în cercetarea lingvistică şi literară“, Academia de Ştiinţe a Moldovei, Institutul de Filologie, Chişinău, pp. 397-402.

http://amprom.uaic.ro/

http://w3.u-grenoble3.fr/dialecto/AMPER/amper.htm

Acknowledgements: The paper was done with support from the European Social Fund through the Sectoral Operational Program for Human Resources Development in the Project Innovative Development and Research Impact through Postdoctoral Programs POSDRU/89/1.5/S/49944.

Keywords: AMPRom, acoustic analysis, prosodic features, intonative typology of the varieties of Romanian language.

L'atlas linguistique audiovisuel du francoprovençal valaisan et les défis du polymorphisme

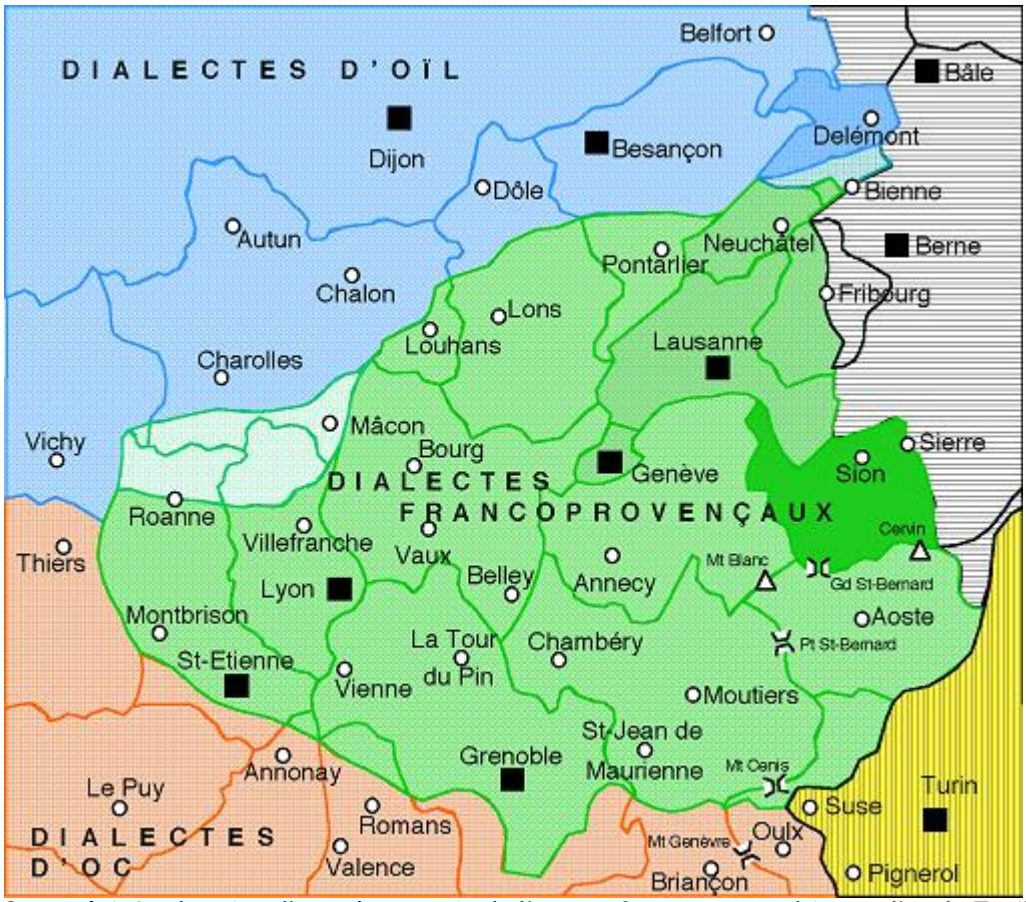
Federica Diémoz et Andres Kristol, Université de Neuchâtel (Suisse)

L'*Atlas linguistique audiovisuel du francoprovençal valaisan* (*ALAVAL*), en voie d'élaboration au Centre de dialectologie de l'Université de Neuchâtel (Suisse), est axé sur des questions de morphologie et de syntaxe, les domaines les moins étudiés de la linguistique francoprovençale. Il s'appuie sur des données dialectales recueillies par un questionnaire et des conversations dirigées qui laissaient une grande liberté aux témoins dans leur manière de formuler leurs énoncés; la totalité de nos enquêtes a été enregistrée par caméra vidéo (trois à cinq heures d'enregistrements par témoin). Le questionnaire utilisé était construit sur le principe de la redondance : l'analyse linguistique repose non pas sur des énoncés «types», mais sur de nombreuses occurrences du même phénomène. L'*ALAVAL* est ainsi en mesure de documenter l'importante variation interne qui caractérise les parlers francoprovençaux – et vraisemblablement toutes les langues vivantes faiblement normées : un même témoin a souvent produit trois ou quatre formes alternatives d'une même forme verbale, d'un même pronom clitique, et de nombreuses allomorphies pragmatiques, alors que les relations entre enquêteurs et informateurs sont restés relativement stables pendant toute la durée de l'enquête. La nature de nos données – énoncés complets, intégralement conservés et accessibles aux utilisateurs de l'atlas, sous forme de clips vidéo transcrits et traduits – nous a obligés à trouver des solutions innovatrices dans l'analyse et dans la représentation cartographique. Dans notre communication, nous discuterons les difficultés résultant du polymorphisme intrinsèque de nos données, et les solutions adoptées: comment cartographier par exemple l'expression du sujet indéterminé dans une langue où une demi-douzaine de tournures alternatives sont en concurrence, tout en garantissant la lisibilité des cartes à l'écran de l'ordinateur, et ceci même pour des utilisateurs daltoniens ? Quatre pages «modèles» de l'*ALAVAL* peuvent être consultées sur le site internet du Centre de dialectologie de l'Université de Neuchâtel (http://www2.unine.ch/dialectologie/page-8174.html); la totalité des données (97 cartes actuellement disponibles) est trop volumineuse (plus de 30 GO) pour être aisément consultée en ligne. Nous en prévoyons une publication sur clé USB, accompagnée d'un volume de commentaires.

1 Généralités

Le projet d'Atlas *linguistique audiovisuel du francoprovençal valaisan ALAVAL* est en voie d'élaboration au *Centre de dialectologie et d'étude du français régional* de l'Université de Neuchâtel depuis 1994[1]. Il s'agit d'un atlas linguistique proprement *audiovisuel* qui cherche à tirer pleinement profit des ressources que l'informatique a mises à notre disposition depuis une vingtaine d'années dans le domaine du traitement de la vidéo. Sauf erreur de notre part, c'est le premier projet d'atlas linguistique de cette nature, après les atlas purement sonores qui ont ouvert la voie, *l'ALD* (Goebl et al. 1998s.) ou le projet *Vivaldi* (Bauer 1995, Kattenbusch 1995, etc.).

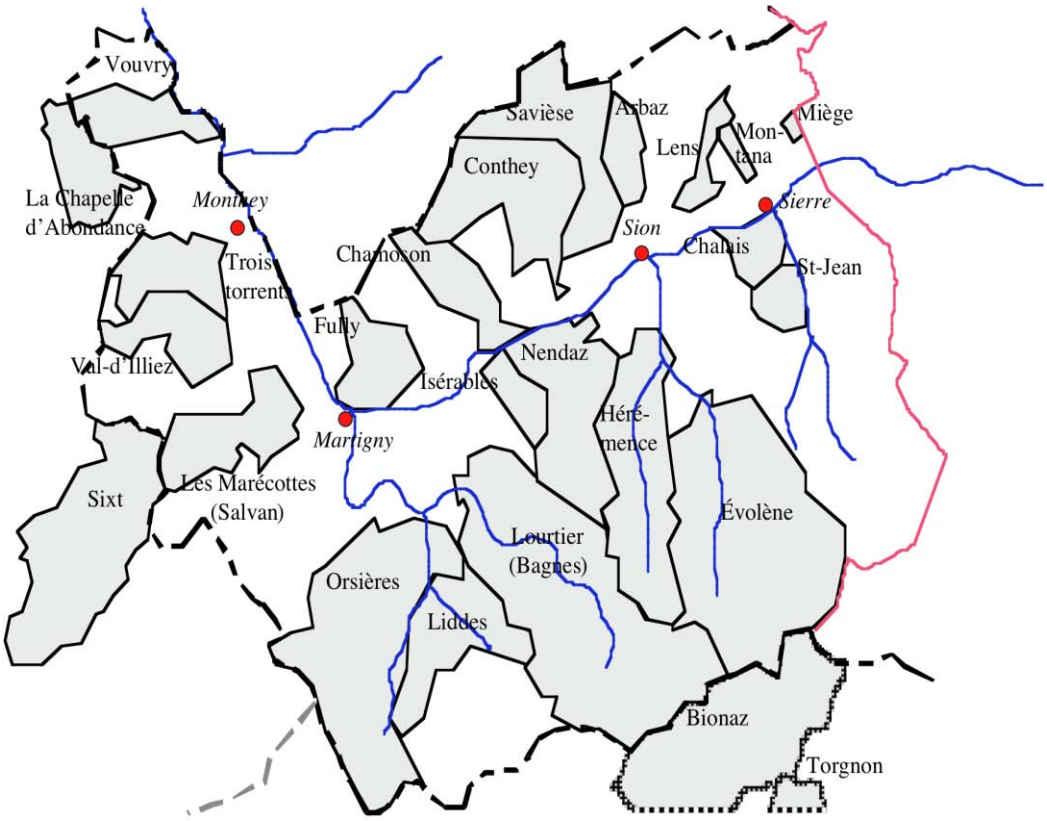
[1] Les travaux ont bénéficié du soutien matériel de la communauté européenne grâce à un projet Interreg en collaboration avec la Vallée d'Aoste et son propre projet d'atlas linguistique, ainsi que d'un financement par le *Fonds national de la recherche scientifique* suisse. Toutes les publications consacrées au projet sont disponibles en ligne à l'adresse http://www2.unine.ch/cms/lang/fr/pid/8174.



Carte n° 1: Le domaine d'enquête au sein de l'espace francoprovençal (carte d'après Tuaillon 1972)

Le projet *ALAVAL* est consacré à une «micro»-région de tradition linguistique franco-provençale située dans la haute vallée du Rhône en amont du lac Léman (en vert foncé sur la carte n° 1): c'est la partie romane du Canton du Valais (Suisse), la région la plus orientale (avec la Vallée d'Aoste) de l'espace galloroman, à la frontière linguistique avec l'allemand au nord et à l'est, et le piémontais au sud-est. C'est la dernière région de la Suisse romande où une telle entreprise est encore possible *et fait sens*: en Suisse romande, les parlers vernaculaires ont disparu presque partout depuis plus de 50 ans – et dans les deux autres régions où il nous reste des locuteurs (en Gruyère, canton de Fribourg, et en Ajoie, canton du Jura), la variation diatopique est faible, alors qu'elle est très marquée en Valais.

Nos enquêtes se sont déroulées entre 1994 et 2001. Depuis, nous travaillons à la transcription et à l'analyse linguistique des matériaux recueillis, ainsi qu'à l'élaboration des cartes.



Carte n° 2: Le réseau d'enquêtes de l'ALAVAL

Notre réseau comprend 25 points d'enquête (communes marquées en gris sur la carte n° 2) : 21 localités en Valais, auxquels s'ajoutent deux communes valdôtaines et deux communes en Haute-Savoie, pour interconnecter nos données avec les atlas linguistiques voisins. À chaque point d'enquête, nous avons enregistré deux témoins, une femme et un homme, avec deux caméras vidéo, analogiques au début, numériques vers la fin. Les enquêtes individuelles ont duré entre 3 et 5 heures. Tous les témoins ont été enregistrés chez eux, dans leur cadre familier, avec des effets très positifs pour la spontanéité de leurs réponses – c'était important pour nos objectifs – mais avec des problèmes parfois très nets pour la qualité du son et de l'image.

Actuellement, nous disposons d'un corpus entièrement numérisé d'environ 17000 énoncés de longueur très variable, environ 350 énoncés par témoin. Les travaux de transcription sont encore en cours, parallèlement à l'exploitation linguistique des matériaux. Environ deux tiers des matériaux sont transcrits.

2 Le questionnaire

Dès ses débuts, notre entreprise a été axée sur des questions de morphologie et de syntaxe. Cela s'explique par l'état de la recherche sur le francoprovençal qui possède depuis longtemps de bons travaux de phonétique historique – même s'il reste du travail à faire – et d'importants travaux de lexicologie et de lexicographie, alors que la description morphologique et syntaxique est restée peu développée. Pour de nombreuses questions, notre atlas constitue la première analyse scientifique de la morphosyntaxe du francopro-

vençal valaisan, qui ne possède aucune tradition grammaticale et qui n'a jamais été standardisé.

Notre questionnaire a été conçu dans une optique semi-ouverte, que nous comparons volontiers aux deux parties d'un concours de patinage artistique: les formes imposées (pour pouvoir récolter un corpus d'énoncés comparables) et l'expression libre. Pour obtenir un corpus de données relativement proche de l'usage naturel de nos témoins[2], nous avons veillé à formuler notre questionnaire dans un français régional aussi réaliste que possible, dans une langue proche de l'usage quotidien de nos informateurs. De même, nous avons évité de donner à notre questionnaire une tournure trop scolaire, en formulant des énoncés qui suscitent parfois le sourire – ou alors qui permettaient à nos informateurs de répondre avec une certaine distanciation ironique: on ne peut pas travailler avec des locuteurs dialectophones d'un certain âge comme on le ferait avec une population d'étudiants.

Pendant les enquêtes, pour éviter dans la mesure possible l'apparition d'artéfacts, nous avons complètement renoncé à «extorquer» les formes recherchées à nos témoins. Il s'agit là d'un véritable parti-pris méthodologique: si nos informateurs reformulaient leur réponse sans utiliser la forme que nous attendions, tant pis… nous passions à la prochaine question. C'est une des raisons pour lesquelles nous avions construit notre questionnaire sur le principe de la redondance systématique, ce qui devait nous permettre de recueillir quand même l'information recherchée. Par ailleurs, l'absence de certaines formes attendues dans certains parlers – et les stratégies d'évitement déployées par nos témoins – se sont parfois révélées comme hautement significatives d'un point de vue linguistique.

Mais la principale raison pour laquelle nous avons cherché à recueillir toutes les informations de manière redondante se trouve ailleurs. Quand on travaille sur une langue non standardisée telle que le francoprovençal, le phénomène de la variation interne est constant, dans tous les domaines du système linguistique. Par conséquent, si on veut documenter la variation interne de nos parlers, il faut s'en donner les moyens. Si on ne pose une question qu'une seule fois, on ne reçoit qu'une seule réponse… et c'est ce qui produit les cartes apparemment «homogènes» (absence de variation interne dans les parlers individuels) dans la plupart des atlas traditionnels.

Les mêmes phénomènes grammaticaux (et les mêmes éléments lexicaux) se retrouvent donc à plusieurs reprises dans différentes parties de notre questionnaire[3]. Ainsi, pour la 3e personne du pluriel de l'imparfait du verbe *acheter*, nous demandons à nos témoins de «traduire» (ou de reformuler à leur manière) l'énoncé suivant:

«Mes parents n'**achetaient** pas de jambon; nous avions des cochons nous-mêmes.»

Et ailleurs dans le questionnaire:

«Les gens **achetaient** les chaussures chez le cordonnier.»

[2] C'est ce que nous avons appelé «l'élaboration d'un corpus semi-spontané de langue naturelle» (Kristol 1998).

[3] Pour ne pas attirer l'attention de nos témoins sur les objectifs de notre démarche – tout en leur précisant que c'était leur langue qui nous intéresse, et qu'aucune question personnelle de nature «compromettante» ne serait posée – nous avons focalisé nos questions sur les réalités matérielles de la vie traditionnelle de l'espace alpin.

Et bien sûr, «les parents» et «les chaussures» se retrouvent dans deux ou trois autres phrases, et ainsi de suite. De plus, avec deux informateurs pour chaque parler, cela double l'information. Dans la mesure où nos témoins répondent effectivement aux questions comme nous l'avions prévu – en réalité, ils nous proposent souvent des formulations alternatives («*ah non, nous n'achetions pas les chaussures chez le cordonnier; c'est le cordonnier qui venait à la maison et les faisait*»[4]) – la redondance de l'information nous permet de dépasser le stade de l'information aléatoire. Et lorsque les deux témoins, à plusieurs reprises, nous donnent une forme inattendue, bizarre à première vue, nous avons le droit de conclure qu'il ne s'agit pas d'un simple *lapsus linguae*.

Relevons au passage que notre questionnaire est formulé de telle manière que chaque énoncé nous permet une exploitation multiple: rien que dans le premier exemple mentionné ci-dessus, on aborde la question du possessif (*mes parents*), de la négation (*n'achetaient pas*), de l'expression du partitif (*pas de jambon*), de la formation du pluriel (*mes parents, des cochons*), de la première et de la troisième personne du pluriel de l'imparfait (*avions, achetaient*) …

Sur cette base, nous avons commencé à établir des cartes cumulatives qui réunissent toutes les occurrences d'une même forme dans notre corpus. C'est que, sur une carte qui se contente de cartographier des énoncés individuels, il est impossible de se rendre compte de la variation. Seules des cartes cumulatives sont capables de rendre justice au fonctionnement réel des parlers que nous étudions.

Et le résultat a dépassé toutes nos attentes. On aurait pu penser que la redondance des questions, la répétition des mêmes formes était un «luxe». En réalité, elle est devenue une source d'informations de toute première importance, en particulier pour certains phénomènes linguistiques de base.

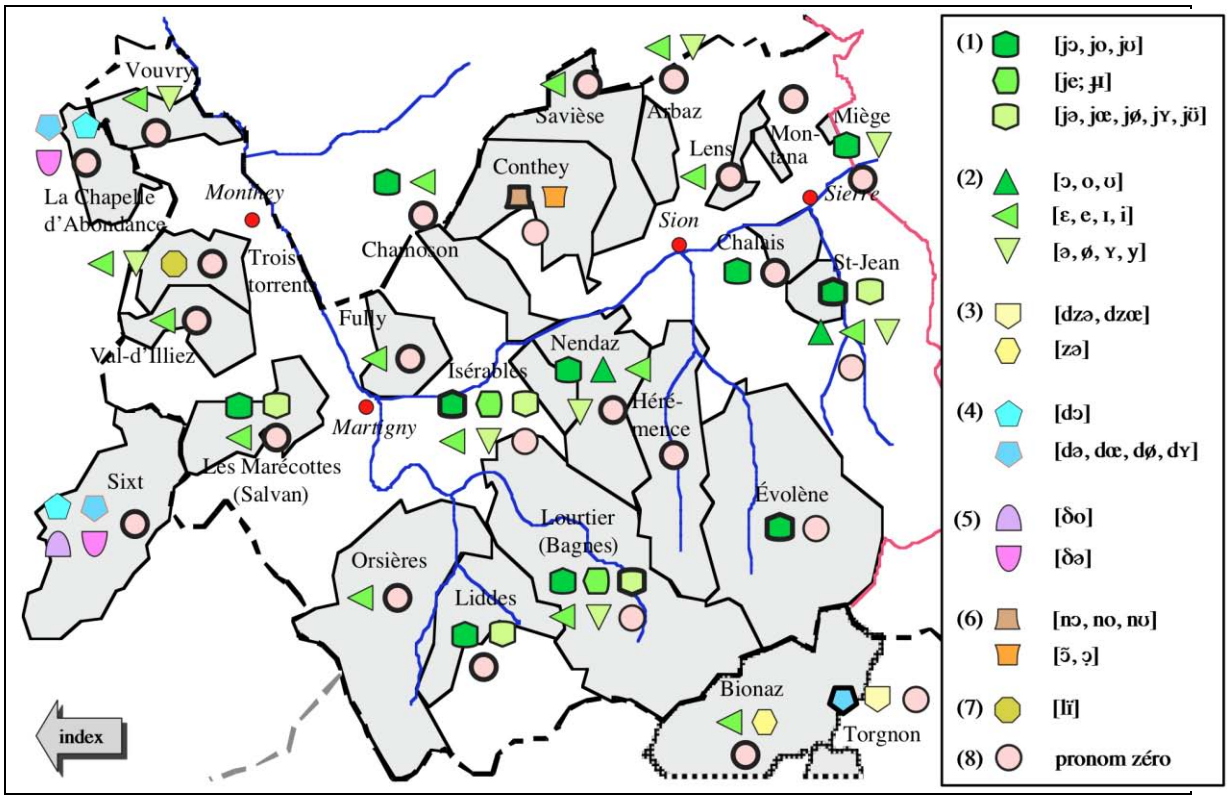
3 Le clitique sujet de la première personne du singulier

La carte n° 3 est le résultat de l'analyse de 2047 énoncés qui contiennent un verbe à initiale consonantique à la 1re personne du singulier (la carte des formes qui se trouvent devant un verbe à initiale vocalique serait très différente)[5].

Un exemple particulièrement significatif, c'est celui du parler d'Isérables où notre corpus comprend 79 phrases qui ont un verbe à initiale consonantique à la 1re personne du singulier. D'habitude, en dialectologie, on s'attend naturellement à la variation diatopique, qui distingue chaque dialecte local de ses voisins. Or, comme le montrent les exemples (1) à (10), la variation est également syntopique; elle se trouve à l'intérieur de chaque dialecte, et à l'intérieur de chaque locuteur.

[4] Cf. l'exemple suivant, enregistré à St-Jean (témoin féminin)
[dɛv'an nɔ prɛ̃ʒ'ẽŋ tɔz'ɔ̃ lɔ kɔrdɔnj'ɛ a - ɑ miʒ'ɔ̃ - pwɛ wï prɛ̃ʒ^{j}'ɛ mɛʒ'ura a - a la fam'ıł - pwɛ faʒj'ɛvɛ lɛ bw'ɔtɛ ɛ̃ntʃ'ɛ nɔ]
Autrefois nous prenions toujours le cordonnier à .. à la maison .. puis il prenait mesure à .. à la famille .. puis il faisait les souliers chez nous.

[5] Les trois cartes qui font l'objet d'un commentaire dans cette contribution peuvent être consultées en ligne sur le site du Centre de dialectologie de l'Université de Neuchâtel, à partir de l'adresse https://www2.unine.ch/dialectologie/page-8174.html. La consultation demande cependant beaucoup de patience: la carte du clitique sujet de la 1re personne du singulier comprend 107 clips vidéo, celle de l'article défini pluriel (point 6, ci-dessous) 264 clips. Le temps de téléchargement des données sera long. C'est la raison pour laquelle nous pensons distribuer la version définitive de l'atlas sur une clé USB, qui accompagnera le volume de commentaires.



Carte n° 3 : l'expression du «je»

À Isérables, les formes qui correspondent au «je» français sont multiples – mais elles peuvent aussi manquer complètement. Ainsi, notre corpus comprend:

• 40 occurrences pour [jɔ], [jʊ] (et quelques allophones)

(1) jo bˈaːlɔ̈ pɐst sːɪ aɲˈajɐ
Je bâille parce que je suis fatiguée.

(2) jʊ̈ mːɐ̯ sy tˈʷɛːrsɑ ɑ tsəʊˈɪlːɛ
Je me suis tordue la cheville.

• 12 occurrences pour [jə], [jʏ] (et quelques allophones)

(3) jə vˈʏz œ̃nˈo mʊntˈaɲɐ
Je vais en haut à l'alpage.

(4) jö sːi mariˈaːjɐ
Je suis mariée.

(5) kɛ̃ jʏ rʏntrˈaʊɔ tˈarː i pˈar sə mətɛ todˈõ ɛ̃ rːˈadzə
Quand je rentrais tard, mon père se mettait toujours en colère.

• 1 occurrence pour [je]

(6) ɛ dzˈɔːɪ də fjˈeːtəː je mˈɛtɔ ɔ kostˈym pɔr ɑ ã mˈɛsːa
Les jours de fête .. je mets le costume pour aller à la messe.

• 8 occurrences pour [i], [ɪ], [ɛ] (voyelles antérieures)

(7) nˈa - i mə soʊeɲˈɛʊə pɑ mi d̥ saː - istwˈeːr
Non .. je ne me rappelais plus de cette .. histoire.

(8) ɛ̯ mɛt ɛː mɑ̃ sø ɛ z ˈaːnsɛ
Je mets les mains sur les hanches.

• 2 occurrences pour [ə]

(9) ə ʊˈëzə twˈɛ̯ːdr ˈɛɪ o ku

Je vais tordre lui le cou (rire).

• et 16 occurrences pour le clitique sujet zéro

(10) Ø mə sʊvənːdri tɔdˈɔ̃n də tɛ

Je me souviendrai toujours de toi.

Soulignons que, dans ce cas précis, tous les facteurs de variation mentionnés habituellement dans la recherche d'orientation sociolinguistique sont neutralisés. Toutes les réponses proviennent de la même informatrice. Elles ont été enregistrées en l'espace d'une demi-journée, au cours de la même enquête. La variation ne peut pas s'expliquer par un changement de situation ou de contexte. En plus, le [jo] de (1), le [jʊ] de (2), le [ɛ] de (8) et le [ə] de (9) proviennent de la même partie du questionnaire, consacrée thématiquement aux parties du corps. Par ailleurs, nous avons aussi testé d'autres facteurs de variation, et ils se sont révélé nuls: ainsi, comme le montrent les exemples (2), (7) et (10), la présence d'un pronom régime devant le verbe n'est pas pertinente pour la présence ou la forme du clitique sujet. Et bien sûr, on retrouve la même variation chez le témoin masculin.

4 Problèmes techniques ; présentation des données

On imagine facilement les problèmes que pose cette extraordinaire variation pour la cartographie de nos résultats. Si nous tentions de faire comme les atlas traditionnels, c'est-à-dire d'inscrire les différentes transcriptions phonétiques à l'endroit des points d'enquête, la carte deviendrait complètement illisible. C'est une des raisons pour lesquelles nous avons décidé de travailler avec des symboles.

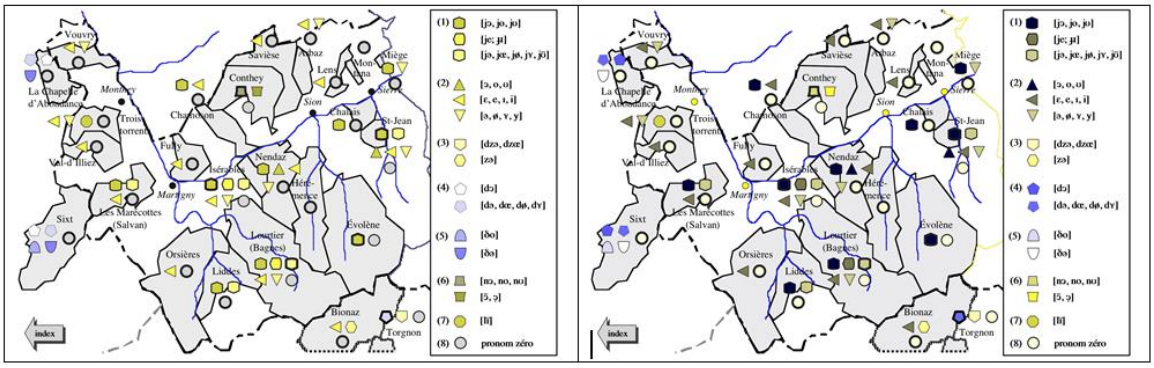
Au début de notre projet, nous avons simplement distingué les différentes formes sur la carte par des pastilles de différentes couleurs. Jusqu'au jour où un de nos étudiants[6] nous a rendus attentifs au problème des daltoniens. À ce moment-là, nous avons refait toutes nos cartes en doublant l'information «couleur» par la forme des symboles (et nous vérifions le résultat pour les deux types de vision daltonienne, les protanopes qui ne voient pas le rouge, et les deutéranopes qui ne voient pas le vert (cf. carte n° 4). En fait, grâce à la forme des symboles, même une carte en noir et blanc reste parfaitement lisible[7].

Mais ce n'est pas tout. En réalité, en définissant les formes et les couleurs de nos symboles, nous passons inévitablement de la multitude des formes attestées à un certain regroupement. Dans toutes les langues humaines, la variation individuelle en *discours* est infinie. Mais on sait bien que notre cerveau est capable d'identifier des unités discrètes dans le continuum des formes prononcées. Pour cartographier les données, il convient donc de passer à un niveau d'abstraction supérieur: nous cherchons à créer des familles de couleur et des familles de symboles relativement proches pour indiquer quelles sont les formes que nous proposons de regrouper; nous essayons de symboliser la parenté

[6] Nous tenons à remercier ici M. Manuel Riond qui a contribué ainsi de manière significative à la cartographie de notre atlas.

[7] Le polymorphisme des formes enregistrées (5 ou 6 variantes, voire davantage, pour de nombreux parlers) rend difficile, à notre avis, une cartographie «dynamique», géoréférenciée, de nos données par Google Maps, telle qu'elle est préconisée actuellement par plusieurs projets atlantographiques en cours. Seul le placement manuel des différents symboles dans l'espace cartographique garantit une lisibilité optimale des données pour l'ensemble de la carte (sans nécessiter des opérations de zoom sur les points individuels) et, lorsque cela s'impose, la possibilité d'indiquer des isoglosses de nature morphosyntaxique sur nos cartes (cf. à ce sujet la carte n° 8 ci-dessous).

plus ou moins grande des formes individuelles. Sur la carte n° 3, on distingue ainsi les formes qui commencent par [j] et les formes qui sont purement vocaliques. Dans les deux familles, la pointe du symbole va vers le haut pour les voyelles [o] et [ʊ], à gauche pour [e] et [i], et en bas pour les voyelles arrondies antérieures [y, ø], et on a adopté la même couleur pour les [je] et les [e]. Ensuite, il y a un symbole spécifique pour chacune des autres formes.



Carte n° 4: vision protanope **vision deuteranope**

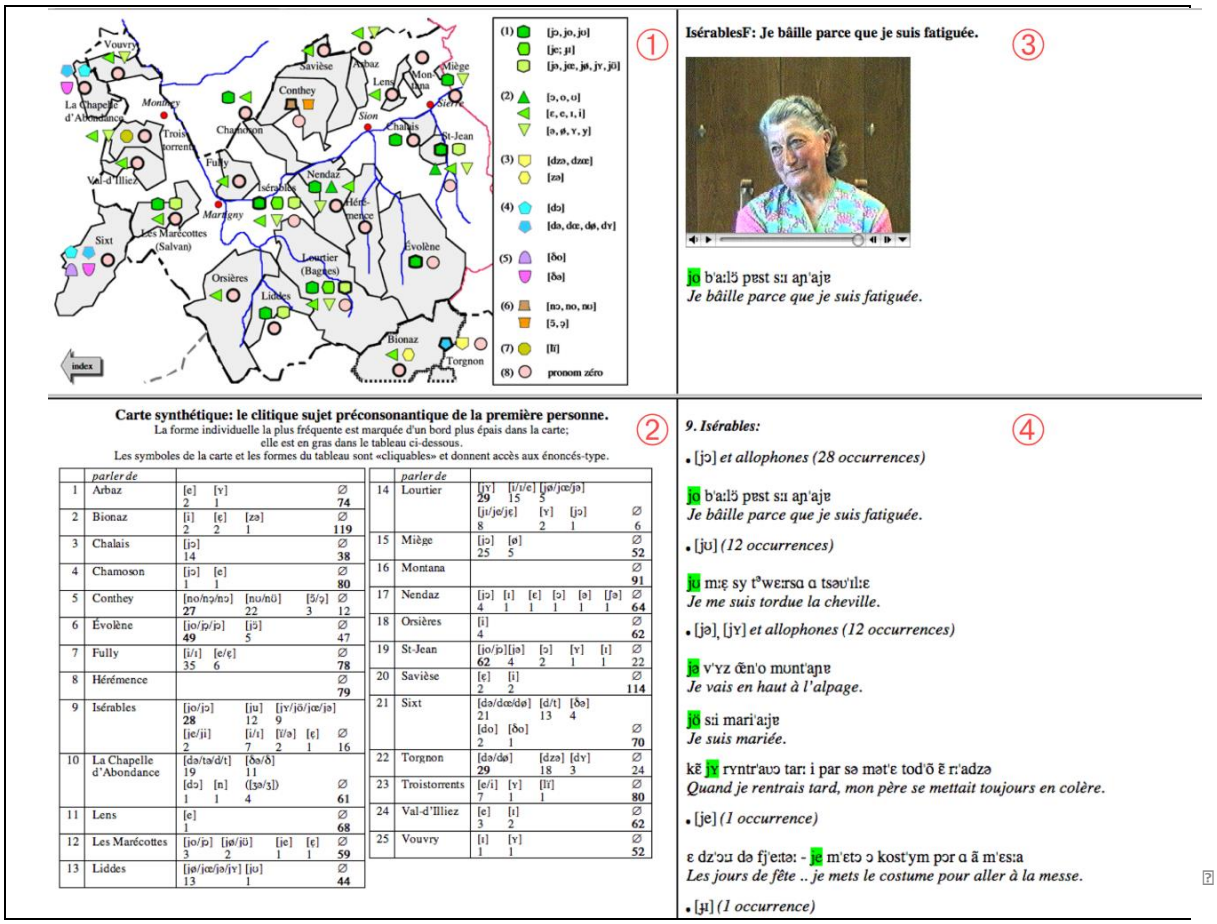
Tout ce travail se fait de façon manuelle, et c'est le travail du linguiste. Le logiciel capable de faire ce genre d'analyses n'a pas encore été développé. Chaque carte est le résultat d'un long tâtonnement et de discussions au sein de l'équipe, pour parvenir à un maximum de lisibilité et de pertinence linguistique. Notre principal effort est de mettre un ordre raisonnable et justifié dans les données observées. Nous sommes constamment obsédés par *une* question: non pas «*Est-ce qu'il y a* un système derrière la multitude des formes?», mais «*Quel est* le système qui permet de faire fonctionner, qui permet de gérer cette diversité dans la communication humaine, au sein de chaque communauté locale?»

Bien sûr, dans toute cette démarche, nous restons dialectologues, et nous restons attachés aux réalités observées. C'est la raison pour laquelle, à côté des cartes interprétées, nous présentons toujours aussi les matériaux bruts originaux: toutes les «pages» de notre atlas se composent de quatre «fenêtres» interconnectées, qui offrent un accès multiple aux données (carte n° 5).

- La fenêtre n° 1 («carte») contient la carte interprétée, cliquable, dans laquelle à chaque symbole graphique correspond un clip vidéo accompagné de sa transcription complète en API et d'une traduction littérale qui s'affichent dans la fenêtre n° 3.
- La fenêtre n° 2 («grille») comprend le titre de la page, un tableau avec la transcription phonétique et, lorsque c'est significatif, les indications statistiques précises pour l'élément sur lequel porte l'intérêt de la page. Chaque transcription individuelle de ce tableau est à son tour cliquable et renvoie à un clip vidéo qui s'affiche dans la fenêtre n° 3.
- La fenêtre n° 3 («clip») sert à afficher les clips vidéos individuels, avec les énoncés dialectaux transcrits, traduits et parfois annotés. Le titre du clip correspond à l'énoncé théoriquement prévu dans le questionnaire; en réalité, celui-ci est souvent très différent de l'énoncé réellement obtenu, qui se trouve dans la transcription phonétique. La reproduction de l'image et du son com-

mence automatiquement à l'ouverture de la fenêtre; elle peut être répétée à volonté.

- La fenêtre n° 4 («liste») présente la liste complète des énoncés qui ont servi à l'élaboration de la carte, dans l'ordre alphabétique des localités. Chaque transcription est à son tour cliquable; lorsque c'est judicieux, l'élément soumis à l'analyse est mis en relief par la couleur du symbole qui lui correspond dans la carte.



	parler de								
1	Arbaz	[e] 2	[ʏ] 1					∅ 74	
2	Bionaz	[i] 2	[ɛ] 2	[zə] 1				∅ 119	
3	Chalais	[jɔ] 14						∅ 38	
4	Chamoson	[jɔ] 1	[e] 1					∅ 80	
5	Conthey	[nɔ/nɔ̃/nɔ] 27	[nʊ/nʊ̃] 22	[ɔ̃/ɔ] 3				∅ 12	
6	Évolène	[jo/jɒ/jɔ] 49	[jɔ̃] 5					∅ 47	
7	Fully	[i/ɪ] 35	[e/ɛ] 6					∅ 78	
8	Hérémence							∅ 79	
9	Isérables	[jo/jɔ] 28	[jʊ] 12	[jv/jʊ̈/jœ/jə] 9					
		[je/ji] 2	[i/ɪ] 7	[ï/ə] 2	[ɛ] 1			∅ 16	
10	La Chapelle d'Abondance	[də/tə/d/t] 19	[ðə/ð] 11						
		[dɔ] 1	[n] 1	([jə/ʒ]) 4				∅ 61	
11	Lens	[e] 1						∅ 68	
12	Les Marécottes	[jo/jɔ] 3	[jø/jʊ̈] 2	[je] 1	[ɛ] 1			∅ 59	
13	Liddes	[jə/jœ/jə/jʏ] 13	[jʊ] 1					∅ 44	

	parler de								
14	Lourtier	[jʏ] 29	[i/ɪ/e] 15	[jə/jœ/jə] 5					
		[jɪ/je/jɛ] 8	[ʏ] 2	[jɔ] 1				∅ 6	
15	Miège	[jɔ] 25	[ø] 5					∅ 52	
16	Montana							∅ 91	
17	Nendaz	[jɔ] 4	[ɪ] 1	[ɛ] 1	[ɔ] 1	[ə] 1	[jə] 1	∅ 64	
18	Orsières	[i] 4						∅ 62	
19	St-Jean	[jo/jɒ][jə] 62	[ɔ] 4	[ʏ] 2	[ɪ] 1	[ɪ] 1		∅ 22	
20	Savièse	[ɛ] 2	[i] 2					∅ 114	
21	Sixt	[də/dœ/dø] 21	[d/t] 13	[ðə] 4					
		[dɔ] 2	[ðo] 1					∅ 70	
22	Torgnon	[də/dø] 29	[dzə] 18	[dʏ] 3				∅ 24	
23	Troistorrents	[e/i] 7	[ʏ] 1	[lï] 1				∅ 80	
24	Val-d'Illiez	[e] 3	[ɪ] 2					∅ 62	
25	Vouvry	[ɪ] 1	[ʏ] 1					∅ 52	

Carte n° 5 : Les quatre «fenêtres» des cartes de l'*ALAVAL*

Bien entendu, comme tous les atlas multimédias de dernière génération, l'*ALAVAL* permet à l'utilisatrice et à l'utilisateur de ne jamais être obligé à faire une confiance aveugle aux transcripteurs: toutes les données originales – son et image – restent disponibles en permanence et peuvent être confrontées à la transcription proposée qui devient ainsi un simple outil de travail et ne constitue plus la dernière instance, invérifiable.

Techniquement, les pages de l'Atlas sont réalisées de façon modulaire en langage HTML élémentaire (avec l'inclusion d'une routine en JavaScript), pour éviter tout problème de compatibilité avec les différents navigateurs Internet actuellement disponibles. Pour les clips vidéo, nous avons adopté le format QuickTime™ d'Apple qui était le seul disponible à l'époque où nous avons commencé nos travaux[8]. Les transcriptions ont été

[8] Chose rare en informatique – 20 ans, c'est presque une éternité – ce format est toujours disponible, largement répandu et peut être téléchargé gratuitement pour tous les systèmes d'exploitation actuellement sur

réalisées à l'aide de plusieurs logiciels d'analyse du son (SoundEdit, BiasPeak, et plus récemment Amadeus Pro 2) qui présentent tous la particularité de ne pas dissocier la piste vidéo de la piste son; la possibilité de vérifier sur l'image la gestuelle des témoins et en particulier la forme et la position des lèvres nous a souvent aidés à mieux comprendre les énoncés spontanés et à désambiguïser des séquences sonores qui posaient problème[9].

Tous les éléments individuels qui composent les pages de l'atlas (carte, tableau des formes[10], mais aussi les transcriptions) sont réalisés au format GIF (Graphical Interchange Format) «pixellisé», pour éviter tout problème de compatibilité: ils peuvent être lus sur n'importe quel système d'exploitation actuellement disponible (Mac, Windows, Linux), sans demander aux utilisateurs d'installer sur leurs machines la police Unicode spécifique que nous avons développée pour nos besoins[11].

5 La première personne du présent du verbe *avoir*

Si dans la carte du «je» présentée ci-dessus, l'analyse ne dépasse pas le niveau phénoménologique parce que nous n'avons pas encore réussi à découvrir un principe d'organisation systémique permettant de gérer cette variation, il est possible d'identifier certains secteurs du système morphosyntaxique de nos parlers qui admettent une analyse plus poussée, malgré la profonde variation interne qui les caractérise. C'est ce que nous montrerons ci-dessous sur la base de deux exemples concrets.

Comme l'a montré le polymorphisme du «je» à Isérables (carte n° 3), l'emploi du clitique sujet de la 1re personne est facultatif en francoprovençal valaisan (avec une fréquence d'emploi variable selon les parlers, et parfois, au sein d'un même parler, entre nos deux témoins).

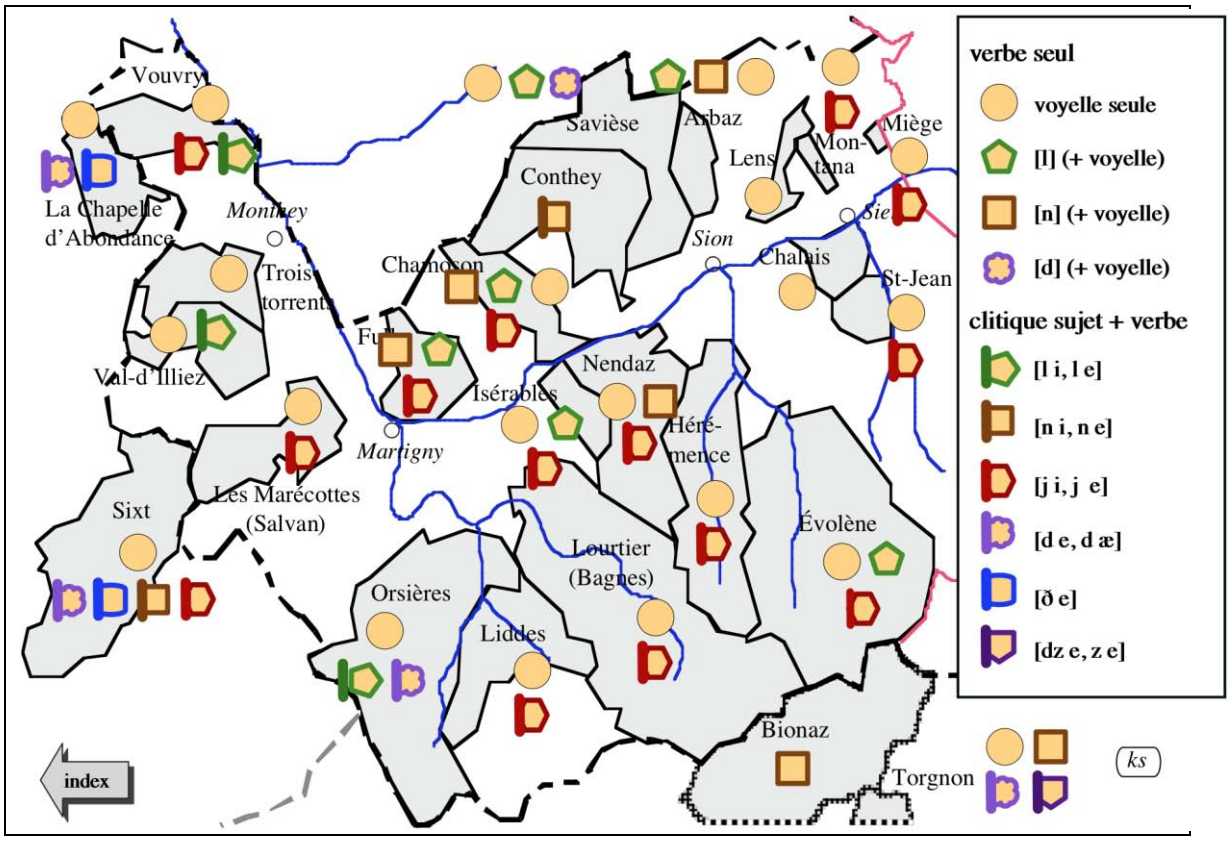
La carte n° 6 ci-dessous illustre la situation de la 1re personne de l'indicatif présent pour le verbe auxiliaire *avoir*. Cet exemple nous permettra d'illustrer à quel point l'analyse morphosyntaxique individuelle de chaque parler est essentielle.

le marché. Malgré plusieurs changements d'infrastructure technique (processeurs, supports de stockage), nous n'avons jamais dû transcoder nos données pour adopter un nouveau format.

[9] Certains collègues nous ont demandé si au moment de la transcription, nous procédions à un balisage grammatical de nos données. Notre réponse a été négative, pour des raisons qui apparaîtront dans ces lignes: chaque parler de notre domaine possède sa propre grammaire que nous sommes en général les premiers à analyser. Or, l'analyse grammaticale d'un phénomène déterminé n'est possible qu'au moment où nous disposons de la totalité des transcriptions pour un parler donné.

[10] Les cartes et les tableaux constituent des «image maps» (statiques) cliquables.

[11] Cette police qui – outre l'alphabet de l'API – nous permet de citer les graphies originales de l'*Atlas linguistique de la France ALF* (système Rousselot-Gilliéron), les *Tableaux phonétiques des parlers de la Suisse romande* (GAUCHAT et al. 1925), le système d'écriture particulier du *Glossaire des patois de la Suisse romande GPSR* et la transcription de l'*Atlas linguistique de l'Italie et de la Suisse méridionale AIS* (système Boehmer-Bourciez) peut être téléchargée librement sur le site du Centre de dialectologie de l'Université de Neuchâtel, à l'adresse https://www2.unine.ch/dialectologie/presentation.



Carte n° 6: Les formes de la 1re personne présent singulier d'*avoir*

	parler de							
1	Arbaz	le 34	l 1	ne 3	ni 1	e 4		tot. 43
2	Bionaz	nɪ 37	nĩ 3	nɛ 1	ny 1			tot. 42
3	Chalais	e 12	ɛ 2	eɪ 2	i 4			tot. 19
4	Chamoson	nɪ 24	ne 2	lɪ 4	j i 2	ɪ 1		tot. 33
5	Conthey	n ɪ 37						tot. 37
6	Évolène	j e 15	e 13	ɪ 9	lɪ 1			tot. 37
7	Fully	nɪ 24	ne 10	ny 1	n 1	lɪ 2	j i 1	ɪ 1 tot. 40
8	Hérémence	ɪ 30	e 2	ei 1	j i 2			tot. 35
9	Isérables	j e / j ɛ 34		j i 1	e 4	i 1	le 1	tot. 41
10	La Chapelle d'Abondance	d e 22	ð e 3	i 1	e 3	ɛ 1		tot. 30
11	Lens	e 31	ɪ 2	ɛ 1				tot. 34
12	Les Marécottes	j e / j ɛ 34		j ɪ 3	i 1			tot. 38
13	Liddes	j e 17	e 9	ɪ 7	j ʏ 2	j ɪ 1		tot. 36

	parler de							
14	Lourtier	j e 18	j ɪ 11	j ə 5	i 3	e 2		tot. 39
15	Miège	e / eɪ 22	i 11	j ɪ 6	j e 3			tot. 42
16	Montana	e / ɛ 29	i 2	j e / i e 8				tot. 39
17	Nendaz	ɪ 35	e 4	ni 5	j e / j i 2			tot. 46
18	Orsières	l i 27	l e 1	d i / d e 3	ɪ 3			tot. 34
19	St-Jean	j e / j ɛ 19	e 16	ɪ 3				tot. 38
20	Savièse	i 23	e 8	y 1	le / leɪ 10	j e 2		tot. 44
21	Sixt	d ɛ / d æ 37	ð ɛ 15	n ɛ 1	j ɛ 1	ɛ 2		tot. 56
22	Torgnon	d e 22	dz e 6	n ɪ / n e 7	i / e 3			tot. 38
23	Troistorrents	i 28	e 11	y 1				tot. 40
24	Val-d'Illiez	i 26	e 11	l e 2				tot. 39
25	Vouvry	j e 10	l e 6	e 9	i 2	œ 1	ø 1	tot. 29

- La plupart des parlers sont caractérisés par la possibilité d'exprimer le «j'ai» par une simple voyelle; dans certains parlers (Lens, Chalais, Troistorrents), c'est la seule solution attestée. Le plus souvent, il s'agit d'une voyelle antérieure comprise entre [i] et [ɛ], [y] et [ə], mais on trouve aussi des réalisations en [æ], et la variation interne à chaque parler, quant à la nature de la voyelle, est souvent importante. Pour que la carte reste lisible, nous avons dû renoncer à indiquer le timbre de la voyelle; sinon, le nombre de variantes pour chaque parler individuel devenait ingérable. En revanche, cette précision se trouve bien sûr dans le tableau qui accompagne la carte.
- À côté de cette forme purement vocalique, la plupart des parlers possèdent pourtant aussi des formes précédées d'une consonne, et c'est ici que les choses se compliquent. Quant à la nature de l'élément prévocalique, il peut s'agir de plusieurs consonnes dentales/alvéolaires ([d, ð, n, l]), de [dz, z] et de [j]. Alors que [d, ð, dz, z] et [j] ont sans doute une origine étymologique (< lat. EGO; cf. MARTIN 1974 et KRISTOL 2009), [n] et [l] sont d'origine analogique. Or, selon les parlers, une même séquence phonétique telle que [ni], [ne] ou [le] doit être interprétée soit comme «verbe seul» (c'est-à-dire avec une consonne agglutinée), soit comme «clitique+verbe».

Nous nous limiterons ici à un seul cas de figure, à savoir les voyelles précédées de la consonne [n]. Dans les différents parlers, la séquence phonétique [ni] ou [ne] (avec quelques allophones : [nɪ], [nɛ], etc.) doit être analysée de *quatre* manières différentes :

1° Le [n] peut correspondre au *nous* de la première personne du pluriel, transféré de manière analogique à la première personne du singulier aussi. Le seul parler valaisan qui connaît ce phénomène, c'est celui de Conthey, où on dit:

(11) jëʁ aprˈï mœ̣- mïødzˈɔː **no** ʃi itˈɑː ʏ fɛ̃ (ContheyF)
[littéralement :] *Hier après-mi.. midi NOUS suis été aux foins.*

Devant verbe à initiale vocalique, [no] se réduit à [n][12], et par conséquent on dit :

(12) **n** adzˈøtɔ o aθˈe a a lɛterˈi (ContheyM)
J'achète le lait à la laiterie.

Lorsqu'on trouve ce même [n] à la 1re personne du verbe *avoir*, on conclura donc sans hésitation qu'il s'agit du même clitique sujet: c'est la forme réduite, prévocalique, qui correspond à la forme pleine [no], caractéristique pour le parler de Conthey VS :

(13) **n** ɪː θˈɛ̃kɑ̃t θɛ̃k ɑ̃ (ContheyF)
J'ai cinquante-cinq ans.

2° Dans plusieurs parlers, le [n] reflète l'adverbe pronominal qui correspond au *en* du français (< lat. INDE). Ce cas de figure peut être illustré par le parler d'Évolène. Celui-ci possède un clitique sujet facultatif [jɔ] (avec quelques variantes allophoniques) devant consonne, et [j] devant voyelle. Ce pronom apparaît dans un peu moins de la moitié des occurrences (env. 44%).

(14) **jʊ̯** vˈeʒœ̣ frɑ̃ŋ ˈɔrə ok sɛ̣lˈi (ÉvolèneM)
Je vais droit maintenant à la cave.
(15) **j** atsˈëtə lɔ lasˈe ɐ̨ la lœ̣tɐ̨rˈik (ÉvolèneF)
J'achète le lait en la (à la) laiterie.

[12] Nous n'avons trouvé aucune mention de ces formes chez MARZYS 1964. Elles sont en revanche bien attestées dans les *Tableaux phonétiques* (GAUCHAT et al. 1925, col. 47, 109, 192, 340, etc.).

Devant les formes du verbe *avoir*, on trouve le clitique prévocalique [j] (avec 15 occurrences dans notre corpus) ou la forme verbale seule (22 occurrences).

(16) **j** e ʃɔʃɑ̃tɕ un an (ÉvolèneF)
J'ai soixante et un ans.
(17) **e** ʒ'ami vjuk siŋ ʒ ɔʃ d yŋ kɔ (ÉvolèneF)
Je n'ai jamais vu cinq ours d'un coup.

Par conséquent, dans ce parler, lorsqu'on relève un [n] devant la première personne du verbe *avoir*, il ne s'agit en aucun cas d'un clitique sujet, mais de l'adverbe pronominal qui correspond au français *en* (< lat. INDE) :

(18) dë ᵗsaɕ'æɲə **n** ɪ ʒ ʊ mɣ̈ʒʲ'a (ÉvolèneF)
*Des châtaignes, j'**en** ai eu mangé.*

3° Le [n] peut représenter le *ne* de la négation. C'est le cas des parlers du Chablais valaisan[13] qui ont conservé un *ne* de négation facultatif devant le verbe, à la différence des parlers du Valais central qui utilisent des formes du type ['viɲɔ pɑ] «je viens pas», sans «ne», comme en français parlé ordinaire. Par conséquent, dans ces parlers, un [n] devant une forme verbale de la première personne à initiale vocalique peut représenter soit la négation *ne*, soit le *en* de INDE, comme à Évolène.

Concrètement, le parler de Val-d'Illiez possède un clitique sujet rare qui est [i] ou [e] en position préconsonantique (5 occurrences sur 67 attestations), [l] ou [j] en position prévocalique (9 occurrences sur 44 attestations) :

(19) v'ʏzɔ ɑ lɑ mɔ̃t'aɲa (Val-d'IlliezF; clitique sujet zéro)
Je vais à l'alpage.
(20) ɪ wa k lö gam'ɛ̃ ɛːɑ y ji døv'ã ny œʀ (Val-d'IlliezF)
Je veux que les enfants aillent au lit avant neuf heures.
(21) amɛʀ'i bɛ̃ tʊʁn'a vɛʁ - ma mɪz'ɔ̃ jɔ sa neː (Val-d'IlliezF[14])
J'aimerais bien revoir.. ma maison où je suis née.
(22) kɑ̃ **j** ɪʁɔ gɑm'in **l** am'ɑvɔ pa lö z ɪpïn'ɑʁdɛ (Val-d'IlliezF)
Quand j'étais gamine j'aimais pas les épinards.

À la première personne du verbe *avoir*, on trouve la forme verbale seule (avec 37 occurrences dans notre corpus) et deux fois [l i]

(23) **i** lʏ kwe kə bɑ (Val-d'IlliezF)
J'ai le cœur qui bat.
(24) aː **l i** dikʀəv'ɛ na – ɔ̃ mwe də fʁöm'ɛ dã mɔ̃ kuʁt'i (Val-d'IlliezF)
Ah, j'ai découvert une .. un tas de fourmis dans mon jardin.

Par conséquent, lorsque nous rencontrons une forme [n i] (ou [n e]) dans ce parler, le [n] correspond sans doute à la négation *ne* ou à l'adverbe pronominal *en* :

(25) n **i** ʒam'i jy sɛ̃k uʁs d ɑ̃ k- d ɔ̃ k'uᵊ (Val-d'IlliezF)
Je n'ai jamais vu cinq ours d'un c.. d'un coup.
(26) le tɕet'aɲɛ n **e** pʀʏ ʑ y mẽjdʒʲ'a (Val-d'IlliezF)
Les châtaignes j'en ai assez eu mangé

4° Dans plusieurs parlers, enfin, le [n] peut être une simple consonne agglutinée, sans la moindre fonction morphologique. C'est ce que l'on observe par exemple dans les parlers d'Arbaz ou de Bionaz. Ces deux parlers possèdent un clitique sujet de la première personne du singulier dont l'emploi est extrêmement rare. À Arbaz, sur 77 formes ver-

[13] Le Chablais est la région située immédiatement au sud de l'embouchure du Rhône dans le lac Léman.
[14] Clitique sujet zéro devant verbe à initiale vocalique ([amɛʀ'i] 'j'aimerais') et consonantique ([sa] 'je suis').

bales à initiale consonantique, nous n'avons enregistré que 3 occurrences d'un clitique [ʏ] ou [e]. À Bionaz, sur 70 formes verbales à initiale consonantique, nous n'avons trouvé que 4 occurrences d'un clitique sujet [ɐ̯] ou [i]. Dans les deux parlers, aucune forme verbale à initiale vocalique n'est précédée d'un élément potentiellement pronominal, à l'exception notable de la première personne du verbe «avoir».

L'analyse détaillée des occurrences de «j'ai» (cf. le tableau de la carte n° 6) donne le résultat suivant :

	parler de						
1	Arbaz	le	l	ne	ni	e	tot.
		34	1	3	1	4	43
2	Bionaz	nɪ	nĩ	nɛ	nʏ		tot.
		37	3	1	1		42

À Bionaz, la totalité des formes attestées présente un [n] initial. La situation est un peu plus complexe à Arbaz, où nous trouvons 35 formes avec [l], 4 formes avec [n] et 4 formes sans consonne initiale.

Face à ce résultat, il n'y a que deux explications possibles, à notre avis :

a) soit le verbe *avoir*, dans ces deux parlers, possède un comportement morphosyntaxique complètement atypique, ayant généralisé à la première personne du singulier un clitique sujet [l] ou [n] qui n'apparaît devant aucun autre verbe;

b) soit il se comporte comme tous les autres verbes, mais la forme a été «renforcée» par l'agglutination d'une consonne initiale qui est évidemment d'origine analogique.

Ce qui nous fait surtout pencher pour la deuxième interprétation, outre l'invraisemblance de l'hypothèse (a), c'est l'observation suivante :

Dans les parlers où, selon notre analyse, une consonne prévocalique représente un clitique sujet, elle peut être séparée du verbe par un autre élément proclitique, en particulier par des pronoms régimes, comme dans l'exemple suivant :

(27) de blˈɛtɛ **d** ã **e** y plãto (La Chapelle d'AbondanceF)
Des blettes j'en ai eu planté.

Par contre, dans les parlers dans lesquels la consonne s'est agglutinée au verbe, cette agglutination semble inhiber l'apparition d'un autre élément proclitique :

(28) dɛ̈ tsahˈanʏ **le** ʒ y mĩzjˈa (ArbazF)
Des châtaignes j'ai eu mangé.

Alternativement, dans les cas analogues, le parler de Bionaz rejette carrément tous les clitiques objets à la fin du groupe verbal :

(29) də tsatˈɑɲɛ **nɪ** ajˈaw{ɔ} mœndzˈʊ nɛ̃ (BionazF)
*Des châtaignes j'ai eu mangé **en**.*

(30) **ny** baʎˈa læɪ də tsˈɪːflɑ ɑ̃ŋ tsˈɪːflɑ (BionazM)
*J'ai donné **lui** de gifle une gifle.*

Voilà pourquoi, dans la carte de l'Atlas qui résulte de cette analyse, une même séquence peut recevoir une interprétation divergente, selon les parlers. Et on comprend pourquoi il nous a été impossible de baliser notre corpus au moment de faire les transcriptions. Au-delà du polymorphisme qui caractérise nos parlers, nous sommes cons-

tamment confrontés au fait qu'un même élément, apparemment, reflète en réalité plusieurs phénomènes linguistiques différents.

6 L'article défini pluriel

Il y a quelques mois, après avoir dépouillé la totalité de notre corpus, nous nous sommes trouvés confrontés à une liste d'environ 950 occurrences de l'article défini pluriel, masculin et féminin (environ 40 occurrences pour chaque parler), ce qui – en faisant abstraction de la fréquence des formes individuelles – nous a donné le tableau suivant des formes observées brutes, encore sans la moindre interprétation, mis à part les couleurs qui permettent de regrouper plus facilement les formes qui se correspondent:

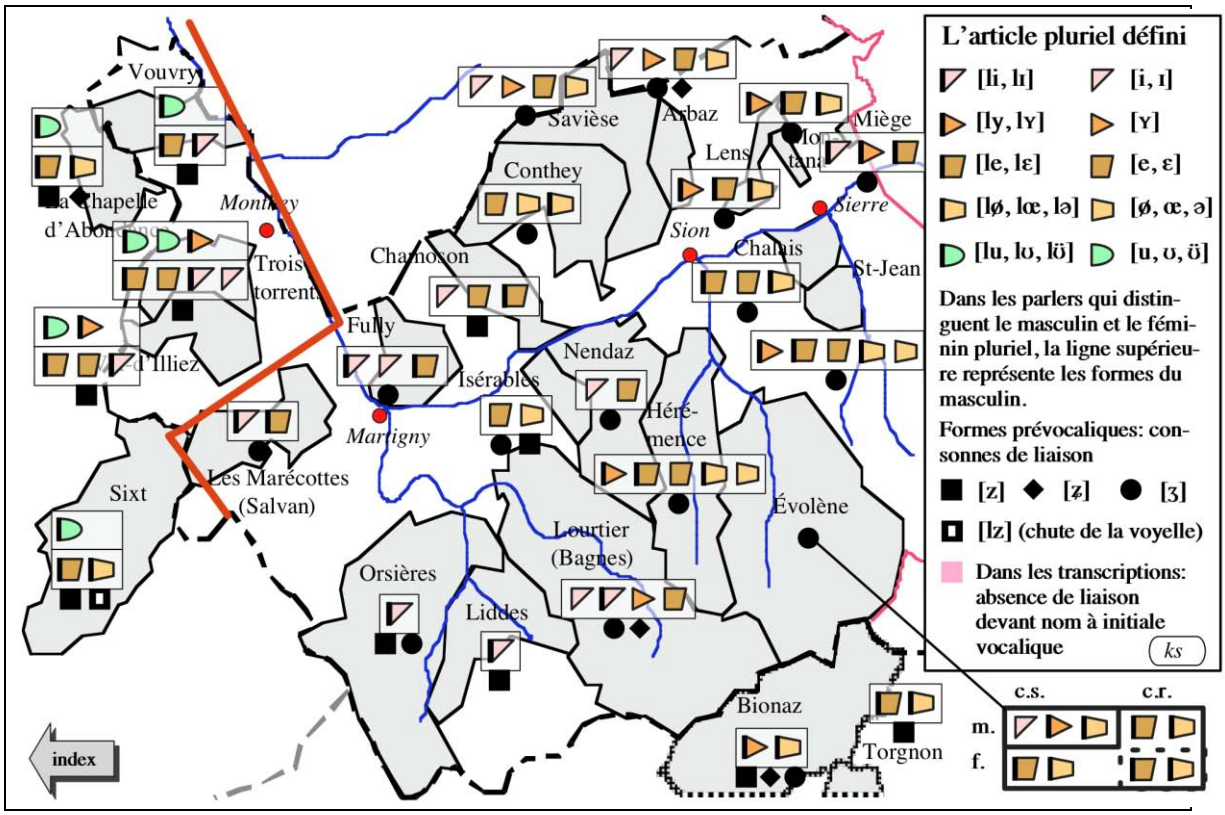
	parler de	témoin fém.	témoin masc.
1	Arbaz	ɪ ʏ e̞ ɛ̞ œ ə	e̞ ɐ̞ ɛ̞ ø œ
2	Bionaz	lʏ lø lœ lə	lʏ lø lə
3	Chalais	lɐ̞ ɐ̞ lœ	le lɐ̞ lɐ̈ lœ
4	Chamoson	ɪ ɪ̈ e̞ ɐ̞ lɛ	e̞ ɛ̞
5	Conthey	e̞ ɛ̞	ɐ̞ œ lœ
6	Évolène	i lʏ lɛ lø lœ lə	lʏʒ lɛ lø øʒ lœ lə
7	Fully	lɪ i le	lɪ
8	Hérémence	lɛ ɛ̞ lœ lə	lʏ le ɛ̞ lœ lə əʒ
9	Isérables	ez ɛ œ ə	e̞ ɛ̞
10	La Chapelle d'Abondance	lʊ le lɛ	luz lʊ le lɛ lə
11	Lens	lɐ̞ lœʒ	lʏ leʒ lɛ lœ
12	Les Marécottes	lɪ le	lɪ le
13	Liddes	li	li
14	Lourtier	i lɪ	i ʏ eːz
15	Miège	lɪ lɛ	lʏ lɛ
16	Montana	lʏ leʒ lɛ lœ	lɛ
17	Nendaz	e̞ ɛ̞	iʒ e̞ ɛ̞
18	Orsières	li	li
19	St-Jean	lʏ lɛ ɛ̞ lə əʒ	lʏ lɛ lœ
20	Savièse	i ʏ ɛ̞ øʒ œ əʒ	ɪ ʏ e̞ ɛ̞ øʒ
21	Sixt	luz lʊ le løz lz	lu lʊ le lz
22	Torgnon	le lɛz lə	lɛ lœ lə
23	Troistorrents	luz lʊ lʏ le	lu u lʊz ʊz lʏ lɪ ɪ le e̞
24	Val-d'Illiez	lu lʊz lʏ lɪ le ez	luz lʊ le lɐ̞
25	Vouvry	lʊ lʊz lɪ le	luz lʊ lɪ le

Tableau n° 7: l'article défini pluriel (masculin et féminin); corpus global des formes observées

Existe-t-il une logique derrière cette prolifération de formes, un système (ou des systèmes) de nature linguistique qui permettent de structurer les données? Dans ce cas précis, la réponse est positive, du moins en partie.

1° En ce qui concerne les formes en [u] marquées en vert dans le tableau, il est facile de voir, en examinant les énoncées transcrits, qu'il s'agit d'un article masculin pluriel (cf. 31), alors que la majorité des autres formes, dans les mêmes parlers, précède un substantif féminin (32):

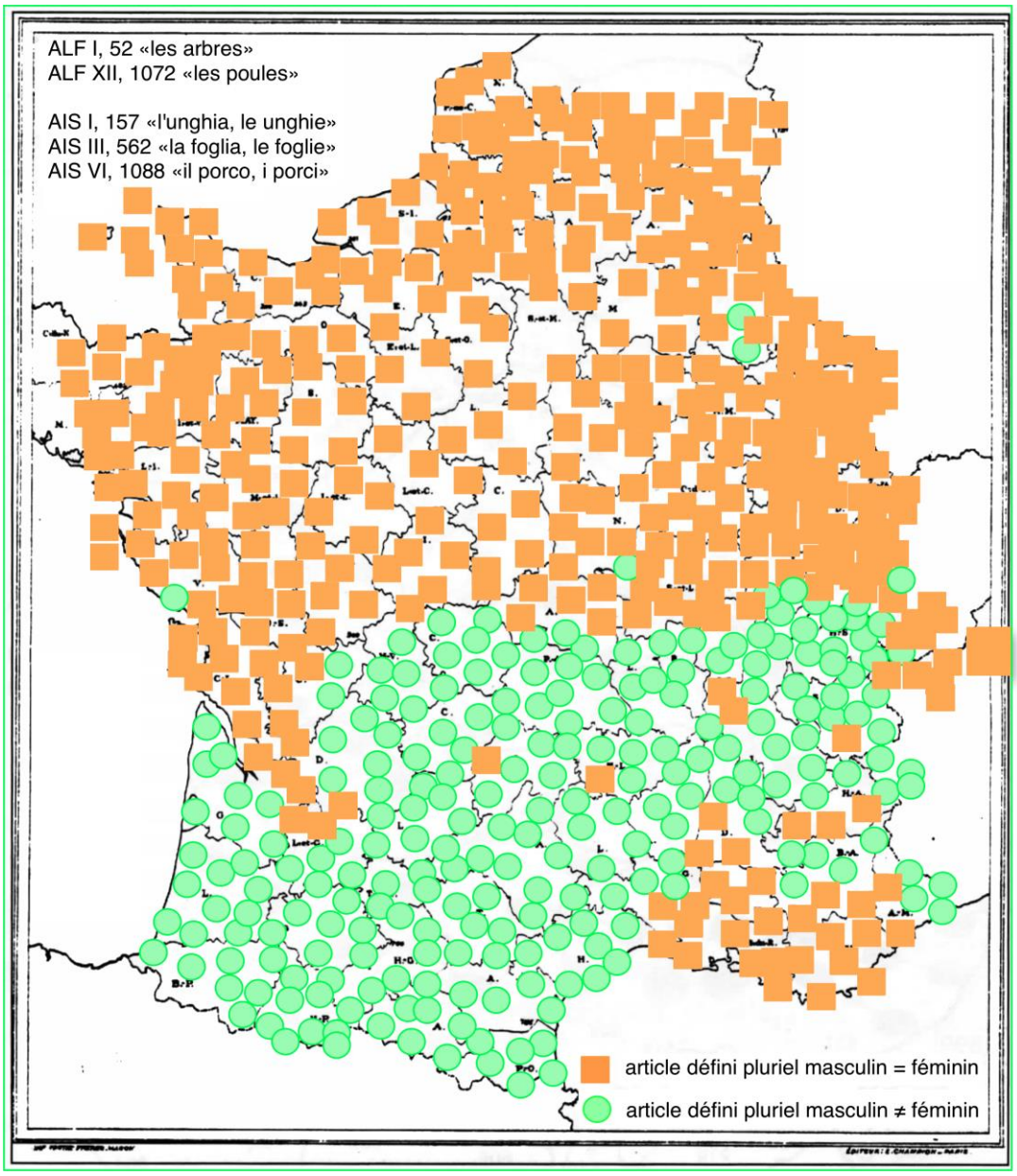
(31) ɔ̃ **lu** tɑsˈɔ̃ l ɑ̃ fe dʏ mˈoːʊ dɛ̃ː **lu** tsɑ̃ (TroistorrentsM)
Oh les blaireaux ont fait du mal dans .. les champs.



Carte n° 8 : le polymorphisme de l'article défini pluriel

	parler de	témoin fém.	témoin masc.
1	Arbaz	ɪ ʏ e̯ ɛ̯ œ ə	e̯ ɛ̯ ɛ ø œ
2	Bionaz	lʏ lø lœ lə	lʏ lø lə
3	Chalais	lɛ̯ ë lœ	le lɛ̯ lë lœ
4	Chamoson	ɪ ï e̯ ɛ̯ lɛ	e̯ ɛ
5	Conthey	e̯ ɛ	ɛ̯ œ lœ
6	Évolène suj. m.	i lʏ	lʏʒ lø lœ
	rég. m.	lɛ lœ	lɛ lø
	suj. f.	lɛ lœ	lœ lə
	rég. f.	lɛ løʒ lœ̯ lə	lɛʒ løʒ øʒ
7	Fully	lɪ i le	lɪ
8	Hérémence	lɛ ɛ lœ lə	lʏ lɛ ɛ lœ lə əʒ
9	Isérables	ez ɛ œ ə	e̯ ɛ
10	La Chapelle d'Abondance m.	lʊ	luz lʊ
	f.	le lɛ	le lɛ lə
11	Lens	lɛ̯ lœʒ	lʏ leʒ lɛ lœ
12	Les Marécottes	lɪ le	lɪ le

	parler de	témoin fém.	témoin masc.
13	Liddes	li	li
14	Lourtier	i lɪ	i ʏ eːz
15	Miège	lɪ lɛ	lʏ lɛ
16	Montana	lʏ leʒ lɛ lœ̯	lɛ
17	Nendaz	e̯ ɛ	iʒ e̯ ɛ
18	Orsières	li	li
19	St-Jean	lʏ lɛ ɛ lə əʒ	lʏ lɛ lœ
20	Savièse	i ʏ ɛ øʒ œ əʒ	ɪ ʏ e̯ ɛ̯ øʒ
21	Sixt m.	luz lʊ	lu lʊ
	f.	le løz lz	le lz
22	Torgnon	le lɛz lə	lɛ lœ lə
23	Troistorrents m.	luz lʊ lʏ l	lu u lʊz ʊz lʏ
	f.	e	lɪ ɪ le e
24	Val-d'Illiez m.	lu löz lʏ	luz lö
	f.	lɪ le ez	le lɛ̯
25	Vouvry m.	lö lʊz	luz lʊ
	f.	lɪ le	lɪ le



Carte n° 9: distinction et neutralisation des deux genres au pluriel de l'article défini[15]

(32) le - **le** dzɛ̃[16] ʁtset'avɑ̃ **le** b'ote ʊe l kaʁdaɲ'eː (TroistorrentsM)
Les .. les gens achetaient les souliers chez le cordonnier.

Il apparait aussi (cf. carte n° 8) que les cinq parlers qui possèdent une forme du masculin pluriel en [u] forment une zone cohérente, dans les parlers du Chablais valaisan, et c'est une zone qui se prolonge en Haute-Savoie voisine. De fait, les parlers du Chablais valaisan constituent la pointe orientale de la grande zone des langues ibéro- et galloromanes qui distinguent le masculin et le féminin pluriel, à la différence du français et des parlers occitans de l'espace provençal qui ont neutralisé l'opposition de genre pour les

[15] Nos remerciements vont à Chiara Marquis qui a élaboré cette carte. Les points isolés de l'espace oïlique qui semblent présenter une distinction des deux genres, d'après le témoignage des cartes dépouillées de l'ALF, demanderaient une analyse spécifique qui ne peut pas être entreprise ici.

[16] En francoprovençal, [dzˈɛ̃] 'gens' est féminin (cf. *GPSR* 8, 258-265, en particulier 265a); il en va de même pour [bˈote] 'chaussures'.

déterminants au pluriel (carte n° 9). Même si cela nous laisse encore un certain polymorphisme pour le masculin **et** le féminin pluriel dans ces parlers, cela clarifie déjà considérablement la situation.

Mais il nous reste encore le gros morceau: ce sont tous les parlers du Valais central et les parlers valdôtains voisins qui ne connaissent pas cette répartition des formes, et dont la majorité est caractérisée par un important polymorphisme[17].

Dans l'état actuel de notre réflexion, nous pensons que la clé, pour la compréhension de cette situation, pourrait venir de l'examen des formes de l'article défini singulier. Comme nous l'avons montré en 2010 au Congrès de linguistique romane de Valence (Kristol 2013), les parlers francoprovençaux de l'Est valaisan sont les seuls, dans le monde latin occidental, qui conservent jusqu'à nos jours un système bicasuel parfaitement vivant et fonctionnel au singulier, un système typiquement occidental qui distingue le cas sujet du cas régime, comme en ancien français ou en ancien occitan, mais en s'appuyant uniquement sur la forme de l'article défini. Nous nous contentons ici d'un seul exemple qui illustre ce phénomène: c'est un énoncé spontané de notre informateur de Montana :

(33) **lï** greːnˈi i ɛ pʊ tənˈiŋ lɔ - **lɔ** bla k õn a batˈup (MontanaM)
Le grenier (sujet) il est pour garder le .. le blé (objet direct) qu'on a battu.

Or, un collègue rédacteur du *Glossaire des patois de la Suisse romande* nous a rendus attentifs au fait qu'il y a au moins un parler de l'Est valaisan qui maintient en principe un système bicasuel de l'article défini au pluriel aussi: à Évolène, la sujet pluriel masculin possède une forme spécifique, alors que celle du régime se confond avec le féminin, sujet et objet.

En effet, en analysant le corpus d'Évolène, nous avons constaté que notre informatrice faisait une distinction parfaite entre le masculin pluriel et le reste du paradigme :

(34) **i** mjo̯ parˈɛns atsëtavɔ̃ pa **lɛ** tsambˈœt - n avˈĩŋ dɛː pwɛʃ mˈimɔ (ÉvolèneF)
Les miens parents (sujet) achetaient pas les jambons (objet direct). Nous avions des porcs (nous-)mêmes.

Notre témoin masculin, en revanche, même s'il connaît et utilise encore parfois une forme spécifique pour le masculin pluriel, montre une tendance très nette à la confusion des formes du sujet et du régime masculin. Si (35) maintient l'opposition casuelle [lʏ] / [lə], celle-ci est neutralisée en (36); [lø] est un allomorphe des formes du régime et du féminin:

(35) o vjɔ tˈɛ̃ŋ jo mə ʃuvˈønio kə lə - **lʏ ʒ** ˈom ɑlˈɑʊən plytˈo fi lɔ̈ lɔ dˈəvɑ̃ŋ - ɛ **lə** fəmˈəlë dˈəʃ la mˈɛjtʃa dˈəri də l iʎˈøʒ (ÉvolèneM)
Autrefois je me souviens que les .. les hommes allaient plutôt sur le devant .. et les femmes vers la moitié arrière de l'église.

(36) **lø** mjɔ̈ parˈẽs atsëtˈɑːvœn pɑ **lə** tsambˈøtᵋ - n avĩŋ dœ̯ pwˈɔʃyr nɔ mˈĩmɔ (ÉvolèneM)
Les miens parents n'achetaient pas les jambons .. nous avions des cochons nous-mêmes.

Signalons en passant que, contrairement aux croyances répandues dans la recherche dialectologique et sociolinguistique, ce ne sont pas toujours les «NORM» (les Non-educated, Old, Rural, Male) qui sont les meilleurs informateurs, les meilleurs conserva-

[17] Étant donné que le Valais central fait partie des régions les plus conservatrices de l'espace francoprovençal et que l'emprise du français en Vallée d'Aoste est certainement plus faible que dans les régions transalpines, nous sommes tentés de penser que la neutralisation du genre, dans ces parlers, n'est pas due à une éventuelle influence du français, mais prolonge la zone lombarde qui connaît le même phénomène.

teurs du dialecte traditionnel: dans notre cas précis, la locutrice – plus jeune – maintient mieux le système bicasuel que l'homme plus âgé.

Mais l'essentiel n'est pas là. Ce qui est significatif, ici, c'est le fait que dans les parlers immédiatement voisins d'Évolène, qui possèdent en principe le même éventail de formes, nous avons pu constater que l'opposition bicasuelle au pluriel n'est plus fonctionnelle, même si elle se maintient parfaitement au singulier. Ainsi, notre informateur d'Hérémence est capable d'utiliser un ancien [lʏ] du cas sujet pour un complément d'objet, ce que le locuteur d'Évolène ne fait jamais dans les exemples disponibles dans notre corpus :

(37) ɪ kɔ̃tˈɑː na dɔzˈɑ̃ŋnɑ ə də ʒ ʁʁɑ̃ndˈɑlːə la ʃ **lʏ** fek (HérémenceM)
J'ai compté une douzaine euh d'hirondelles là sur les fils.

À notre avis, les conclusions qui s'imposent de cette observation sont évidentes – et le phénomène décrit ici n'est pas un cas isolé dans nos matériaux. Le polymorphisme actuel que nous avons relevé dans la plupart des parlers du Valais central résulte sans doute de l'abandon de l'ancienne opposition casuelle, alors que l'éventail des formes disponibles est resté stable. Étant donné l'absence complète d'une norme scolaire ou académique dans nos parlers, il n'y a pas eu la moindre *décantation*, il n'y a pas eu la moindre *sélection* parmi les formes disponibles. Par conséquent, nous observons actuellement de vrais phénomènes de variation libre, une variation dans laquelle les différents allomorphes n'ont aucune assignation spécifique de nature sociolinguistique (âge, sexe ou groupe social). Mais sur une telle base, rien n'empêche évidemment que dans un deuxième temps, dans le cadre de l'évolution linguistique, une telle variation interne puisse être réaffectée et réutilisée de manière fonctionnelle – soit en distinguant des groupes sociaux au sein d'une même communauté, soit encore, comme nous l'observons parfois dans nos propres matériaux, pour mieux se distinguer des locuteurs des villages voisins qui partagent en principe les mêmes structures linguistiques, mais qui opèrent des choix divergents dans les virtualités offertes par le système.

7 Quelques conclusions

Une première conclusion qui s'impose à nous, dans notre travail, c'est que l'utilité des atlas linguistiques audiovisuels n'est plus à prouver. Le recours constant au document audiovisuel enregistré, image et son, améliore considérablement nos conditions de travail. De plus, la convivialité des atlas linguistiques de nouvelle génération – et la fiabilité des matériaux linguistiques qu'ils mettent à notre disposition, grâce à la bande son originale – est nettement meilleure que les atlas «papier» – ce qui ne diminue pas les mérites de nos prédécesseurs qui ne pouvaient faire mieux, et qui n'avaient pas toutes les ressources que l'informatique a mises à notre disposition.

Ce qui n'est plus à démontrer non plus, à notre avis – mais là, nous n'avons rien inventé – c'est l'intérêt des cartes interprétées par rapport aux cartes qui se contentaient d'inscrire les données brutes dans l'espace géographique, dans la mesure où les données originales restent intégralement conservées.

Mais le point le plus important, c'est la question que nous avons soulevée dans le titre de cette communication. Comme nous croyons l'avoir montré, le polymorphisme de nos données n'est pas seulement un défi pour l'analyse linguistique et pour la représentation cartographique des données. Il devient lui-même porteur d'information. Dans la mesure

où nous nous donnons les moyens de l'enregistrer – en posant les questions qu'il faut, lors des enquêtes – et dans la mesure où nous ne l'écartons pas comme une sorte de nuisance ou de phénomène parasite, il peut nous permettre de mieux comprendre les systèmes linguistiques que nous analysons, et il peut nous renseigner sur certains mécanismes de l'évolution linguistique.

Dans ce contexte, l'intérêt heuristique de l'approche géolinguistique devient évident. Il ne suffit pas de rendre compte de la diversité du langage humain dans sa variation géolinguistique – et de la diversité des microsystèmes morphosyntaxiques qui coexistent en synchronie dans un même espace linguistique. Le véritable intérêt de l'approche géolinguistique, c'est qu'elle peut nous permettre de *comprendre* les faits apparemment obscurs que nous observons dans un dialecte donné grâce aux évolutions qui se déroulent sous nos yeux et dans la même synchronie dans un dialecte voisin.

Bibliographie

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

ALD = Goebl, Hans (1998-2012): *Atlant linguistich dl ladin dolomitich y di dialec vejins = Atlante linguistico del ladino dolomitico e dei dialetti limitrofi = Sprachatlas des Dolomitenladinischen und angrenzender Dialekte* / Helga Böhmer... [et al.] materialia collegerunt... ; Hans Goebl opus omne curavit, Wiesbaden: L. Reichert/Strasbourg: Société de linguistique romane

ALF = Gilliéron, Jules / Edmont, Edmond (1902-1910) : *Atlas linguistique de la France*, Paris

Bauer, Roland (1995): «Vivaldi-Sicilia. Documentazione sonora dei dialetti siciliani», in: Giovanni Ruffino (a cura di), *Percorsi di geografia linguistica. Idee per un atlante siciliano della cultura dialettale e dell'italiano regionale*, Palermo: 543-550

Gauchat, Louis/Jeanjaquet, Jules/Tappolet, Ernest (1925) : *Tableaux phonétiques des patois suisses romands*, Neuchâtel: Attinger

GPSR = Gauchat, Louis et al. (1924-) : *Glossaire des patois de la Suisse romande*, Neuchâtel: Attinger / Genève: Droz

Kattenbusch, Dieter (1995) : «Atlas parlant de l'Italie par régions: VIVALDI», in: *Estudis de lingüística i filologia oferts a Antoni M. Badia i Margarit*, Barcelona 1995: 443-455

Kristol, Andres (1998) : «La production interactive d'un corpus semi-spontané: l'expérience ALAVAL», in: M. Mahmoudian/L. Mondada (ed.), *Le travail du chercheur sur le terrain*. Questionner les pratiques, les méthodes, les techniques de l'enquête, Cahiers de l'ILSL 10, Lausanne: Université de Lausanne, 91-104 ; http://www.unine.ch/ dialectologie/ALAVAL_articles/Kristol_1998.pdf

Kristol, Andres (2009) : «La morphosyntaxe du pronom personnel sujet de la première personne du singulier en francoprovençal valaisan: comment manier le polymorphisme d'une langue dialectale?», in: Frechet, Claudine (éd.), *Langues et cultures de France et d'ailleurs*. Hommage à Jean-Baptiste Martin, Lyon: Presses universitaires de Lyon: 195-216

Kristol, Andres (2013) : «Le francoprovençal, laboratoire des virtualités linguistiques de la Romania occidentale: le système bicasuel des parlers valaisans.» Conférence plénière, in: Casanova, Emili/Calvo, Cesareo (ed.) : *Actes del 26é Congrés internacional de lingüística i filologia romàniques*, València, 6-11 septembre 2010, vol. 1, Berlin: 341-361.

Martin Jean-Baptiste (1974) : «Le pronom personnel sujet de la première personne du singulier en francoprovençal», *RLiR* 38: 331-338

Marzys, Zygmunt (1964) : *Les pronoms dans les patois du Valais central. Étude syntaxique*, Berne: Francke

Tuaillon, Gaston (1972) : «Le francoprovençal. Progrès d'une définition», *Travaux de Linguistique et de littérature* 10/1: 293-339

20 Jahre digitaler Sprachatlas VIVaio Acustico delle Lingue e dei Dialetti d'Italia (VIVALDI)

Fabio Tosques & Michele Castellarin, Humboldt-Universität zu Berlin

Zu den zahlreichen schützenswerten kulturellen Gütern in Italien gehören zweifelslos auch die vielen Dialekte und Minderheitensprachen. Während sich Literatur in Texten, Architektur in Bauwerken und Musik in der klanglichen Darbietung ausdrückt, leben Sprachen besonders von der akustischen Realisierung ihrer Sprecher. Bedauerlicherweise liegen im Bereich der romanischen Sprachgeographie allgemein zugängliche Daten bisher fast ausschließlich in gedruckter Form vor. Wenn man einmal absieht von einer Reihe von Schallplatten (v.a. zu italienischen Dialekten), die jedoch erhebliche Nachteile aufweisen (teilweise schlechte Aufnahmequalität, Abnutzung, besonders aber geringe Kapazität), gibt es bisher für Italien kein systematisch erhobenes authentisches auditives Sprachmaterial, das für Forschungszwecke zur Verfügung stünde. Eine Ausnahme bildet der 1998 erschienene Atlas des Dolomitenladinischen (ALD), zu dessen Lie¬fer¬um¬fang auch 3 CD-ROM bzw. 1 DVD gehören (vgl. Bauer/Goebl/Haimerl 2005 und http://ald1.sbg.ac.at/). Die Dokumentation der sprachlichen Einzigartigkeit Italiens hat sich das Projekt VIVALDI (VIVaio Acustico delle Lingue e dei Dialetti d'Italia / Akustischer Sprachatlas der Dialekte und Minderheitensprachen Italiens) zur Aufgabe gestellt. Das Hauptziel besteht darin, authentisches und aktuelles phonetisches Material zu sammeln und der Wissenschaft und Forschung zur Verfügung zu stellen, das besonders die kontemporären Einflüsse des Standarditalienischen auf die untersuchten Dialekte und Minderheitensprachen sichtbar und hörbar machen soll, den Vergleich der vielen verschieden Dialekte untereinander erlauben soll und auch den Vergleich auf phonetischer Ebene mit älteren Sprachstufen zulässt (z.B. AIS, Atlas Italiens und der Südschweiz).

1 VIVALDI – Von der Idee bis zur Realisieurng der digitalen Versionen

Sprachatlanten zählen im Bereich der Sprachgeographie und Geolinguistik zu den wichtigsten Arbeitsmitteln in der Forschung. Als Forschungsinstrument hat auch in Italien der Sprachatlas eine lange Tradition, wobei die bisherigen ausschließlich in gedruckter Form erschienen sind. Im Umfeld des ALD I (*Atlante linguistico del ladino dolomitico e dei dialetti limitrofi*)[1] entstand die Idee eines akustischen Sprachatlasses, der alle Regionen Italiens abdeckt.

Die ersten Sprachatlanten, die über die romanische Sprachgeographie hinaus methodologische Maßstäbe setzten, sind der ALF (*Atlas linguistique de la France*) aus den Jahren 1902 bis 1910 (Gilliéron & Edmont 1902-1910) und der AIS (*Sprach- und Sachatlas Italiens und der Südschweiz*) aus den Jahren 1928 bis 1940 (Jaberg/Jud 1928-40), die in Form von gedruckten, zu mehreren großformatigen Bänden gebundenen Sprachkarten vorliegen, jedoch kein authentisches Tonmaterial enthalten.

Neuere Initiativen wie beispielsweise das in den 80er Jahren von Sobrero geleitete Projekt NADIR (**N**uovo **A**tlante del **D**ialetto e dell'**I**taliano per **R**egioni, vgl. Sobrero 1991) oder das von Ruffino initiierte Projekt des sizilianischen Sprachatlasses (ALS, vgl. Ruffino 1986) stellen bis heute der Öffentlichkeit vom gesammelten erhobenen Material nichts bzw. lediglich gedruckte Kostproben zur Verfügung.

Mit der Schallplatte steht zwar seit den zwanziger Jahren des 20. Jahrhunderts ein Medium zur Verfügung, das eine Konservierung und größere Verbreitung sprachlichen Materials ermöglicht, aber die Aufnahmepraxis war umständlich und kostspielig, so dass

[1] http://www.sbg.ac.at/rom/people/proj/ald/ald_home.htm

sie kaum in der Feldforschung eingesetzt wurde. Es gibt eine Reihe von Schallplatten – z.B. zu italienischen (Cortelazzo 1974-88) oder zu schweizerdeutschen (Hotzenköcherle/Brunner 1972-76) Dialekten –, die jedoch erhebliche Nachteile aufweisen: schlechte Aufnahmequalität, Abnutzung, besonders aber geringe Kapazität sowie keine direkte Verbindung des visuellen Elements (Karte) mit dem auditiven (Ton).

Mit der Einführung der CD und später der DVD wurde zwar das Problem der Kapazität geringer und die Einführung von Aufnahmegeräten wie MiniDisc und DAT-Recorder erleichterte die Aufnahmepraxis. Dennoch sind die inzwischen digital aufgezeichneten und digitalisierten Tondokumente, die seit den 1990er Jahren auf CD erschienen sind, in der Praxis nur schwer erhältlich.

Heute findet man natürlich zahlreiche dialektale Kostproben in Portalen wie Youtube usw., die jedoch völlig unsystematisch dargeboten werden und daher für wissenschaftliche Zwecke weitgehend unbrauchbar sind. Neuere Projekte wie das von Google unterstützte Projekt *endangered languages*[2], welches die weltweit bedrohten Minderheitensprachen katalogisieren möchte, sind ebenfalls unsystematische Sammlungen von Video- und Audioaufzeichnungen der jeweiligen Sprachen und für die wissenschaftliche Arbeit schwer zu verwenden.

Die Einführung moderner, computerunterstützter Verfahren der Tonaufnahme, -konservierung, -verarbeitung und -wiedergabe erlaubte eine qualitativ hochwertige, verlustfreie und kostengünstige Verbreitung der Sprachdokumente. Der Sprachatlas ALD I erschließt der Sprachgeographie diese Neuerungen erstmals in größerem Rahmen (Goebl 1998): Die abgefragten Wörter und Sätze (Stimuli) werden im Computer in einer Datenbank systematisch nach genau definierten Kriterien erfasst, welche die Transkription der einzelnen Dialektantworten unterstützt, die Verwaltung der Sonderzeichensätze übernimmt, über ein Koordinatensystem eine Zuordnung zur geographischen Lokalisation herstellt und die Daten über das Postscript-Format (heute PDF-Format) direkt in die Karte exportiert. Darüber hinaus werden die Tondokumente auf DVD mit direktem Zugriff auf die einzelnen Stimuli bereitgestellt. Damit wurden die technischen Möglichkeiten umfassend genutzt und mit Hilfe des Computers eine direkte Verknüpfung von Ton, Karte und Transkription realisiert.

Den Einsatz gängiger Software zur Erstellung von Sprachkarten kleinräumiger Sprachatlanten sowie die Bearbeitung von Tondokumenten beschreiben Harder und Boller (Harder & Boller 1996), wobei von ihnen die Verbindung von Transkription und Ton noch nicht realisiert, aber als in der Zukunft realisierbar gesehen wird (vgl. Müller & Köhler & Kattenbusch 2001).

Seit Ende der 1990er Jahre werden mit der immer schnelleren Verbreitung des Internets gehäuft digitale Sprachatlanten als Online-Version realisiert. Als eines der ersten funktionierenden online-Projekte ist hier in jedem Fall VIVALDI zu nennen, welches seit 1999 die erhobenen Daten kontinuierlich im Netz zur Verfügung stellt. Die Internetversion des ALD-I – mit dem das Projekt VIVALDI im ständigen Wissens- und Informationsaustausch steht – basiert beispielsweise komplett auf der im Projekt VIVALDI entwickelten Softwareumgebung. Zahlreiche andere Projekte, die das Ziel verfolgen, Sprachkarten mit authentischen Audiodaten zu verknüpfen, haben seit kurzem mit der Entwicklung von eigenen Lösungen begonnen und zeigen auch verschiedene Beispielkarten. Zu nen-

[2] http://www.endangeredlanguages.com

nen wäre hier beispielsweise das Projekt ALMURA (*L'Atlas linguistique multimédia de la région Rhône-Alpes*)[3], welches sich an VIVALDI (und in der Konsequenz am ALD-I) orientiert:

> En général, ces atlas [ALMURA, Anm. d. Verf.] disposent d'un module phonétique qui sur interrogation de l'internaute reproduit la prononciation d'un item donné. Cela nous paraît particulièrement réussi pour les projets ALD-I et Vivaldi qui complètent la reproduction de l'enregistrement du témoin par la transcription phonétique de l'item (Reisdoerfer 2009).

Neueste Publikationen, wie die beiden in der Reihe *Handbuch der Sprach- und Kommunikationswissenschaft* erschienenen Bände *Language and Space* verweisen auf die Bedeutung und die nicht von der Hand zu weisenden Vorteile von multimedialen Sprachatlanten im Internet:

> The advantages of internet publication are obvious. Multimedia and interactivity offer map users new ways of perceiving areal variation. Internet publications are readily accessible to the public. They allow for the integration of different research projects and access to a large quantity of data. In contrast to finalized print publications, internet publications have a dynamic character since, in principle, they remain open for corrections and additions (Girnth 2010, S. 117).

Seit 1999 existieren von VIVALDI eine online-Version und eine CD-ROM bzw. seit 2007 eine DVD, die stets weiterentwickelt werden. Stand 1999 nur die Region Sizilien zur Verfügung, ist seither kontinuierlich fast jedes Jahr eine weitere Region veröffentlicht worden. Aktuell sind elf Regionen abgeschlossen und weitere fünf in Bearbeitung, in 230 Orten wurden Interviews mit Sprechern durchgeführt (Stand: 08/2013). VIVALDI bietet inzwischen gut 100.000 Datensätze online (Audio-Dateien und Transkriptionen).

Im Wesentlichen hat sich VIVALDI folgende Ziele gesetzt:

1. Sammlung aktueller Dialektdaten in allen 20 Regionen Italiens.
2. Nutzung elektronischer Datenträger (CD-ROM/DVD) und moderner Medien (Internet) zur schnellen Bereitstellung der Daten in Form von Audio-Dateien.
3. Transkriptionsvorschläge, somit Möglichkeit des Vergleichs der aktuellen Daten mit den Daten des AIS (*Sprach- und Sachatlas Italiens und der Südschweiz*, vgl. Jaberg/Jud 1928-40) und des ALI (*Atlante Linguistico Italiano*, vgl. Bartoli 1995-2008).
4. Flächendeckende Dokumentation der italienischen Dialektlandschaft.
5. Wissenschaftliche Dokumentation der Basisdialekte.
6. Verwendungsmöglichkeit in der linguistischen Forschung.

Mit der L.N. (nationales Gesetz) 482 von 1999 zum Schutz der historischen Minderheitensprachen in Italien war Italien eines der ersten Länder, die die Minderheitensprachen tatsächlich gesetzlich schützen. Das liegt zum einen daran, dass in Italien nach wie vor eine Vielzahl an Minderheitensprachen aktiv gesprochen werden und dass zum anderen erkannt wurde, dass jene, die weniger als 10.000 Sprecher haben, in naher Zukunft vom Aussterben bedroht sein könnten. In Artikel 2 der L.N. 482/1999 werden die zu

[3] http://w3.u-grenoble3.fr/almura/index.php.

schützenden Minderheitensprachen explizit genannt: „[...] la Repubblica tutela la lingua e la cultura delle popolazioni albanesi, catalane, germaniche, greche, slovene e croate e di quelle parlanti il francese, il franco-provenzale, il friulano, il ladino, l'occitano e il sardo." Von dieser Gesetzesinitiative ausgenommen sind aber sämtliche Dialekte, die in Italien aktuell gesprochen werden (vgl. Toso 2006, 67). Jede noch so gut gemeinte Gesetzesinitiative kann aber alleine das Aussterben einer Sprache nicht verhindern. Davon betroffen sind besonders die Minderheitensprachen mit wenigen 100 oder 1.000 Sprechern wie beispielsweise das Okzitanische in Guardia Piemontese (ca. 340 Sprecher, vgl. Toso 2006, 89) oder die kroatischen Kommunen in Molise mit weniger als 2.000 Sprechern (vgl. ebd. 127).

Komplizierter stellt sich die Situation bei den Dialekten dar, die – abgesehen von vereinzelten regionalen Initiativen – keinerlei staatlichen Schutz genießen. Das betrifft zum einen dialektale Sprachinseln wie das Galloitalische in der Basilikata (vgl. Toso 2006, 87) oder auf Sizilien (vgl. ebd. 159), aber auch viele Kommunen auf der Halbinsel, die von starker Abwanderung betroffen sind. Eine Vielzahl italienischer Dörfer besteht nur noch aus wenigen hundert Einwohnern und deren Bevölkerungsprognose ist mehr als pessimistisch. So werden wir in den nächsten Jahrzehnten vermutlich erleben, dass mit dem Aussterben der Dörfer auch die jeweiligen Dialekte dem Untergang geweiht sind.

Besser ist die Situation für das Tabarchino auf Sardinien, einen ligurischen Dialekt, der dort noch von ca. 10.000 Sprechern in den beiden Kommunen Calasetta und Carloforte gesprochen wird. Problematisch ist hier, dass die kleine Minderheit zwar durch regionale Gesetze[4] geschützt wird, nicht hingegen von dem wichtigen nationalen Gesetz 482. Das könnte u. U. zum Aussterben der Sprache führen, auch wenn aktuell davon nicht die Rede sein kann, da die Sprecher ein starkes traditionelles Sprachbewusstsein pflegen. Das zeigt, dass die dialektale Situation zwar nicht hoffnungslos ist, aber eine Vielzahl an Dialekten möglichst schnell systematisch dokumentiert werden sollten. VIVALDI ist hier bereits sehr gut vorangekommen und hat die technischen und wissenschaftlichen Möglichkeiten und Voraussetzungen, dies zu realisieren, und zwar in einem sehr kurzen Zeitplan.

Seit 1992 wurde in 230 Orten der jeweilige Dialekt bzw. die Minderheitssprache dokumentiert und publiziert. Daraus resultiert ein Korpus von ca. 100.000 Tondateien und Transkriptionen mit einer Gesamtlänge von ca. 100 Stunden, das Dank der für VIVALDI entwickelten Software dem Nutzer bequem und übersichtlich in Form von Karten oder Listen zur Verfügung gestellt wird. Es bildet so die Basis für zahlreiche Publikationen, Abschlussarbeiten und Forschungsvorhaben, deren Zahl in den letzten Jahren stetig steigt. Mit dem Abschluss der Erhebungen würde erstmalig das gesamte Dialektkontinuum Italiens mit authentischem Tonmaterial abgebildet werden und könnte so die Grundlage für weitere zukünftige Forschungsarbeiten bilden.

Dass das Projekt zusehends an Bekanntheit und Bedeutung gewinnt, wird zum einen durch die Nutzerzahlen der online-Version deutlich (ca. 1.000 Zugriffe/Monat – Google Analytics, vgl. Abbildung **1**) und zum anderen wird in der einschlägigen Fachliteratur VIVALDI häufig zitiert und als funktionierendes Beispiel für einen sprechenden interaktiven Sprachatlas angeführt. Es handelt sich schließlich um eine systematische Sammlung einer nicht-zufälligen Stichprobe von italienischen Dialekten mit einem definierten

[4] L.R. 26/1997.

Fragebuch und ist zudem auf den HU-Seiten[5] weltweit hörbar. Das ist ein absolutes Novum, da ansonsten i.d.R. „nur“ die gedruckten Transkriptionen veröffentlicht werden und wurden.

Während Forscher das gesammelte Material als Grundlage für weiterführende Untersuchungen und Analysen der Dialekte und Minderheitensprachen Italiens nutzen können, bietet VIVALDI Laien, Schülern und Studenten, Einheimischen und Ausgewanderten, Angehörigen der Minderheitensprachen und Einwohnern aus den Ursprungsländern die Möglichkeit, sich die Originalstimmen anzuhören und so eine sprachliche Zeitreise zu unternehmen, die dank neuester Technologien heute tatsächlich im Bereich des Möglichen liegt. Wir freuen uns, dass wir Teil dieser dokumentarischen und wissenschaftlichen Erhebungen sein dürfen, deren Wert von Jahr zu Jahr steigt. Und vielleicht gelingt es uns, einen Teil des immensen Kulturschatzes Italiens zu bewahren, zu dokumentieren, verfügbar zu machen und zu analysieren, da einige der von uns aktuell erhobenen Sprachschätze vom Sprachentod bedroht sind. Wohl wissend, dass das Aussterben von Sprachen nicht verhindert werden kann, möchten wir doch unseren Teil dazu beitragen, dass die Stimmen der Vergangenheit (und Gegenwart) nicht in Vergessenheit geraten und so die Grundlage für aktuelle und zukünftige Forschungsvorhaben bilden.

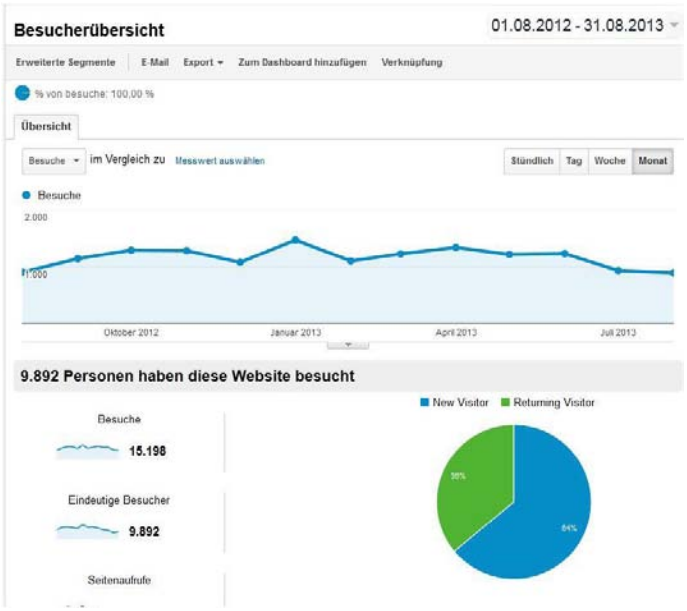


Abbildung 1: Screenshot von Google Analytics im Zeitraum vom 01.08.2012 bis 31.08.2013: Zu sehen ist, dass monatlich mindestens 1000 Besucher die VIVALDI-Seiten aufrufen.

Bisher wurden die folgenden Regionen aufgenommen, bearbeitet und veröffentlicht:

- 1992: Sizilien (erste Probeaufnahmen, die teilweise seit 1999 wiederholt wurden)
- 1999-2000: Sardinien
- 2001-2002: Ligurien
- 2003: Aostatal
- 2003-2004: Umbrien

[5] http://www2.hu-berlin.de/vivaldi

- 2005-2007: Trentino-Südtirol
- 2005: Molise
- 2007-2008: Piemont
- 2009-2010: Friaul-Julisch Venetien
- 2009-2012: Venetien
- 2012: Basilikata

Seit 2009 sind Toskana, Apulien, Kalabrien, Lombardei und Kampanien in Bearbeitung. Die in diesen Regionen aufgenommenen Dialekte sind bereits online verfügbar. Es fehlen noch: Emilia-Romagna, Marken, Latium und Abruzzen.

Seit dem Jahr 2011 können die Minderheitensprachen mit dem gesonderten Visualisierungsmodul PALMI (Panorama Acustico delle lingue minoritarie d'Italia) in ihrer Gesamtheit dargestellt werden (s. Abbildung 2). Zum ersten Mal stehen Tondokumente von den aktuell in Italien gesprochenen Minderheitensprachen der Allgemeinheit zur Verfügung.

Es ist uns natürlich bewusst, dass mit dem Fragebuch, das speziell die Entwicklungen in den italienischen Dialekten untersuchen soll, insbesondere die nicht-romanischen Sprachen nicht vertieft analysiert werden können. Andererseits können mit den gut 350 Antworten in jedem Fall unidirektionale Interferenzen vom Italienischen bzw. der lokalen Dialekte auf die Minderheitensprachen herausgearbeitet werden.

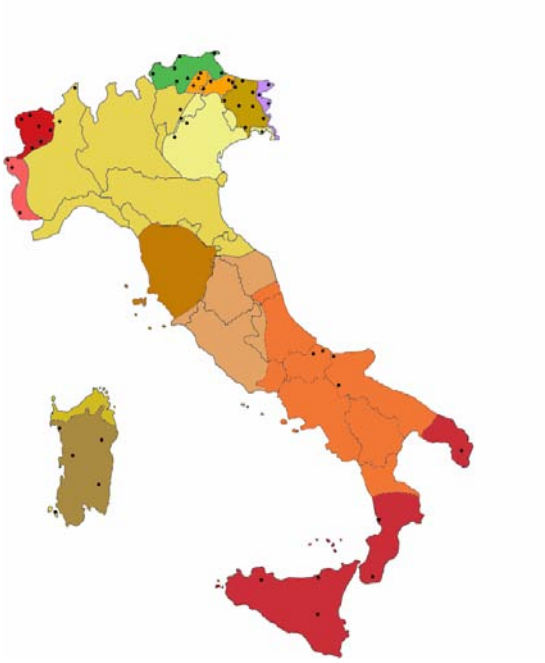


Abbildung 2: Startseite von PALMI, eine Übersicht von den verzeichneten Minderheitensprachen

Die Wahl der bisherigen Aufnahmeorte berücksichtigt hauptsächlich die grobe dialektale Gliederung Italiens (Nord-, Mittel-, Süditalien). Die Auswahl der bisher aufgenommenen Regionen erfolgte auch immer mit dem Ziel, eine romanische oder nicht-romanische Minderheitensprache zu dokumentieren.

Nach der Erhebung der Daten im Feld und der digitalen Aufarbeitung und Nachbereitung dauert es in der Regel höchstens sechs Monate bis zur Veröffentlichung im Internet. Neben der zeitaufwendigen Erhebung der Daten im Feld, der Konzeption und Erstellung des Fragebuchs sowie der Veröffentlichung der Daten im Netz und auf DVD wurden die folgenden Vorarbeiten geleistet: Entwurf der Datenbank, Konzeption der Software zur Präsentation der Daten und Erstellung der Webseiten.

Die relationale Datenbank von VIVALDI bildet das Herzstück für die Präsentation und Publikation des im Feld gewonnenen Materials. Nachdem die Daten eingegeben wurden, werden diese mit Hilfe verschiedener Softwaremodule ausgelesen. Die Eingabe betrifft einerseits die Transkription und andererseits die geschnittenen Tondateien aus dem Aufnahmekontinuum. Dafür wird im Projekt die Software Goldwave benutzt, mit deren Hilfe die einzelnen Antworten aus der Aufnahme, die in der Regel zwei bis drei Stunden dauert, geschnitten werden. Goldwave hilft auch beim Segmentieren der einzelnen Antworten, um die Transkriptionen zu verbessern und/oder zu kontrollieren.

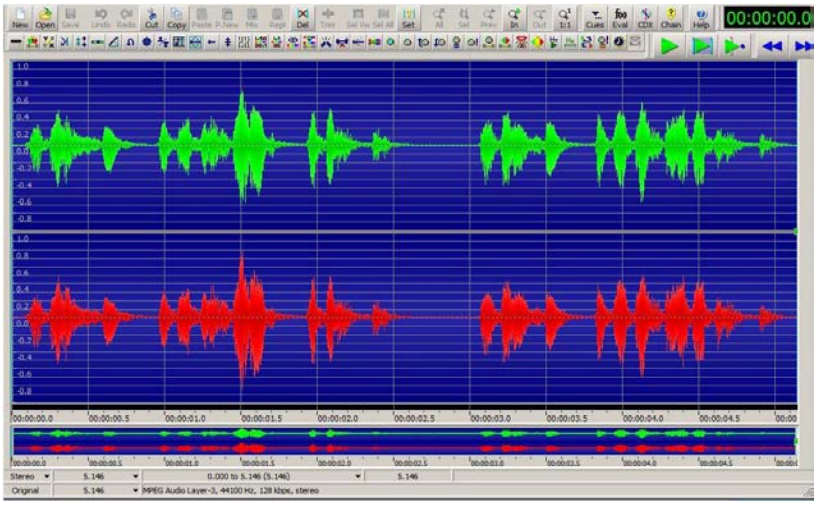


Abbildung 3: Screenshot Goldwave vom Stimulus Nr. 83 „Adesso devi andare a destra“ von Ortsnr. V08 (Verona)

Bei schwer zu identifizierenden Lauten wird zur Unterstützung auf das Programm Praat zurückgegriffen, da hier alle drei Dimensionen (Dauer, Frequenz und Intensität) sichtbar werden und dem Phonetiker beim Erkennen der Laute behilflich sind. So erleichtern die bei Praat dargestellten Formanten die Identifizierung von Vokalen und Diphthongen.

Die Transkriptionen werden zunächst in eine Excel-Tabelle, dem im Projekt entwickelten VIVALDI-Transkriptionsassistenten (VivTKA) eingetragen, der die Codes in die jeweiligen Transkriptionszeichen umsetzt. Dafür mussten spezielle Module für Excel entwickelt werden, die die Umsetzung vornehmen.

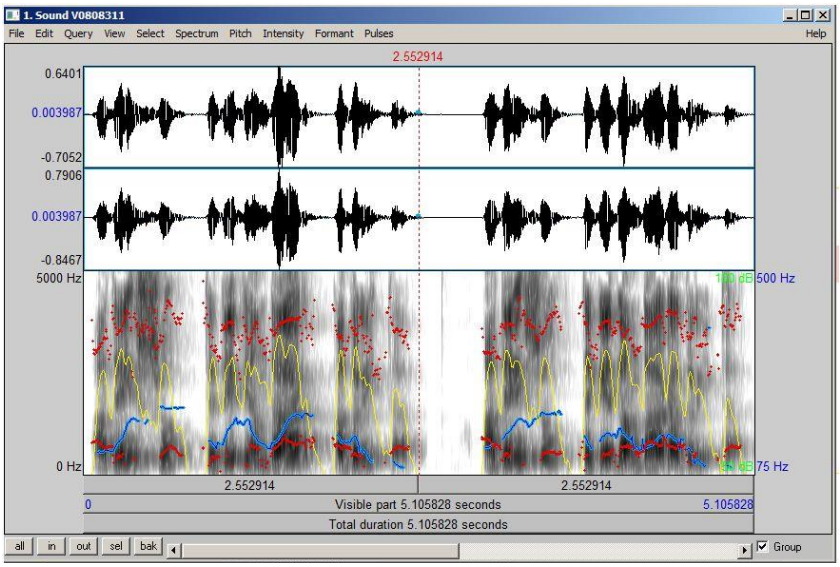


Abbildung 4: Screenshot Praat vom Stimulus Nr. 83 „Adesso devi andare a destra“ von Ortsnr. V08 (Verona)

Die Aufbereitung des Tonmaterials beruht darauf, dem linguistisch forschenden Benutzer keine subjektive Transkription anzubieten, damit er nicht auf die eines anderen Wissenschaftlers angewiesen ist. So entfällt z.B. auch die fehleranfällige Übertragung von einem in ein anderes Transkriptionssystem, mit dem der romanistische Linguist ständig konfrontiert ist. Benutzer können das Tonmaterial z.B. auch für eigene phonetisch/akustische Analysen verwenden.

A	B	X	Y	Z	AA	AB	AC
	Ort Nr.	**V07**			**V08**		
	SprecherIn Nr.	**1**			**1**		
	Version	**1**			**1**		
	Ort Name	**Pozzale**			**Verona**		
	Bemerkung	fertig			fertig		
	Sprachatlas Ort						
	WWW Ort						
StNr.	**Stimulus**	Code	Zeichen	Notiz	Code	Zeichen	Notiz
1	l'acqua	al \3D\1ga	*al ɛ́ga*		l ákwa	*l ákwa*	
2	l'acqua è calda	al \3D\1ga l(r) é	*al ɛ́ga ˌlʳ é cáu̯ða*		l ákwa l é kálda	*l ákwa l é kálda*	
3	l'agnello	al k\3e\1tín	*al kẹ́tín*		l a\6n\3D\1l	*l a ñɛ́l*	
4	l'aglio	l ái	*l ái*		l áy\2O\1	*l áyọ*	
5	agosto	ag\2d\6s\1t\2o\	*agóṣtọ*		agó\6s\1t\2O\1	*agóṣtọ*	
6	l'ala	l ála	*l ála*		l ál\6a\1	*l álɑ*	
7	alto	\f\1á\6w\1t\2o\1	*áu̯tọ*		ált\2O\1	*áltọ*	
8	altro	á\6w\1tro	*áu̯tro*		áltr\2O\1	*áltrọ*	
9	dammi un altro pezzo	dá m in un á\6v	*dá m in un áu̯tro tṣọ*		dá m\3E\1 n ált	*dá mẹ n áltrₒᵃ tṣ́kọ*	
10	l'anca	l \7n\1n\6c\1a	*l ã́nća*		l á\6N\1k\6a\1	*l áŋkɑ*	
11	l'angelo	l \7n\1nd\2O\1l	*l ã́ndọl*		l án\6g\1el\5U\	*l ángelu̯*	
12	l'anno	l á\6N\1	*l áŋ*		l án\2O\1	*l ánọ*	
13	aprile	aprí\3e\1	*aprílẹ*		apr\4n\1l	*aprí͡l*	
14	l'argento	l ar\6d\3D\1nto	*l arδɛ́nto*		l arg\3G\1nt\2C	*l argɛ́̃ntọ*	

Abbildung 5: Eingaben im VIVALDI Tanskriptions Assistenten (VIVTka)

Nach der Endkontrolle der eingetragenen Transkriptionen mit den Tondateien wird aus den Excel-Daten eine Exportdatei für die Datenbank erzeugt. Die Daten werden dann in die Datenbank eingespielt, worauf die im Projekt entwickelte Software den Ort mit der Tondatei und der Transkription verknüpft. Entweder kann die Antwort auf die gestellte

Frage durch einen Klick auf den jeweiligen Ort in der Karte angehört werden, wobei die Transkription dann über der Karte erscheint, oder die Antworten stehen mit Ortsname, Transkription und Tondatei als Liste unter der Karte zur Verfügung.

Abbildung 6: Die Region Piemont im Projekt VIVALDI

Für die Präsentation wurden mit Hilfe von LAMP (Linux, Apache, MySQL, PHP) die Internetseiten entwickelt. Die Karte selbst wurde als Java-Applet realisiert. Dabei wurde stets darauf geachtet, dass standardisierte und freie Software zum Einsatz kommen. Die Seiten werden regelmäßig gepflegt, zum Einen werden sie mit neuen Daten aus neuen Regionen bzw. Orten aktualisiert und zum anderen werden sie durch Korrekturen der Tranksriptionsvorschläge und das Ersetzen von toten Links durch funktionierende auf dem neuesten Stand gehalten. Diese stetigen Verbesserungen und Aktualisierungen sind auch ein wesentlicher Vorteil von online-Sprachatlanten im Vergleich zu gedruckten Atlanten.

Seit gut sieben Jahren steht dem Nutzer auch die „Einzelauswahl von Orten und Stimuli“ zur Verfügung, die besonders für wissenschaftliche Arbeiten von größtem Nutzen ist. Der Nutzer kann sich bei der Einzelauswahl eine Treffermenge bestimmter Orte und Stimuli selbst zusammenstellen. Dies erleichtert besonders die Vergleichsmöglichkeiten von Antworten in verschiedenen Orten, wie beispielsweise das Frankoprovenzalische in Piemont und jenes in Apulien, oder den Vergleich der galloitalischen Dialekte untereinander usw.

VIVALDI Maps ist seit gut fünf Jahren online. Hier werden die Daten aus der Datenbank von VIVALDI mit den Geodaten von Google gemesht. Dadurch entsteht eine andere Form der Präsentation: Der Nutzer hat die Möglichkeit sich einen Gesamtüberblick über die aufgenommenen Orte im gesamten Raum zu verschaffen. Auch bei VIVALDI Maps werden die Transkriptionen angezeigt und die Tondateien abgespielt.

Abbildung 7: Startbildschirm von VIVALDI Maps

VIVALDI nutzt die von Girnth (2011, S. 117) genannten Vorteile von Sprachatlanten im Internet in vollem Umfang. Alle aufgenommen Daten werden schnell bearbeitet, transkribiert und zeitnah der Öffentlichkeit zur interaktiven Nutzung zur Verfügung gestellt.

PALMI ermöglicht eine schnelle und unmittelbare Visualisierung aller Sprachinseln und sprachlichen Minderheiten in Italien, die mit den Transkriptionen und Audiodateien von VIVALDI verknüpft werden und so die wissenschaftliche Beschäftigung mit diesen Minderheitensprachen vereinfacht.

Mit der Vollendung VIVALDIs wird so das gesamte Kontinuum der italienischen Dialektlandschaft hör- und sichtbar. Untersuchungen zu aktuellen phonetischen Phänomenen werden damit ermöglicht. Daher sind die Dokumentation und die systematische Archivierung von authentischem basilektalen Tonmaterial für den Gesamtraum Italien, das nicht nur für kurzfristige, sondern auch für langfristige Untersuchungen erhoben worden ist, das primäre Ziel von VIVALDI.

2 Zur Methodik

2.1 Erstellung des Fragebuchs

Das Fragebuch ist so konzipiert, dass alle phonetischen Besonderheiten im Bereich Vokalismus und Konsonantismus, die bei der Entstehung der Dialekte von Bedeutung sind, berücksichtigt sind. Das heißt: Ausgehend von den lateinischen Ursprungsformen wird für alle potentiellen Entwicklungen mindestens ein Vertreter (Etymon) abgefragt.[6] Im

[6] Eine komplette Etymaliste mit Hinweisen auf die AIS-Karten und VIVALDI-Stimuli befindet sich im An-

Bereich Vokalismus bedeutet dies beispielsweise, dass für die zehn lateinischen Vokalphoneme und Diphthonge die heutigen Ausprägungen in den jeweiligen Dialekten überprüft werden. Dabei berücksichtigt wird natürlich auch die konsonantische Umgebung, in der die Vokalphoneme im Lateinischen stehen. Gleiches gilt für die Konsonanten und Konsonantenverbindungen. Auch hier befindet sich im Fragebuch mindestens ein Vertreter für das Auftreten im Anlaut, im Inlaut und in bestimmten Umgebungen (z.B. *muta cum liquida* uvm.).

Losgelöst vom Kontext könnten einige Stimuli auf den ersten Blick als fragwürdig erscheinen, z.B. (88) „io dico“ und (89) „lui dice“. In diesem wie in ähnlich gelagerten Fällen geht es aber nicht um die konkrete lexikalische Realisierung, sondern um die phonetischen Variationen wie z.B. die Realisierung des intervokalischen stimmlosen velaren Verschlusslautes. So kann im norditalienischen Raum aus dem velaren Verschlusslaut auch bei der ersten Person Singular eine Veränderung zum Frikativ stattfinden.

Abbildung 8: "io dico" in verschiedenen Dialekten (Auschnitt: VIVALDI)

So entstehen aus den gut 300 Fragen (Originalkarten) zur Phonetik bei einer genauen Taxierung der phonetischen Phänomene mindestens 1.000 Arbeitskarten. Beispielsweise beim Stimulus Nr. 60 „la chiave“ (< CLAVE(M): Hier stellen sich die Fragen:

- Bleibt der konsonantische Nexus *cl-* im Anlaut erhalten?
- Wie entwickelt sich der Haupttonvokal *a*?
- Bleibt der intervokalische stimmhafte Frikativ erhalten?
- Wie entwickelt sich der Auslautvokal?

D.h. aus der einen Grundkarte „la chiave“ können so vier Arbeitskarten entstehen. Die folgende Tabelle zeigt exemplarisch einige Entwicklungsmöglichkeiten auf, die aus dem lateinischen Etymon CLAVE(M) entstehen können.

Ortsname	San Severo (Apulien)	Moggio (Friaul)	Claut (Friaul)	Verona (Venetien)	Cencenighe (Venetien)	Reba (Venetien)
la chiave (< CLAVE(M))	*a čę́və*	*la klɑ̃́f*	*la tʂá*	*la ćávę*	*la tʂɑ́f*	*la klę́*

2.2 Auswahl der Orte

Die Auswahl der Orte erfolgt nach festen Kriterien. Höchste Priorität haben die Aufnahmeorte des AIS-Punktenetzes. Grundsätzlich wird versucht, sämtliche im AIS verzeichneten Orte neu aufzunehmen. Sind in einer Region jedoch nur wenige AIS-Punkte dokumentiert – besonders im süditalienischen Raum ist das AIS-Punktenetz relativ dünn –

hang.

werden die ALI Punkte herangezogen, um einen Vergleich mit älteren Daten zu ermöglichen. Darüber hinaus wird immer versucht, in den einzelnen Regionen die dort vorhandenen sprachlichen Minderheiten bzw. Sprachinseln zu katalogisieren.

2.3 Sprecherwahl und Aufnahmesituation

Die Informanten müssen über eine hohe dialektale Kompetenz verfügen, sollten im Ort aufgewachsen sein und möglichst wenig Zeit außerhalb der Region verbracht haben. Das bedeutet: Es muss ein Sprecher[7] gefunden werden, dessen erste Muttersprache der Dialekt ist und der diesen in seiner alltäglichen Kommunikation verwendet.

Bei der Wahl der Sprecher ist weniger die Schulbildung oäd. von Bedeutung[8], sondern vielmehr dass sie die Pragmatik verstehen, d.h. die Sprecher eine basilektale Entsprechung des italienischen Stimulus korrekt und klar aussprechen und bereit sind, den Fragebogen vor der Befragung zu bearbeiten[9], damit alle gefragten Teile vollständig beantwortet werden.

VIVALDI geht auch der Frage nach, inwieweit sich im Gesamtraum Italien die Dialekte italianisiert haben. Besonders in den 1970er und 1980er Jahren wurden die Dialekte geradezu verteufelt und beispielsweise aus den Schulen verbannt bzw. verboten. Auch in den Familien wurde immer weniger im Dialekt kommuniziert. Daher interessiert sich VIVALDI besonders für jene Generation, die in den 1960er und 1980er Jahren aufgewachsen ist und zu dieser Zeit die Schule besucht hat. Das sind Personen, die heute im Alter zwischen 45 und 65 Jahren sind.

Die Informanten dürfen natürlich keine Sprach- und Artikulationsfehler (wie z.B. Lispeln) haben. Von größter Wichtigkeit ist ein komplettes, möglichst natürliches, Gebiss, damit beispielsweise die Sibilanten korrekt artikuliert werden. Ältere Informanten (> 70 Jahre) haben oft kein natürliches Gebiss und keine stabile und klare Stimme, um die Laute deutlich auszusprechen. Bei jüngeren Sprechern ist in der Regel das Kriterium „erste Muttersprache = Dialekt“ nicht mehr gegeben.

Meistens reicht die Befragung eines Vertreters des Ortsdialektes aus. Zeigen sich bei diesem jedoch Unsicherheiten und Diskrepanzen, so wird ein zweiter Vertreter – wenn nötig auch weitere – befragt. Die Befragung mehrerer Sprecher erfolgt aber nicht aus soziolinguistischen Motiven. Bei VIVALDI geht es weniger darum, die verschiedenen Varietäten auf kleinem Raum zu untersuchen, wie es beispielsweise Ziel soziolinguistischer Studien ist, sondern darum, den Basilekt der jeweiligen Ortspunkte im Untersuchungsnetz zu dokumentieren. Die Erforschung des Varietätengefüges kann und will VIVALDI nicht leisten. Das wäre schon wegen der Größe des Untersuchungsgebiets – ganz Italien – mit den derzeit vorhandenen Mitteln nicht erreichbar. Ziel ist und bleibt, die phonetischen Besonderheiten im gesamten Raum Italien zu dokumentieren.

[7] Mit „Sprecher“ sind natürlich sowohl die weiblichen als auch die männlichen Informanten gemeint. Aus Gründen der Lesbarkeit verzichten wir auf die umständlichen Formulierungen, wie „Sprecher/Sprecherin“ oder „SprecherIn“ usw.

[8] Darauf weisen schon Jaberg/Jud (1928, S. 190): „Bildung gefährdet den Bestand der Mundart, weil ihr Vehikel die Schriftsprache ist“. Sujets von natürlicher Intelligenz stellen unabhängig vom Bildungsstand für den Dialektforscher eine gute Quelle für die Datenerhebung dar.

[9] Der Informant bekommt den Fragebogen mindestens einen Tag vor dem Interviewtermin und hat somit Zeit, die Fragen, Sätze und das Gleichnis vorzubereiten und durchzuarbeiten.

Feste Regeln, die zum Auffinden der Gewährsleute führen, gibt es nicht. Die meisten Aufnahmepunkte sind bezüglich der Einwohnerzahl so überschaubar, dass dort von der Gemeinschaft gute Hinweise auf den optimalen Informanten gegeben werden. Solche Hinweise können in der Gemeindeverwaltung, Bibliotheken, Kirchen, Bars und (falls vorhanden) Hotels gewonnen werden. In größeren Städten gestaltet sich die Informantensuche meistens schwieriger. Hier gibt es aber Kultur-Organisationen, die einem weiterhelfen. Ansonsten ist der *Assessore alla cultura* von Amts wegen verpflichtet, den Dialektforscher zu unterstützen. Wer frühzeitig weiß, wo und wann er eine Aufnahme macht, kann auch per e-Mail zuvor sein Kommen bei der Gemeinde ankündigen und u.U. den Fragebogen vorab schicken.

Viel schwerer als das Auffinden eines geeigneten Informanten gestaltet sich häufig die Suche nach einem geeigneten Aufnahmeort. Da es sich, wie erwähnt, um einen phonetischen Sprachatlas handelt, sind nach Möglichkeit alle störenden Geräusche abzuschirmen. Zu diesem Zweck werden stets ein Richtmikrophon und eine „Schallbox“ verwendet.

Abbildung 9: Typische Aufnahmesituation bei Aufnahmen für das Projekt VIVALDI

2.4 Transkriptionssystem und Transkriptionen

Bei der Auswahl des Transkriptionssystems lagen besonders pragmatische Gesichtspunkte zugrunde. Drei Anforderungen wurden an das System gestellt:

a) direkte Vergleichsmöglichkeit mit dem Transkriptionssystems des AIS, da dieser nach wie vor das Referenzwerk der italienischen Sprachatlanten ist.

b) Einfache Schreib- und Lesbarkeit sowohl für Projektmitarbeiter als auch für wissenschaftlich Interessierte.

c) Da der Atlas über seinen wissenschaftlichen Zweck hinaus auch dokumentarischen und landeskonservatorischen Wert hat, soll der Zugang zu den Transkriptionen auch für nicht-Linguisten ermöglicht werden.

Nur das AIS-System scheint alle drei Anforderungen zu erfüllen. Aus diesem Grund wurde dem Internationalen Phonetischen Alphabet (IPA) das AIS-System vorgezogen, welches nach und nach an die eigenen Bedürfnisse angepasst wurde. Darüber hinaus wird das System in vielen Standardwerken zur italienischen und romanischen Sprachge-

schichte verwendet (z.B. Tagliavini 1968), sowie in renommierten älteren Einführungen wie Walther von Wartburg (1970) und Gerhard Rohlfs (1966) und sollte daher bekannt sein. Sämtliche Transkriptionszeichen werden auf den VIVALDI Seiten artikulatorisch beschrieben und sind daher einfach konsultierbar.

Transkribiert werden aus dem Fragebuch alle Stimuli des phonetischen Teils. Nicht transkribiert wird dagegen das „Gleichnis vom verlorenen Sohn", da aktuell die dafür benötigten Arbeitskräfte fehlen. Das Transkriptionsverfahren erfolgt in drei Etappen:

1. Vor Ort, direkt nach der Aufnahme, um Unstimmigkeiten beseitigen und eventuelle Zweifel sofort mit dem Informanten klären zu können.
2. Eingabe in die Datenbank mit Hilfe des VIVALDI Transkriptionsassistenten.
3. Endkontrolle vor der Publikation im Internet und auf DVD. Hier wird noch einmal jede Frage von wenigstens zwei Mitarbeitern genau geprüft, Flüchtigkeitsfehler, die bei der Eingabe in die Datenbank auftreten können, werden eliminiert und die Verknüpfung von Ton und Transkription kontrolliert.

Trotz aller Sorgfalt (Kontrolle mit Goldwave und Praat) möchten wir darauf hinweisen, dass es sich dabei um einen Transkriptionsvorschlag handelt, da uns durchaus bewusst ist, dass der individuelle Höreindruck eine gewisse Toleranz bezüglich des Gehörten zulässt. Auch die technische Umgebung (Lautsprecher, Kopfhörer usw.) können u. U. zu leichten Variationen der auditiven Wahrnehmung führen. Um dieses Problem zu beseitigen, bieten wir Interessierten sämtliche Tondateien auf Nachfrage auf DVD an, die dann von diesen kontrolliert werden können.

Bezüglich der Präsentation der Daten folgen wir festen Prinzipien, die auch u.a. von Girnth (vgl. 2010, S. 117) postuliert wurden:

- *Principle of quality* – Daten sollen transparent sein und die sprachliche Realität möglichst genau abbilden.
- *Principle of quantity* – es soll darauf geachtet werden, dass keine Datenlücken entstehen und die Karten den Raum vollständig abbilden.
- *Principle of thematic relevance* – Daten müssen systematisch erhoben werden und die Karten dürfen nur für das Untersuchungsobjekt relevante Daten enthalten.
- *Principle of modality* – Erhebung und Präsentation sollen für die Nutzer nachvollziehbar sein.

Die schnelle Bearbeitung der gesammelten Daten ermöglicht ihre zeitnahe Veröffentlichung im Netz – ein weiteres Prinzip, das sich VIVALDI zum Ziel gesetzt hat und seit Jahren strikt verfolgt. Damit wird vermieden, dass die Daten bei der Publikation schon veraltet sind bzw. Datenfriedhöfe entstehen.

Aus Zeit- und Kostengründen ist es zwar derzeit nicht möglich, das Gleichnis durchgehend zu transkribieren, dennoch dient es Untersuchungen als Kontrollfunktion für vorher abgefragte Stimuli oder um spezielle syntaktische Phänomene im Kontext eines durchgängig gesprochenen Textes zu analysieren (vgl. Mensching 2012). Dann sollte nicht vergessen werden, dass das Gleichnis vom verlorenen Sohn in den vergangenen hundert Jahren immer wieder abgefragt wurde, u. a. auch im ALI, und daher eine ausgezeichnete Quelle für den diachronen Vergleich unterschiedlichster Phänomene bietet.

Die Erfahrung zeigt, dass die Darbietung der gesprochenen Sprache entspricht, obschon wir Wert darauf legen, dass die Informanten das Gleichnis mindestens einen Tag vorher bekommen und vorbereiten können. Ein unabdingbares Kriterium für die systematische Arbeit mit einem längeren Text, damit die Vergleichbarkeit gewährleistet wird. Aus technischen Gründen wird das Gleichnis nach der Aufnahme in 24 Teile segmentiert, was den spontanen Sprechfluss während der Aufnahme jedoch keineswegs behindert.

3 Umgang mit den im Projekt erzielten Forschungsdaten

Die im Projekt systematisch erhobenen und gesammelten Forschungsdaten werden für einen Zeitraum von mindestens zehn Jahren auf Servern der Humboldt-Universität zur Verfügung stehen. Darüber hinaus werden im Projekt regelmäßig, d.h. alle 6 Monate DVDs von allen Daten erzeugt. Diese werden, falls gewünscht, an interessierte Forschungseinrichtungen und Forscher geschickt, die die im Projekt VIVALDI erhobenen Daten nutzen möchten. Natürlich haben diese auch die Möglichkeit, die aktuellsten Daten vom Server der HU kostenlos zu nutzen. Bis auf Weiteres werden wir immer eine aktualisierte DVD (Lebensdauer ca. 30 Jahre) der Institutsbibliothek zur Verfügung stellen, die für alle interessierten Nutzer im Bibliothekskatalog verzeichnet ist und ausgeliehen werden kann.

Es ist uns durchaus bewusst, dass 30 Jahre, besonders im Bereich Sprache und Sprachgeographie, einen relativ kurzen Zeitraum darstellen. Daher sind wir an zwei Archive herangetreten, die es sich zur Aufgabe gemacht haben, Sprachdaten zu sammeln, zur Verfügung zu stellen, zu dokumentieren und zu archivieren. Dies ist zum einen das Projekt *Documentation of Endangered Languages* (DOBES)[10], das besonders für die bedrohten Minderheitensprachen in Italien von Interesse ist. Zum anderen sind wir in engem Kontakt mit der Leitung des Projekts *Gra.fo* (Scuola Normale Superiore di Pisa/Università degli Studi di Siena), das sich darauf spezialisiert hat, sämtliche Dialektdaten der Toskana zu sammeln und zu archivieren. Es wäre natürlich für eine langfristige Archivierung wünschenswert, wenn in den nächsten Jahren Archive entstehen würden, die es sich zur Aufgabe machen, analoge und digitale Sprachdaten zu sammeln, zu archivieren und wie oben beschrieben zur Verfügung zu stellen.

Literaturverzeichnis

Bartoli, Matteo Giulio (Hrsg.) (1995-2008): *Atlante linguistico italiano*. 7 Bände. Roma.

Bauer, Roland/Goebl, Hans/Haimerl, Edgar (2005): *Der „sprechende" Ladinienatlas*, San Martin de Tor.

Calamai, Silvia (2012): „Ordinare archivi sonori: il progetto Gra.fo", in: Rivista Italiana di Dialettologia, 35, 135-164.

Cortelazzo, Manlio (Hrsg.) (1974–88): Profilo dei dialetti italiani. 17 Bände, z.T. mit Schallplatten. Pisa.

Gilliéron, Jules/Edmont, Edmond (1902–1910): Atlas linguistique de la France (ALF). 10 Bände. Paris.

[10] http://dobes.mpi.nl/

Girnth, Heiko. (2010): „Mapping language data“, in: Lameli, Alfred/Kehrein, Roland/Rabanus, Stefan (Hrsg.): Language and Space. Volume 2 Language Mapping, Berlin, Boston, 98-145.

Goebl, Hans (Hrsg.) (1998): Sprachatlas des Dolomitenladinischen und angrenzender Dialekte, 1. Teil (ALD I). 4 Kartenbände, 3 Indexbände, 3 CD-ROM. Wiesbaden.

Harder, Andreas/Boller, Fred (1996): Sprachgeographie und PC. Sprachkarten, Datenorganisation, Tonproben mit Mikrorechnern. Kiel.

Hotzenköcherle, Rudolf/Brunner, Rudolf (Hrsg.) (1972–76): SDS-Phonogramme. Tonaufnahmen für den Sprachatlas der deutschen Schweiz. 16 Langspielplatten. Zürich.

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 (1928–40): Sprach- und Sachatlas Italiens und der Südschweiz. 8 Bände. Zofingen.

Mensching, Guido (2012): „Anmerkungen zur sardischen Syntax anhand des Vivaio Acustico delle Lingue e dei Dialetti d'Italia (VIVALDI)“, in: (Das) diskrete Tatenbuch – Festschrift für Dieter Kattenbusch zum 60. Geburtstag, Berlin. [online: http://www2.hu-berlin.de/festschrift-kattenbusch/mensching-sardisch-syntax.html].

Müller, Marcel Lukas/Köhler, Carola/Kattenbusch, Dieter: „VIVALDI - ein sprechender Sprachatlas im Internet als Beispiel für die automatisierte, computergestützte Sprachatlasgenerierung und –präsentation“, in: Dialectologia et Geolinguistica, 9/2001, 55-68. [online: http://www2.hu-berlin.de/vivaldi/publikationen/vivaldi-sprechender-sprachatlas-2001.pdf].

Reisdoerfer, Joseph (2009): „Géolinguistique et Informatique“, in: Le Monde Blog, http://laurette.blog.lemonde.fr/2009/08/29/geolinguistique-et-informatique/ (Zugriff am: 03.07.2012)

Rohlfs, Gerhard (1966): Einführung in das Studium der romanischen Philologie, Heidelberg.

Ruffino, Giovanni (Hrsg.) (1986): „Prospettive di lavoro per un atlante linguistico-etnografico della Sicilia“, in: Atti della tavola rotonda, Palermo, 11 ottobre 1985. Palermo.

Sobrero, Alberto A./Romanello, Mariateresa/Tempesta, Immacolata (1991): Lavorando al NADIR: un'idea per un atlante linguistico. Galatina.

Tagliavini, Carlo (1968): Le origini delle lingue neolatine, Bologna.

Toso, Fiorenzo (2006): Lingue d'Europa. La pluralità linguistica dei Paesi europei fra passato e presente. Milano.

Wartburg, Walther von (1970): Einführung in die Problematik und Methodik der Sprachwissenschaft, Tübingen.

Wieland, Katharina/Plikat, Jochen/Küster, Lutz: „VIVALDI – eine reiche Fundgrube für Lehramtsstudierende? Fremdsprachen- und hochschuldidaktische Überlegungen“, in: Carola Köhler/Fabio Tosques (Hg.), (Das) diskrete Tatenbuch. Digitale Festschrift für DIETER KATTENBUSCH zum 60. Geburtstag, Berlin 2012. [online: http://www2.hu-berlin.de/festschrift-kattenbusch/kwp-vivaldi-didaktik.html].