Master Thesis

submitted within the UNIGIS MSc programme
at the Centre for GeoInformatics (Z_GIS)
Salzburg University

Geolinguistic GIS Applications: Aspects of Data Quality in Mapping Lesser-Used Languages

by

Mag. phil. Ulla Briscoe
9812423/U40136

A thesis submitted in partial fulfilment of the requirements of
the degree of
Master of Science (Geographical Information Science & Systems) – MSc (GISc)

Advisors:

Dr Adrijana Car, University of Salzburg

Dr Ljuba Veselinova, University of Stockholm

Graz, November 2009

Science Pledge

By my signature below, I certify that my thesis is entirely the result of my own work. I have cited all sources I have used in my thesis and I have always indicated their origin.

Ulla Frisch

Graz, 20 November 2009

Place, Date

Signature

Abstract

The use of Geographical Information Systems (GIS) in the field of geolinguistics is gaining momentum, but is being held back by an apparent lack of critical analysis with regard to data sources, data quality, uncertainty and data documentation. This thesis investigates which of these key issues arise in geolinguistic applications by investigating the location, dialect situation and estimated number of speakers of five lesser-used languages in a sample application. The data sources are sociolinguistic survey reports which are representative of the kind of materials geolinguists work with.

The results show that the quality of geolinguistic data to be used in GIS is above all affected by a lack of primary data. The subsequent re-use of data which were originally collected for a different purpose results in a discrepancy between the two concepts employed by the data producer and data user. This constitutes a source of error in addition to the intrinsic uncertainty of the original data which was introduced during linguistic data collection. The aspect of intrinsic uncertainty includes *inter alia* bias due to the political or social background and motivation of both the interviewer and respondent – factors whose assessment and measure is rather complex. Furthermore, lacking documentation of the data collection process from a GIS perspective compounds decreasing data quality by not allowing the user to establish whether or not the data are fit for the respective use. The other major contributing factor is the scarcity of reference data to critically assess the consistency of thematic and positional data accuracy as well as data completeness.

Table of Contents

List of Figures

List of Tables

1. Introduction

1.1 Motivation

Geographic Information Systems (GIS) and linguistics are engaged in a relatively young relationship which until now has mainly focused on the results of individual linguistic GIS applications, rather than their development. The resulting lack of research from a GIS perspective includes the area of data quality, which in itself has an established place in general GIScience and is the subject of numerous conferences and publications (see for instance Shi, Fisher and Goodchild (eds.) 2002; Devillers and Jeansoulin (eds.) 2006).

It is therefore my intention to address this weak spot by examining geolinguistic data sources and the quality of their data. The focus on lesser-used languages was motivated by the fact that these elements of the world's linguistic heritage are usually not as well examined and documented as more widely used languages, and as such offer a more challenging investigation. The choice of sample languages from Ghana and Papua New Guinea reflects the intention of covering two distinct parts of the world. I expect to consolidate aspects of the two disciplines of GIS and linguistics by providing an analysis of data quality aspects from which future geolinguistic GIS applications can grow, hence making linguistic data fit for – and fit into – GISystems.

1.2 Task

The objective of the thesis at hand is the investigation and analysis of aspects of data quality in geolinguistic applications mapping lesser-used languages in Geographical Information Systems and Science. As such it aims to contribute to establishing general guiding principles for contemporary language mapping which have so far not been adequately addressed.

1.3 Approach

Although my research is targeted mainly at linguists, the investigation is conducted from a GIS perspective, rather than a linguistic one. This includes the development of the fundamental GIS data modelling process and the database design stages of conceptual, logical and physical modelling for my sample application.

An examination of current approaches to language mapping such as the GIS in Linguistics project (GISLI; http://ling-map.ling.su.se/website/index.html) and the Language and Location: A Map Annotation Project (LL-MAP; http://www.llmap.org/; still under development at the time of writing in September 2009) forms the basis from which I adapt a conceptual design for my own investigation as appropriate. The results of the conceptual stage demonstrate what the desired output is, which data are needed to arrive at this output, and which data are actually available. The data are then investigated while bearing in mind the quality requirements for correctly mapping the location of lesser-used languages in a GISystem and for providing relevant additional information such as the number of speakers or dialects.

The data sources and data for the sample languages of Chakali and Safaliba (Ghana), as well as Uyajitaya, Ambakich and Sam (Papua New Guinea) are representative of the kind of data linguists work with and as such allow me to focus on any data quality issues that arise. The main data sources are the sociolinguistic survey reports published by SIL International (formerly known as the Summer Institute of Linguistics). SIL International focuses *inter alia* on the documentation of lesser-used languages and is the publisher of the Ethnologue, “an encyclopedic reference work cataloging all of the world’s 6,912 known living languages” (http://www.ethnologue.com/; Lewis 2009). There is currently no defined number of speakers from which a language is considered a ‘lesser-used language’ (see also Ambrose and Williams 1991: 309). Dahl (2006: 3) suggests languages of less than 10,000 speakers be given this label, a threshold I have adopted in this thesis.

Sources of uncertainty in the data and data sources are identified based on Longley, Goodchild, Maguire and Rhind’s (2005) and Brimicombe’s (1997) approaches to categorising uncertainty in GIS. Longley et al. (2005: 128-153) advocate the classification of stages where uncertainty is introduced into three filters. The focus in this thesis is on filters U1 and U2 which describe uncertainty introduced in the conception of geographic phenomena and in the measurement and representation of these phenomena respectively. Brimicombe (1997: 115-116), on the other hand, promotes the division into four broad categories. The two types relevant to this investigation are *intrinsic uncertainty,* which is found in data due to the original data collection process, as well as *inherited uncertainty,* which occurs due to the use of secondary data.

The International Organization for Standardization's (ISO) standard 19113 Geographic Information – Quality Principles (ISO 2009: 50-51) acts as a reference where appropriate, thereby allowing the drawing of conclusions about which data quality issues regarding positional and thematic accuracy, as well as data completeness, are crucial for effective geolinguistic research and the use of such data in GISystems. The discovered shortcomings and strengths of the data and data sources used in this application are a first pointer towards recommending and adopting standards and criteria which have so far mostly been used in the natural, technical and commercial sciences for linguistic research.

1.4 Expected results

As there is currently no common, well-founded model in place which describes and illustrates the quality issues arising from the use of geolinguistic data for mapping the distribution of lesser-used languages in GIS, I aim to establish such a set of aspects in this thesis. The key questions I set out to answer are:

- Does the use of geolinguistic data sources and their data quality pose special challenges to GIS developers and users?
- What are these challenges and problems with regard to mapping lesser-used languages and linguistic GIS applications in general?
- Where are elements of uncertainty introduced into geolinguistic data in GIS applications?

I expect to contribute thereby to geolinguistic research by raising awareness of the effects of data quality for both spatial as well as attribute data in geolinguistic research. Moreover, I compile data for some of the world's lesser-used languages (Chakali, Safaliba, Uyajitaya, Ambakich and Sam), thus providing a new dataset to be added to the Language Map Server, which was set up by Dr Ljuba Veselinova and Prof Östen

Dahl of Stockholm University’s Department of Linguistics (see http://ling-map.ling.su.se/website /index.html).

1.5 Intended audience

During my initial research into geolinguistic GIS applications it became apparent that GISystems in linguistic research are mostly used by either linguists who often have no professional training in GIScience but have acquired skills, mainly in software use, autodidactically, or by linguists collaborating with geographers, IT specialists and statisticians. While the first frequently appear reluctant to exploit the full potential of GISystems and Science because of modest background knowledge of spatial science, the latter rely on outside help and collaboration with other specialists who in turn may have little or no background in linguistics. It is therefore the objective of my thesis to provide linguists with helpful pointers for using linguistic data in research by means of GIS, so that they can confidently yet critically assess and use geolinguistic data.

I am aware that some of the terminology common to GIS specialists may be new to researchers with an academic background in linguistics, which is why I have explained some core concepts of GIS and provided references for further study in cases which seemed relevant to me. Linguists cannot be expected to be fully qualified statisticians or IT specialists, yet I hope to have struck the right balance between what geolinguists with little or no previous GIS experience can be expected to acquire and what is essential when using GISystems and applying GIScience.

1.6 Issues which are not discussed

Despite the apparent need for established references and guidelines for language mapping in general, I only concentrate on data used for mapping lesser-used languages. I am aware that this constitutes only a fraction of the potential areas in which GISystems can be employed - fields such as dialectometry or historical linguistics offer great opportunities for such research, yet including them would go beyond the thematic and temporal scope of this thesis.

Choosing a certain selection of sample data and data sources clearly also imposes limits on an analysis, yet the aspects of mapping lesser-used languages are covered as comprehensively as the scope of this thesis allows. Although multilingualism does occur in the areas investigated, only the languages under investigation are mapped. I understand that especially the use of GIS opens up new ways of mapping multilingualism, yet this is not the focus of this thesis.

The comparison and analysis of geolinguistic data and quality standards is conducted in a way that is accessible for linguists. The data quality of, for instance, topographic or political information, which is usually used in addition to geolinguistic data, is not discussed. This thesis does not provide a universally valid ‘recipe’ of how to assess data, compile full quality evaluation reports based on complex data quality measures, or apply statistical models and methods to determine error propagation. I do not consider such an approach viable, bearing in mind that the focus of this thesis is on *data* and considering the current stage at which geolinguistic research using GIS is. Rather, I intend to raise crucial issues of geolinguistic data quality with regard to thematic and positional accuracy, as well as data completeness and elements of uncertainty. Hence I provide a starting point for further, more in depth analysis in linguistic research using GIS technologies.

1.7 Thesis structure

This introductory chapter is followed by a review of relevant literature in chapter 2, discussing landmark papers and the most recent developments in GISystems and Science and geolinguistics as well as data quality. Chapter 3 (“Approach”) lays out the theoretical foundations and methods of my research. It also includes the definitions of certain key terms and a brief description of the software used. The fourth chapter discusses the concept and implementation of my research project in detail, allowing the discussion and analysis of the results in chapter 5.

Finally, the conclusion (chapter 6) contains a brief summary and discussion of the investigation and its results, as well as suggestions of which future paths of research in the field discussed may yield interesting and valuable results. Figure 1 illustrates the individual chapters and key contents for easier orientation:

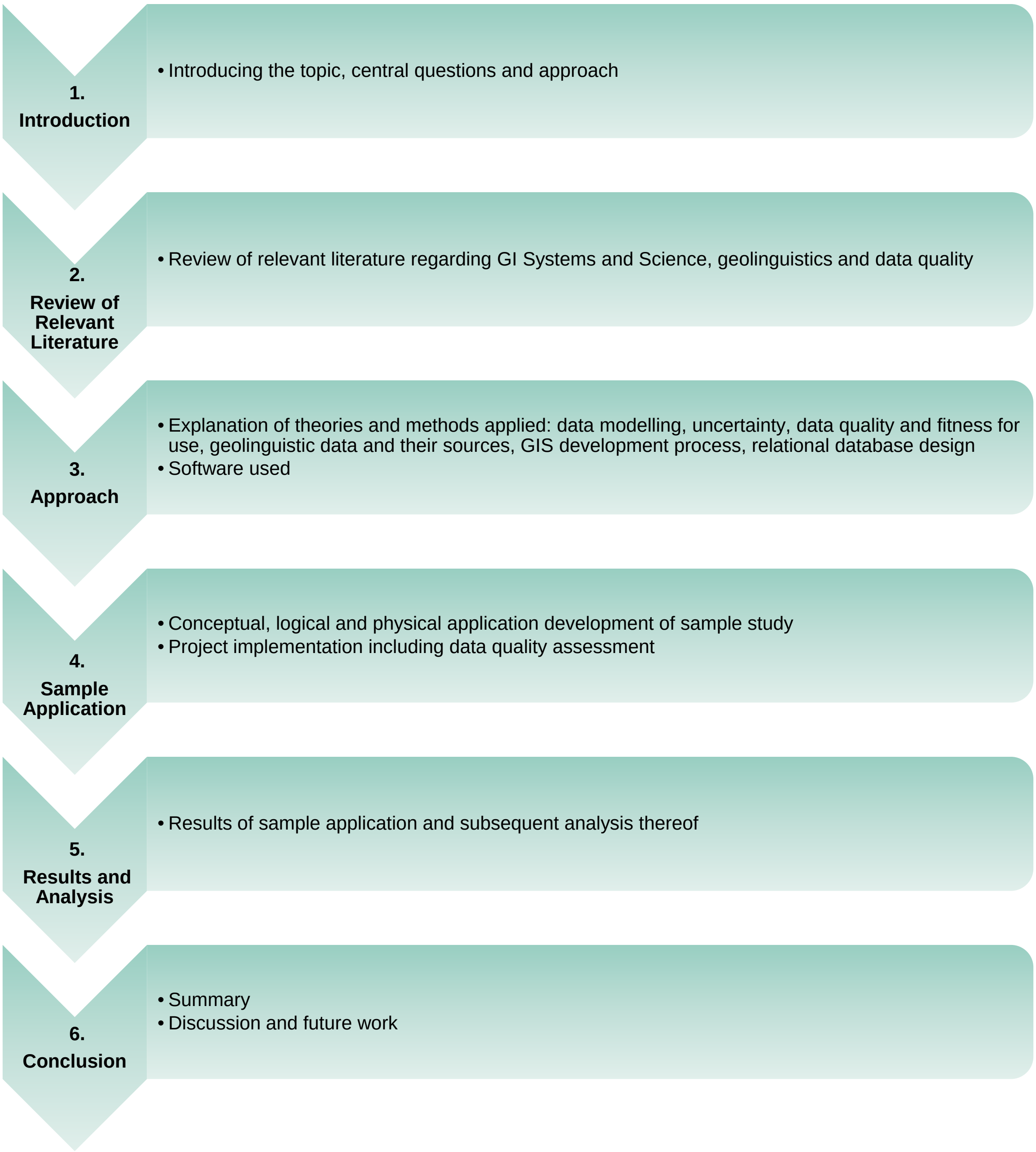

1. Introduction
• Introducing the topic, central questions and approach
2. Review of Relevant Literature
• Review of relevant literature regarding GI Systems and Science, geolinguistics and data quality
3. Approach
• Explanation of theories and methods applied: data modelling, uncertainty, data quality and fitness for use, geolinguistic data and their sources, GIS development process, relational database design
• Software used
4. Sample Application
• Conceptual, logical and physical application development of sample study
• Project implementation including data quality assessment
5. Results and Analysis
• Results of sample application and subsequent analysis thereof
6. Conclusion
• Summary
• Discussion and future work

Figure 1 Thesis structure

2. Review of Relevant Literature

To position my thesis in its wider context and bearing in mind my target audience, I have included some introductory GIS literature in this section, followed by resources dealing with geolinguistics as a discipline and its current use of GIS, as well as aspects of spatial data quality in general. Please note that the references mentioned are not exhaustive due to the limited scope of the thesis, but only represent a selection of the resources available.

2.1 Geographic Information Systems and Science

Over the past decades, the acronym GIS has seen many attempts at definition, none of which seem to comprise all its aspects and potentials. Originally, GIS used to solely describe Geographic Information Systems, the first of which – although basic by current standards – was designed by Tomlinson in 1963. Since then, the progress of technology and the apparent demand and possibilities for spatial analysis have driven the development of GIS to arrive at what Dueker and Kjerne define as "a system of hardware, software, data, people, organizations and institutional arrangements for collecting, storing, analyzing, and disseminating information about areas of the earth." (1989: 7-8). The use of GIS is, however, no longer restricted to the earth – GIS are also employed to investigate for instance other planets or the human body. As such, the term 'spatial information systems and science' may better convey the outline of the field as described by Longley et al. (2005: 8), yet the more commonly used adjective remains 'geographic'. Dueker and Kjerne's definition above lists the major components of a GISystem, which are also described in established textbooks such as Longley et al.'s *Geographic Information Systems and Science* (2005: 18-24).

The fundamental term 'Geographic Information Science' was coined by Goodchild in *Geographic information science* (1992: 31-45), a landmark paper which discusses the need for GIS to consider the issues surrounding spatial data and their processing in GISystems. Following his arguments, a distinction between GISystems and GIScience is now commonly drawn: GISystems describe the technological aspect and are often regarded as a tool, while GIScience comprises aspects of research before and beyond the technological application such as methodology, spatial analysis, interpretation and so on.

Further information about the fundamentals of GISystems and Science are for example available in Korte’s *The GIS Book* (2001), or online in Buckley’s *The GIS Primer* (1997, online), The Geographer’s Craft by the Department of Geography of the University of Colorado at Boulder (2000, online), or an introduction to GIS written by Goodchild in the educational resources at the National Center for Geographic Information and Analysis (1997, online).

2.2 Geolinguistics and GIS

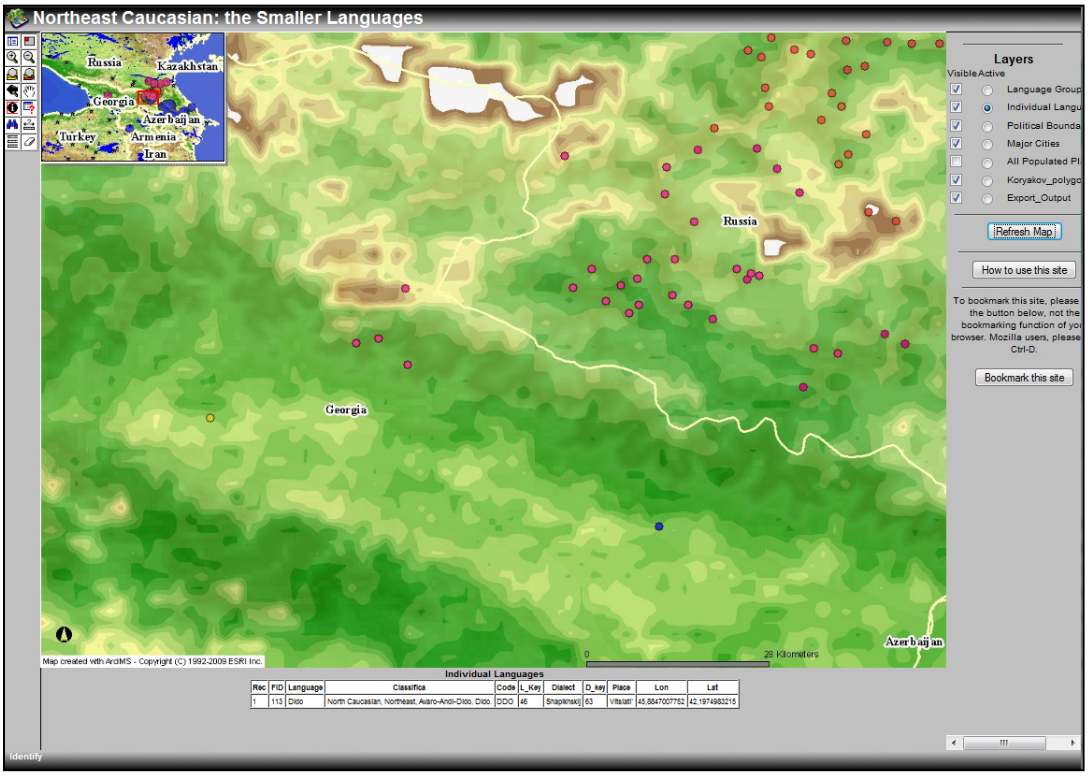
My research into studies dealing with linguistic research and GISystems and Science in general has shown that this combination of disciplines is still in its infancy. Luo, Hartmann, Li and Sysamouth’s conclusion that “GIS mapping and analysis [...] has great potential in linguistic geography research” (2000: 135) is representative of a number of researchers praising the potential of GISystems, while at the same time not yet fully exploiting it. Methods such as spatial analysis or diachronic investigations using GIS are often still treated as novel techniques or experiments (for instance in Fukushima and Heap 2008: 144-145), showing that GISystems in (geo)linguistics are still far from being an everyday tool confidently used by linguists.

Investigating the terminological point of intersection between linguistics and geography yields a variety of terms referring to this area of study. The fuzzy terminology of ‘geolinguistics’, ‘language geography’, ‘geographical linguistics’ or ‘linguistic geography’ used across the scholarly landscape is symptomatic for the diffused outline and origin of this field. I have come across linguistic geography being referred to as “an independent discipline in linguistics” (Fukushima and Heap 2008: 138), as well as Geolinguistics being called “an evolving branch of human geography” (Williams 1996: 63). The thematic overlap of linguistics and geography deals *inter alia* with issues central to both disciplines: concepts such as spatiality (written comprehensively about by Britain 2004, *in press (a) and (b)*), migration, identity, ethnicity, political and policy issues etc. In this thesis, I use the umbrella term geolinguistics and respect it as an interdisciplinary field of research with origins in both linguistics and geography.

Tracing the origins of geolinguistics as a discipline in general and its dealing with data and data sources is not a very long journey. The first key works of language mapping were indeed part of dialectological studies – Wenker's *Sprachatlas des Deutschen Reichs* (unpublished; survey period 1876-1887), which relied on data collected in questionnaires sent out via post, or Gilliéron and Edmont's *Atlas linguistique de la France* (1902-1910), whose data resource were interviews conducted as part of fieldwork (Crystal 1997: 26). Yet in *Linguistic Minorities, Society and Territory* (1991: 298), Ambrose and Williams refer to "the developing discipline of geolinguistics". Although geolinguistic research has in fact existed since before the 1990s, most notably by Breton (e.g. *Géographie des langues* 1976, developed into *Geolinguistics: language dynamics and ethnolinguistic geography* 1991), it seems that the 'discovery' of GIS as a means of expanding the boundaries of geolinguistic research has certainly boosted its attractiveness.

The earliest application of combining geolinguistics and IT was *The Generalized Linguistic Atlas Printin*g System (GLAPS) developed by Ogino in 1975 (Ogino 1980) and C. and Y. Fukushima's *System of Exhibition and Analysis of Linguistic Data* (SEAL) programme in 1983 (Fukushima and Fukushima 1993). Both software systems originated in dialectology, yet nowadays most other geolinguistic applications use readily available software.

One example of such an application using the Environmental Systems Research Institute's (ESRI) software products is the GISLI project at Stockholm University, to which this thesis's data research will contribute. This project is set up and maintained by Dahl and Veselinova as described in *Language Map Server* (Dahl and Veselinova 2005). It aims to map the world's lesser-used languages and to provide additional information such as demographic data. An extract of this map service is shown as an example in Figure 2, displaying the location and available additional information about the lesser-used Caucasian languages of the border area between Russia and Georgia. The display of additional information is possible due to the layer structure of GIS, in which related geographic objects are collected in individual layers, which can then be added or removed depending on the user's requirements.



Layers

Visible	Active	
☑	○	Language Group
☑	●	Individual Langu
☑	○	Political Bounda
☑	○	Major Cities
□	○	All Populated Pl
☑	○	Koryakov_polygo
☑	○	Export_Output

To bookmark this site, please the button below, not the bookmarking function of yo browser. Mozilla users, please Ctrl-D.

Individual Languages

Rec	FID	Language	Classifica	Code	L_Key	Dialect	D_key	Place	Lon	Lat
1	113	Dido	North Caucasian, Northeast, Avaro-Andi-Dido, Dido.	DDO	46	Shapikhskij	63	Vitsiati'	45,8847007752	42,1974983215

Figure 2 Extract from GISLI map service showing the location of some of the lesser-used Northeast Caucasian languages (accessed November 2009)

The distinctive characteristic of GISLI, an extract of which is shown in Figure 2, is the choice of representing the language locations as point features rather than polygons, which are more commonly used in language mapping (cf. Dahl 2006: 3). Contrary to GISLI, the LL-MAP project uses polygons, an issue which is discussed further in subsection 4.1.3.

Other recent applications of GIS in linguistic research include Luo et al.'s *GIS Mapping and Analysis of Tai Linguistic and Settlement Patterns in Southern China*, which describes GIS as the "technology" (2000: 130) used in combination with comparative-historical linguistic research methods. Although methods of spatial analysis are applied in the investigation, little attention is given to reflection on techniques such as the contour interpolation being performed (2000: 133). Six years later, Wang, Hartmann, Luo and Huang provided a more detailed description and analysis of the spatial modelling performed in their analysis of Tai place names (2006). Besides its investigation into toponymy, this paper offers introductory information about techniques such as trend surface modelling (2006: 4) and cluster analysis (2006: 5) and how these were applied to the data under investigation.

However, information about the theory and traditions of spatial analysis and GIS are not yet common in the majority of GIS applications in linguistic research. As such, the approach of using GIS merely as a technology in linguistic applications without theoretical considerations appears to be the rule rather than the exception. Theoretical investigations into (geo)linguistics and GISystems can, in my experience, mostly only be found as brief sub-sections of practical examinations, rather than being recognised in scientific discussions in their own right. It is therefore symptomatic that I have been unable to locate much literature which critically reflects upon which data are being used in geolinguistic projects employing GISystems to ensure the highest possible suitability and 'fitness for use' of linguistic data in general terms.

Barni's paper *From Statistical to Geolinguistic Data: Mapping and Measuring Linguistic Diversity* (2006) proves to be a rare exception to this. She details the different data-collection models applied and describes that all data were geo-referenced and digitally recorded during collection (2006: 5-8). Barni moreover recognises the potential provided by the data collection and storage method not only for synchronic but also diachronic investigations and analyses. However, first hand data acquisition is not always an option for linguists, who often have to rely on data recorded by other parties and in many cases even for different purposes.

2.3 Spatial data quality

When looking at research into spatial data quality in general, there is – contrary to geolinguistic data quality – an abundance of literature available. Fisher, Comber and Wadsworth (2009: 5-8) as well as Devillers, Gervais, Bédard and Jeansoulin (2002: 45) point out the – in my opinion central – problem of the current wide-ranging availability of spatial data and the diverse stratification of data users posing a key issue in assuring data quality. The interchange of spatial information is in many cases no longer an act between well-informed specialists in the production, distribution and usage of datasets. Instead, there is often no direct contact between data providers – let alone data producers – and data users. Dahl (2006: 2-3) and other linguists such as Fukushima and

Heap (2008: 150, 156) stress that current technology offers great opportunities for sharing geolinguistic data, for instance in the form of online and freely accessible datasets. I believe, however, that while it is certainly a viable and noble aim, the same caution is required with a geolinguistic dataset as with any other, particularly if consequences such as policy decisions for language preservation depend on linguistic data originally collected for a different purpose.

One of the most established means to combat this overwhelming situation of spatial data availability are metadata. Metadata are commonly referred to as "data about data" (e.g. Shamsi 2005: 97; Longley et al. 2005: 245), and give the user essential information about a dataset or database. Further information about metadata is provided in subsection 3.1.3. Chrisman (2009: 30) emphasises that metadata are only of use if both the data producers and the data users consciously work with them, a view also supported by Longley et al. (2005: 152).

However, as Zargar and Devillers (2009: 1) point out, many users do not consider consulting the metadata – if available – to see whether a dataset is indeed appropriate for the intended use. Another problem put forward by Fisher et al. (2009) concerns the semantic variety, i.e. the conceptual differences behind the vocabulary describing geographical phenomena among data producers or users, used across the field of GIS and consequently also in metadata. Their quoted example (2009: 14) regarding the wide-ranging standards of the minimum requirements for an area to be considered a 'forest' in various countries (minimum tree height and canopy cover) nicely illustrates this dilemma. While I acknowledge that semantic diversity poses a challenge to both data producers as well as users, I do not believe that Fisher et al.'s suggestion to clarify the meaning of information provided in metadata (2009: 53) can indeed be implemented on a global scale due to reasons of practicability. In my opinion, critical use of available data and the existing metadata are – for time being – as close as data producers and users can realistically get to working with data fit for their purpose.

Devillers and Zargar (2009: 4; see also Zargar and Devillers 2009: 2) examine the latest trend towards developing 'intelligent' software which assesses the quality of used data simultaneously to the spatial operation being performed by a GIS user. This would mean a tremendous reduction in time and effort placed on assessing data and metadata by the individual user – yet as of now, the standardised application of such software is still a pipe dream.

My exploration of the resources has shown that although there is a lot of literature about spatial data quality in general, the ‘thematic overlap’ of linguistics and GIScience still misses the essential analysis and consideration of the peculiarities of spatial linguistic data and their sources – and this is where my research ties in.

3. Approach

This chapter lays the foundation for my investigation by providing information on the theoretical background and on the methods used in the sample application of this thesis. The Methods section is largely based on Longley et al.'s *Geographic Information Systems and Science* (2005), a resource accessible for all newcomers to the field of GIS.

3.1 Theoretical background

3.1.1 Combining linguistics, geography and GIS

Notwithstanding its diffuse outline as a scientific field, geolinguistics is widely regarded as a hybrid between two disciplines: linguistics and geography. While linguistics has its roots in philological studies, other fields such as the social sciences (sociolinguistics) or anatomy and physical sciences (e.g. clinical linguistics or neurolinguistics) have inevitably entered the realm of linguistics, forming an integral part of this science. Geography also comprises a variety of approaches and sub-fields such as behavioural or political geography to name only a few, as well as, of course, physical geography. It is the latter which is seen as the developmental cradle of GIS, as most applications were originally set in the natural sciences.

Contrary to perceptions of physical geographical phenomena, those associated with language and language use are often loaded with prejudice, emotional attachment, issues of imposing political power or ethnic and territorial clashes. The consideration that language plays a vital role in a person's identity has been discussed extensively (see for instance Joseph 2004; Gubbins and Holt, eds. 2002; Fishman, ed. 1999), and has extended into the realm of cartography too. Peeters (1993: 7-8) advocates that it is the sensitive issues of language and its use which makes mapping languages and language variations a rather complex undertaking. Yet the underlying understanding, namely that languages are a phenomenon which varies across space and which is thus one with a spatial dimension, remains.

Williams (1996: 63), whose primary background is in geography, quotes and underlines van der Merwe's (1993: 35) point that geolinguistics as a discipline makes use of many geographical concepts such as *location*, *space* or *interaction*. On the other hand, Britain (2004: 34) criticises that the "geographical dimension of space" is an "almost wholly underexplored dimension in sociolinguistics". This again shows how linguistics and its sociolinguistic element, which in turn is a component of geolinguistics, appears reluctant to use geographical concepts and to merge with them. Yet bearing in mind that even in geography, the scientific discussion about concepts of space is still ongoing (see for instance Thrift 2003 and Kent 2003), it is not surprising that other disciplines such as linguistics attempt to define and use 'space' applying their own parameters.

Despite scholars such as Ambrose and Williams (1991: 301-302) questioning whether linguistics and geography are indeed compatible because their "relationship rather rarely seems like a true 'meeting of minds'", I believe that full compatibility is not necessarily the main goal. The objective should rather be the establishment of a middle ground on which the plethora of methods and options offered by GISystems and Science to linguistics, ranging from data handling and analysis to modelling predictions and beyond, can be effectively exploited without compromising too much on both disciplines' central tenets.

Judging from my research for this thesis, it is my understanding that the main interest of linguists using GIS lies in the output of an application, rather than in how this output was produced. This is understandable, particularly bearing in mind that GISystems are in this case used as tools operating to support research in a certain discipline. However, it is essential to consider the 'science' aspect of GIS rather than merely using the 'system'. These considerations include *inter alia* aspects such as the consistency and compatibility of the datasets used, or in cases where interpolation (i.e. techniques to predict data values at unknown locations) is employed, the choice of the individual interpolation method (see for instance Table 6-2 in de Smith, Goodchild and Longley 2006-2009, online, for a list of methods and their benefits and disadvantages). This is just a fraction of the considerations which GIS users – including linguists – have to take into account, as the decisions based on these factors will inevitably alter the output of a GIS.

3.1.2 Data modelling and uncertainty

At the basis of each GIS application is the user's awareness that any GIS uses models – and as such abstractions and simplifications – of the real world (Longley et al. 2005: 64-65). These abstractions and simplifications are necessary to make the complex nature of reality digitally accessible to an IT system, as well as making it more comprehensible for the human mind (Mark 1999: 81-89). Even analogue maps are models of real world phenomena, as they represent a simplified, selected view of the world. Longley et al. (2005) provide a clear illustration of the levels of abstraction in GIS data models as shown in Figure 3:

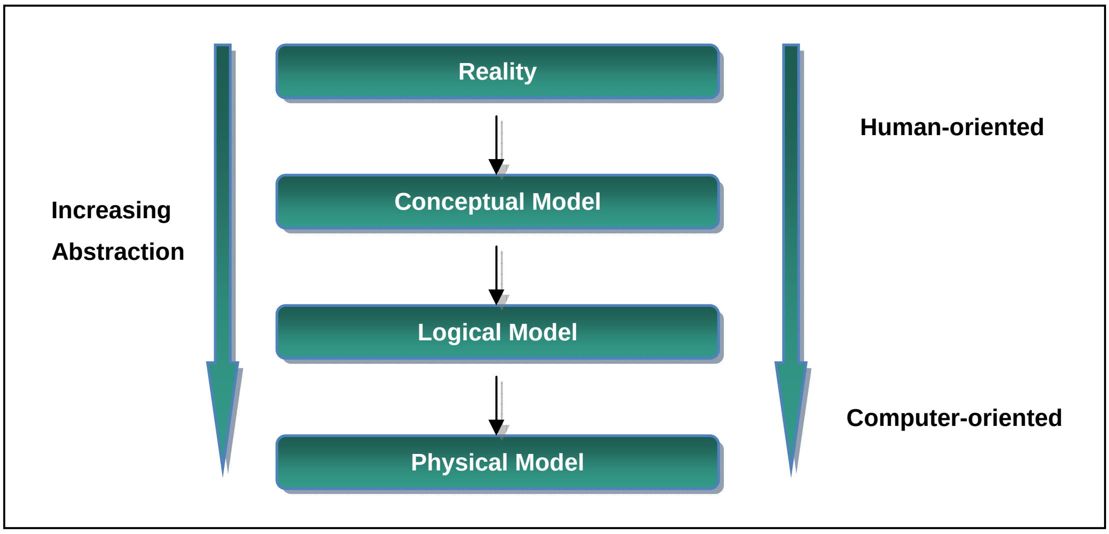


Figure 3 Levels of abstraction relevant to GIS data models (source: Longley et al. 2005: 179)

The three modelling levels shown above effectively constitute the three stages in database design, where increasing abstraction leads from reality to the implementation in a database. These stages are described in detail in the Methods section (3.2) below.

It is in my view essential that geolinguists' attention is drawn to the fact that all (spatial) data – including linguistic data – are to a greater or lesser extent flawed and that these problems will propagate when developing and using GIS applications. This element of uncertainty is used synonymously with 'error' in the literature and in this thesis. 'Error' in this context not only comprises the most obvious meaning of 'mistake' but also 'variation' (see also Heuvelink 1997, online).

Figure 4 taken from Longley et al. (2005: 129) neatly illustrates the filters (U1 to U3) which add uncertainty at different stages during increasing abstraction:

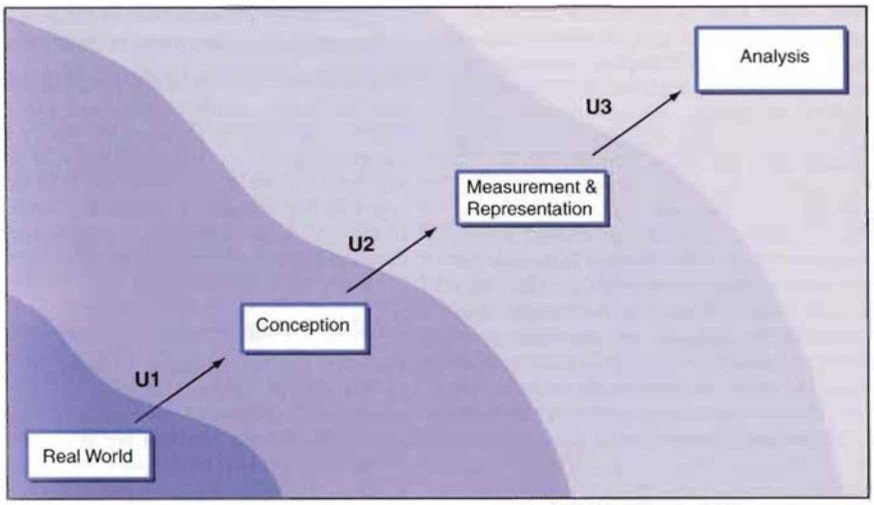


Figure 4 A conceptual view of uncertainty (source: Longley et al. 2005: 129)

Uncertainty introduced at the conceptional stage of spatial phenomena (U1 in the above figure) may result from factors such as problems determining appropriate units of analysis, vagueness and ambiguity. In the case of linguistics, these factors may for instance be semantic differences in the perception of linguistic phenomena. This includes the widely discussed question as to when a dialect is considered a language in its own right. The decision whether or not to assign a certain dialect the status of a language will have inevitable effects on the presence or absence of this specific language/dialect in the representation and subsequent analysis, as well as potential policy decisions depending on these analyses.

Filter U2 adds uncertainty through for example the discrete representation of objects or measurement errors which may for instance change the class to which a certain object is assigned. For the sample application in this thesis, this filter includes the question of how to represent the occurrence of a language: either as polygon or point features (see subsection 4.1.3 for a detailed discussion). Filter U3, which is not discussed in this thesis, increases uncertainty through analysis of inherently uncertain spatial phenomena, which may be tackled through internal or external validation (see Chapter 6 in Longley et al. for details).

Brimicombe (1997), by comparison, describes four levels of uncertainty, as shown in Figure 5:

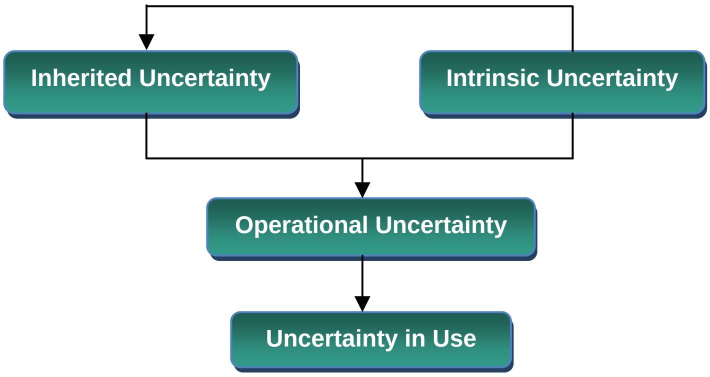


Figure 5 Broad categories of uncertainty encountered in a GIS (source: Brimicombe 1997: 116)

He explains the levels as being introduced at the primary data collection stage (*Intrinsic Uncertainty*), when using secondary data sources (*Inherited Uncertainty*) and while operating on the data using soft- and hardware (*Operational Uncertainty*). The last stage (*Uncertainty in Use*) arises from the GIS users' failure to recognise and deal with the elements of uncertainty introduced during the first three uncertainty levels. It is particularly the first two categories which are of interest in this thesis because at these two stages, specific conclusions about geolinguistic data – in contrast to other data more commonly used in GIS – can be drawn. In the sample application conducted, these two types of uncertainty will include issues arising from the use of geolinguistic data sources such as sociolinguistic surveys and censuses, as well as the use of their data for purposes other than that for which they were originally intended.

To summarise, we have to accept uncertainty as an inevitable predicament resulting from the selective and subjective abstraction process, the data sources, as well as the operations performed on the data and the subsequent interpretation of their results by the user. Locating such sources of uncertainty and being aware of them is of great importance and benefit to developers and users of GIS alike. Longley et al.'s proposed definition of uncertainty as "a measure of the user's understanding of the difference between the contents of a dataset, and the real phenomena that the data are believed to represent" (2005: 128) comprises all stages during the process from conceptualisation to data use. I believe it is important to call attention to this understanding particularly in fields such as geolinguistics, where GISystems and their flaws have so far not been discussed explicitly.

3.1.3 Data quality and 'Fitness for Use'

"Garbage in, garbage out" – this overemployed phrase is often quoted when discussing data quality in GISystems (see for example Korte 2001: 223; Shamsi 2005: 140). It refers to the fact that a GISystem's output can only be as good as its input – i.e. the data. This may sound obvious and rather blunt, but I consider it important to raise geolinguists' awareness of (spatial) data quality.

When thinking about data quality in GIS, one may be tempted to merely refer to the positional accuracy of data. However, there are several equally important issues to be considered which are commonly referred to as 'data quality elements' and which are introduced in subsection 3.2. Generally, data quality can be determined by the degree to which it addresses both the user's requirements and by how detailed and accessible the data's description is. The data description is usually found in the metadata, a term introduced in subsection 2.3 above. User requirements vary from application to application and concern questions such as how detailed the output maps have to be, or which spatial analyses and operations the user wants to conduct with the dataset. The range of spatial as well as statistical analyses offered by GIS which linguistic research can benefit from is extensive – please refer to, for instance, de Smith et al.'s *Geospatial Analysis – a comprehensive guide* (2006-2009), which is available free of charge online.

The definition of data quality as "the difference between a dataset and a universe of discourse" (Jakobsson and Giversen, eds. 2007: 18) as illustrated in Figure 6 forms the basis of the ISO 19113 standard, using a similar concept to the one described by Longley et al. in subsection 3.1.2 above.

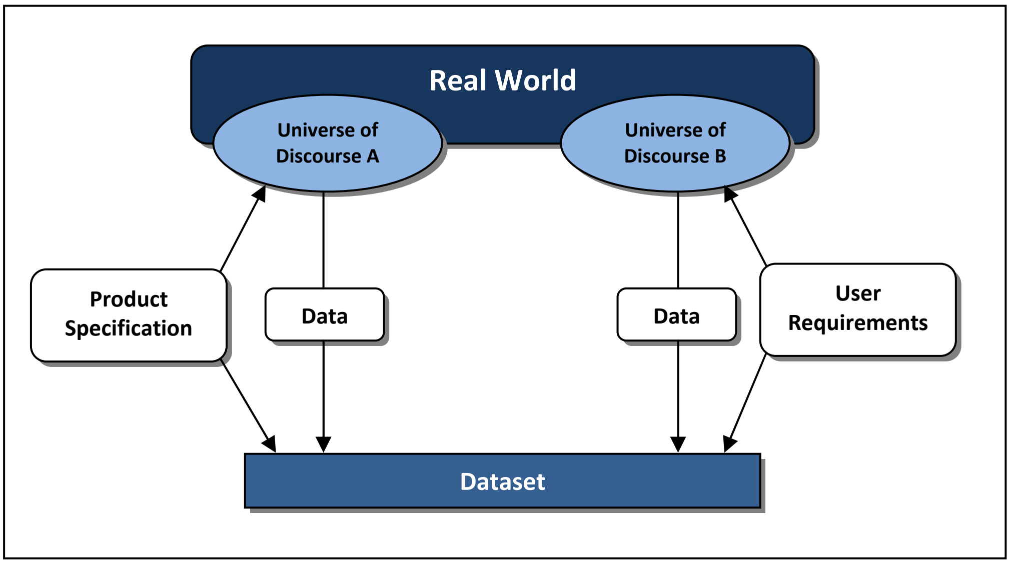


Figure 6 Concept of data quality in ISO 19113 (adapted from Jakobsson and Giversen, eds. 2007: 18)

Since the data producer and the data user are often different people working in and with a different universe of discourse, the gap between their discourses creates an area of uncertainty. This “difference” comprises problems such as semantic discrepancies between the vocabulary used by data producers and users – for the application at hand, such an example could be the definition of ‘dialect’.

One of the catchphrases when discussing data quality is certainly the term ‘fitness for use’. This does not, however, necessarily mean that ‘fitness for use’ is the equivalent of a dataset’s ‘quality’. It stresses the fact that data are usually needed for a specific ‘use’ and a dataset which is ‘fit’ for a certain application or user may not be suitable for another. Moreover, factors such as the accessibility of a dataset – both physical access as well as intellectual access (i.e. being able to fully understand the data) – come into play when assessing whether or not a specific set of data are indeed ‘fit’ or not. Considerations of which requirements a dataset has to fulfil in order to be deemed of use are usually part of the initial stages in a GIS development process (see 3.2 “Methods”) and are as such an indispensable stage in any GIS application. According to Foote and Huebner (1995, online), the fitness for use is usually established bearing in mind how *accurate*, how *precise* and how *complete* the output and therefore the data have to be. This means finding the right balance between two levels: what levels of inaccuracy, imprecision and incompleteness are still acceptable bearing in mind the application’s purpose – and which levels would render the application worthless? It is essential to note that there is indeed a difference between the terms *accurate* and *precise*, as shown in the definitions according to the Open University (online):

> *Accuracy* is a measure of how close a result is to the true value.
>
> *Precision* is a measure of how repeatable the result is.

In addition, precision is also understood to be the amount of detail in a measurement. The varying priorities between applications can be illustrated by the example that for instance applications in road and utility construction depend greatly on precise measurements, while others such as demographic analyses of electoral trends may be able to compromise on the precision level for reasons of cost and time (Foote and Huebner 1995, online).

The term *data quality* does not necessarily mean that data can be assigned the labels ‘good/bad’, or ‘fail/pass’. *Quality* also refers to the features, traits or nature of a dataset, its completeness and other factors, thereby rendering an objective description of the data essential. This description is usually found in the metadata. As such, they are essential not only for providing help in choosing which dataset is the most suitable for an application, but also for supporting and facilitating the sharing of data.

To ensure compatibility and easy data exchange, standards for such metadata have been established. One of the most commonly used ones is the Dublin Simple Core Metadata Standard (ISO 15836) and its set of 15 elements. These metadata elements comprise *inter alia* descriptions of the data creator, the publisher, the data coverage or the data type. Please refer to the Usage Guide at http://dublincore.org/documents/usageguide/, provided by the Dublin Core Metadata Initiative (2005) for more detailed information.

Zargar and Devillers (2009: 2) point out, however, that there are several challenges and problems associated with metadata as a means of quality assessment: Many data users either do not bother to consult the metadata in the first place, or are unable to track down the desired information in the metadata files. Additionally, users are often overwhelmed by the metadata’s technical terminology, thereby rendering the information contained useless. This lack of embracement on behalf of some data producers, distributors and users should, however, not keep anyone from establishing such valuable metadata.

3.1.4 Geolinguistic data and data sources

At the basis of my investigation lies the assumption that although geolinguistic data *can* be used in GISystems and Science, several aspects of these data must require special awareness which I aim to identify. Other disciplines such as meteorology can rely on automated data collection where for instance the temperature or wind speed at a certain location is digitally recorded at set intervals. These methods of data collection obviously have their own drawbacks, yet geolinguistic research has to rely on other sources which may be considered even less ‘reliable’. Data collection in linguistics is mostly done by field linguists who do not focus on recording the exact location of languages. They use surveys, questionnaires or interviews, conducted in written form or orally, as well as observation as methods of data acquisition. This of course has serious shortcomings with regard to reliability, let alone accuracy and precision from a GIS perspective.

Veselinova and Booza (2009: 4-5) show awareness of the problems associated with the use of census data as a resource for investigations into language use. They discuss that although census data provide both the home location as well as languages used by the census respondents, this kind of data fail to take into account factors such as prestige languages, which may influence the respondents’ answers and subsequently the data (see also Pienemann and Keßler 2007: 252). Despite additional problems such as the inexact and incorrect classification of languages (2009: 4-5) reducing the reliability of the census data used, Veselinova and Booza still deem the data ‘fit’ for their purpose of investigating which languages are spoken across the urban area of Detroit. The use of the census data is motivated by the absence of alternative data.

This lack of reliable data and data sources is one of the key problems encountered in mapping lesser-used languages, resulting in geolinguistic applications frequently being developed using whatever data are available. In addition it seems that – apart from very few exceptions such as Barni (2006) – little or no attention is put into considering which methods were applied during geolinguistic data collection, digitisation and interpretation and how these may affect the output.

Parker and Cool (2008: 2) claim that the problem of successfully integrating linguistic data into a GIS has so far been the task of having to draw information from numerous sources and hence of possessing specialised expertise. The LL-MAP project, in which

Parker and Cool participate, is similar to the GISLI project in that it combines language data and non-linguistic information in a GISystem and offers a free Web Map Service (WMS) and Web Feature Service (WFS) to its users. These kinds of services allow users to access maps or parts thereof as well as features from remote databases via the internet (for further information about Web Services please see Doyle and Reed 2001).

Although the LL-MAP project allows, and is designed for, datasets to be added, there is no mention of any data quality or compatibility issues. This is exactly the kind of project where certain guidelines for which data to use, or rather what quality issues to be aware of and how to best describe the data in metadata, would be essential. Drawing data from numerous sources is a necessity for most GIS applications, not just for an undertaking such as the GISLI or the LL-MAP projects.

It is essential to note that ‘pure’ geolinguistic data conforming to the three dimensions of space, time and theme contain information about *what* language or (sub-)dialect is used *where* and *when*. These data are usually not investigated without their context, be it political, ethnic, geographical, social, economic or religious. In dialectology, to name only one example, geographical features such as large rivers or mountain ranges acting as dialect boundaries are commonly established concepts, showing that language variation across space can indeed be partially explained by non-linguistic information. However, the application at hand merely looks at visualising the locations of languages without aiming to explain certain linguistic phenomena, as this would go beyond the scope of this thesis.

3.2 Methods applied in sample application

To arrive at the desired output of any GIS application, be it maps, prediction modelling or other goals, certain well-founded and established steps are recommended. Looking at previous approaches to mapping languages I have come across Fukushima (2008) pointing out that “there are four steps in the process of map-making using a computer”:

1) Electronic data production[1],

2) Sorting and mapping data,

[1] Author’s note: such electronic data production includes for instance the digitisation of analogue maps.

3) Comparing, integrating, superimposing, and linking data, and

4) Publishing linguistic maps

From a GIS perspective, these steps seem rather simplified and in ignorance of fundamental concepts and methods such as the conceptual design of an application, the selection of the appropriate hardware and software, etc. Setting up a GIS project, even if it is 'merely' for "map-making", as described by Fukushima above, involves a far more complex approach – no matter what the application's subject matter may be. Hence linguistic maps would also benefit from the modelling and design stages which are standard in GIS applications in other fields.

Figure 6 illustrates such an approach as suggested by the National Center for Geographic Information and Analysis's (NCGIA) "GIS Development Guide" (n.d., online). Although this guide is mainly targeted at managers in local government implementing a GIS project, I believe it to be a very useful resource and accessible even to GIS novices. It offers a far more comprehensive and better founded approach from a GIS point of view than Fukushima's. Figure 7 shows an example of how a GIS development process can be laid out, bearing in mind that one should first identify *what* a GIS should do and then *how* this objective can be implemented.

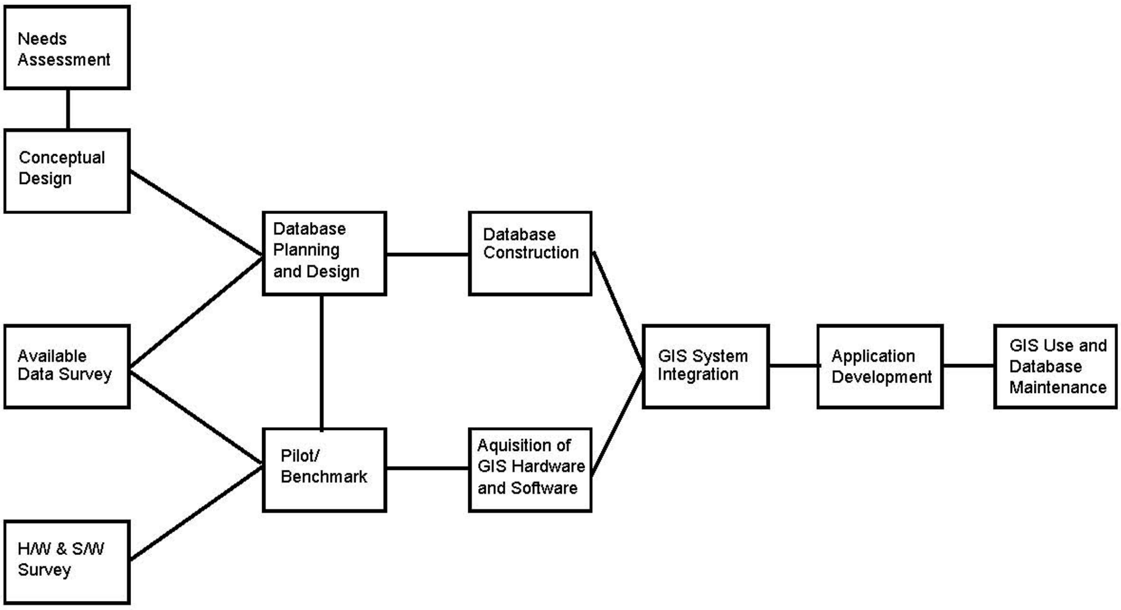


Figure 7 GIS development process (source: GIS Development Guide, n.d.: 11)

It has to be pointed out, however, that these steps are not a recipe which guarantees a successfully functioning GIS at the final stage. The nature of an application, factors such as time, human resources and money allocated, etc. are to be considered in the process of creating a GIS and may alter the layout of a process quite considerably. However, I believe that Figure 7 illustrates the core aspects in the development of a GIS and certainly points out essential stages which Fukushima failed to mention but which any GIS user – no matter how much experience they may have – needs to consider.

For the purpose of the application at hand, which does not focus on a comprehensive step-by-step development of a new GIS but on data quality, I have based my terminology and approach on Longley et al. (2005). The development of my application is conducted as illustrated in Figure 8:

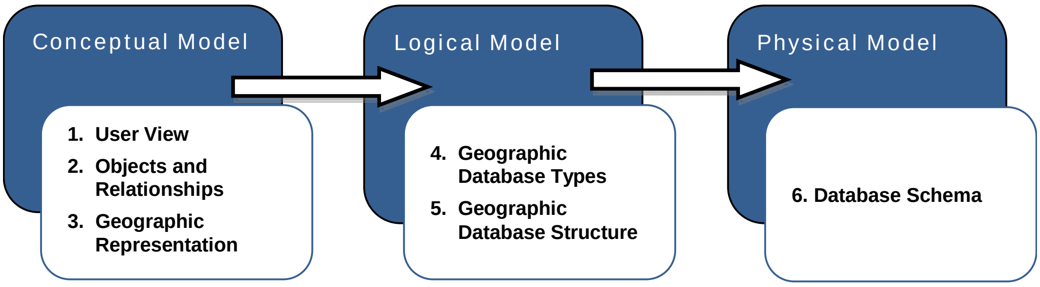

Conceptual Model
1. User View
2. Objects and Relationships
3. Geographic Representation
Logical Model
4. Geographic Database Types
5. Geographic Database Structure
Physical Model
6. Database Schema

Figure 8 Stages in database design (adopted from Longley et al. 2005: 229)

Accordingly, the first stage at the conceptual modelling level is the modelling of the User View which requires the clear identification of the desired GIS functions and its output, the objects to be represented as well as the establishment of which data are required to arrive at these objectives (this is also often referred to as a 'needs assessment', as for instance in the GIS development process in Figure 7). For the application at hand, the desired GIS output is the visualisation of locations at which the sample languages are used, which is then made available online as part of a WMS. The data required to arrive at this output need to include information about language classification, settlement names, dialect variation at the locations, etc. (see subsection 4.1.1 for a detailed description of the User View).

The second stage in Figure 8 ("Objects and Relationships") builds on the developed User View and focuses on the identification of what the objects of interest (entities) are and what the relationship between the entities is. I implement the data model in one of the most commonly used database models: in a relational model. The term relational database was coined by Codd (1970) and describes a database model consisting of tables (i.e. relations), into which the data and their relationships are organised and which allow the information to be linked. Each table consists of rows (also called 'tuples') representing an *entity* and of columns representing an *attribute*. The key terms *entity*, *attribute* and *relationship* are defined as shown in Table 1:

Table 1 Elements of an Entity-Relationship Diagram

Term	Definition
Entity	A real-world object that is distinguishable from other objects (noun)
Attribute	An entity's property
Relationship	Description of how entities relate to each other (verb)

For the sample application used in this thesis, *entities* would be Chakali, Safaliba, Ambakich, Uyajitaya and Sam, as they are instances of the *entity type* 'Language'. Each entity type forms a table. The *attributes* describe the entities, for instance 'language classification' or 'number of speakers' for each of the languages mentioned above. The *relationship* between two entities of different entity types describes their connection. For example, the language Chakali is used at a settlement called Ducie in Ghana. Chakali is an entity of the entity type 'Language', while Ducie is an entity of the entity type 'Populated Place' ('PPL'). Therefore, the entity 'Chakali' *is used* (= relationship) at the entity of Ducie. In the sample application, I make use of an Entity-Relationship Diagram (ERD) to illustrate my model, based on a technique proposed by Chen (1976). For further information about relational databases and ERDs please refer to, for example, chapters 2 and 3 in Sumathi and Esakkirajan (2007).

In step three of the conceptual modelling stage ("Geographic Representation") shown in Figure 8 above, I examine how best to represent the data for my chosen application. Generally, data in GIS can be digitally represented either in vector or raster format. The latter represents information in a grid structure in which each cell is allocated a certain property or attribute. In vector format, on the other hand, the basic geometric primitives in which spatial information is stored are points, lines and polygons (see subsection 2.1.2. in Yeung 1998, online, for further information). Languages, including lesser-used languages, have traditionally been represented as polygons, i.e. as areas with well-defined borders (Dahl and Veselinova 2005: 1-4). An alternative to this approach is the representation of locations on the settlement level as point features, which is the method I used in the sample application (for a detailed discussion see subsection 4.1.3).

The conceptual model's three stages of (1) User View, (2) Objects and Relationships and (3) Geographic Representation as shown in Figure 8 form a core part of my investigation. They are dealt with in more detail than the other two database design stages as they are most relevant to the issues of data quality, data sources and uncertainty of data other than those more commonly used.

The logical modelling stage consists of the definition of database types (step 4) and the specification of the database structure (relational tables in which each line represents one entity and the attributes are recorded in columns), while the physical model in stage 6 ("Database Schema") is the implementation of the database using Structured Query Language (SQL) as the database language. For detailed information about SQL, please refer to chapter 4 in Sumathi and Esakkirajan (2007: 111-213).

An available data survey assesses which of the needed information identified in the conceptual design is in fact available as data, from which sources they may be acquired, and to which degree they will allow the GIS to reach the desired output. The assessment of the data available against the conceptual design of my application and its intended output considers factors of data quality with regard to which extent and how they conform to aspects of the International Organization for Standardization's (ISO) standard 19113 Geographic information – Quality principles (2009: 50-51). Table 2 illustrates the main data quality elements laid out in ISO 19113:

Table 2 Core elements of data quality according to ISO 11913 and their description (adapted from ISO/DIS 19113 draft 2001: 6)

Data quality elements	Description
Completeness	Presence and absence of features, their attributes and relationships (commission vs. omission)
Logical consistency	Degree of adherence to logical rules of data structure, attribution and relationships
Positional accuracy	Accuracy of the position of features
Temporal accuracy	Accuracy of the temporal attributes and temporal relationships of features
Thematic accuracy	Accuracy of quantitative attributes and the correctness of non-quantitative attributes and of the classifications of features and their relationships

The elements most relevant to the application at hand – *completeness*, *positional accuracy* and *thematic accuracy* – are considered in the Sample Application section and discussed in detail in the Results and Analysis chapter. I would like to stress that positional accuracy, which one may consider the most important quality concern when mapping certain phenomena, is only part of a bigger picture of data quality issues.

The settlements at which the respective languages are used will be geocoded, i.e. they will be referenced using coordinate values. GPS coordinates were only recorded in some of the SIL surveys. In cases where the fieldworkers did not record the GPS data but merely included the settlement name, the National Geospatial-Intelligence Agency's (NGA's) GEOnet Names Server (GNS; available online at http://earth-info.nga.mil/gns/html/index.html), a database of geographic names, acts as a reference. This database provides geospatial and toponymic information, thereby allowing the assignment of coordinates to the settlements. Moreover, it provides alternate names for the settlements, which are also recorded in the application's database. An alternative to the GNS is its Russian counterpart Poehali, which is available at http://poehali.org/ (2002-2009). In case of discrepancies between the NGA's recorded data and the position provided in the SIL survey reports, preference is given to the SIL data.

3.3 Software used

To implement my sample application in a GIS, I chose two of the most commonly used software tools whose manuals provide comprehensive, accessible information even for beginners. The database management system used is MySQL 5.1 by Sun Microsystems, Inc., which can be downloaded free of charge (available at http://dev.mysql.com/downloads/). The GIS software is ESRI's ArcGIS 9.3. For more information about these two software products please refer to chapter 1 of the MySQL 5.1 Reference Manual (Sun Microsystems, Inc. 2008-2009) and ESRI (2001-2008) respectively.

4. Sample Application

The following chapter details the stages in the development process of my GIS application with a special focus on the quality of data available and how and where uncertainty is introduced.

4.1 Development of conceptual model

4.1.1 User view

The first step in this GIS application is to determine and define the desired output. The objective of the GIS application at hand is to provide a new dataset for a WMS displaying the location of where the languages Chakali, Safaliba, Uyajitaya, Ambakich and Sam are currently used. Beside the basic spatial information as to *what* investigated language is used *where*, the map needs to contain spatial and thematic information about the languages investigated such as:

- Language classification
- Language code (ISO 639-3)
- Dialects/sub-dialects
- Name of settlements at which language/dialect is used (including alternate names)
- Number of speakers of each language
- Coordinates of locations/settlements

Modelling the use of a language effectively means modelling the location of people who make use of a certain language as a human activity (the focus here is on oral and if applicable written language use, and not merely on written use). However, as humans are mobile and cannot be tracked individually for practical and ethical reasons, their location is modelled based on their reported residence.

The target users may have a variety of potential backgrounds such as in linguistics, history, anthropology, geography, but may also be interested laymen. The layer structure of a GIS and the distribution as an online map service allows users to display information of interest as required.

The WMS should moreover display the current geolinguistic situation rather than show the language distribution at a distant point in the past, therefore requiring data as up-to-date as possible. It is also worth noting that the application at hand does not consider issues such as in which context the language is used, the proficiency of its speakers or whether it is part of a multilingual environment – it is the fact *that* a language is used at a certain location and not *how, by whom, how often* or *when* it is used which is important. Such – from a linguistic point of view clearly important issues – could be included in a more comprehensive and complex map service.

Turning to the requirements which the data need to fulfil, it is essential to consider how *accurate*, how *precise* and how *complete* they have to be to arrive at the desired output. In the case at hand, I set the ideal requirements as follows:

Table 3 Geolinguistic data requirements for sample applications with regard to accuracy, precision and completeness

Accuracy	Essential with regard to whether or not a language is used (yes/no); numbers of speakers should represent actual numbers where available; information such as language family or dialects needs to be accurate; temporal accuracy not as important (number of speakers will not change drastically over the course of hours, but rather over decades); positional accuracy required with regard to settlement names
Precision	Positional: not too critical as the decision was made to map the occurrence of languages at the locations of reported settlement – this location (coordinates) can be looked up; attribute precision should include for instance number of speakers but does not have to go beyond this (for instance no ratio of male/female speakers is needed in the case at hand)
Completeness	Important but not viable – at least all “officially known” locations where languages are used should be included (sensitive issue)

In practice, the situation of limited data availability and general viability of mapping lesser-used languages requires me to compromise on some aspects. Issues such as obtaining accurate demographic data of language users is not feasible, bearing in mind that the data have to be readily available and considering the temporal and financial scope of this project. Obtaining a spatially, temporally and thematically complete, accurate and precise dataset of where a specific language is used is in fact impossible, as the movement of speakers at this scale is unpredictable – what if, put bluntly, two speakers of Safaliba were to emigrate to New York for several years, use the language while in the US, and then return to Ghana or move elsewhere? Such precision would go beyond the purpose of the application at hand, yet the limitation of not being able to track it in the model needs to be mentioned.

4.1.2 Objects and relationships

Due to the lack of data sources other than the SIL reports, and the subsequent inability to compare data and attempt to infer values, I decided to call the main entity type of interest "EthnologueInfo". In this I followed Veselinova's approach to date in order to illustrate to the user which source was used. Consequently (and in contrast to Veselinova), I have also named the Dialect entity type "EthnologueDialect", as it again represents the classification of dialects as reported in the Ethnologue. The entity type "PPL" stands for Populated Place, i.e. the settlements at which a language is reported to be spoken. The relationships between the entities are also described as illustrated in the ERD below:

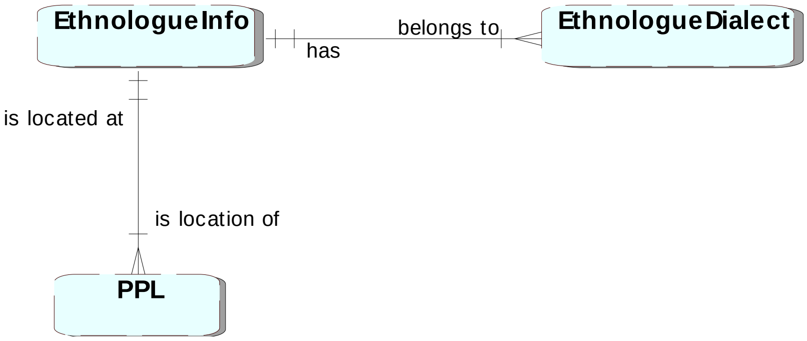


Figure 9 Entity-Relationship Diagram used in sample application

4.1.3 Geographic representation

When considering how the entities of interest are to be represented in the GIS, the key question is how to effectively represent the occurrence of use of a certain language at all. Examining previous geolinguistic GIS applications, one can see how varied the outputs can be, even if the original conceptional goal – for instance to map the location of certain lesser-used languages – is identical. Dahl and Veselinova (2005: 1-5) describe current language mapping as being dominated by the tradition of representing languages as polygons, a view which was cemented by my own research for this thesis.

The decision to represent the location as polygons was taken in both the LL-MAP project, as well as the Global Mapping International's (GMI) World Language Mapping System (http://www.gmi.org/wlms/). In contrast, the developers of UNESCO's Interactive Atlas of the World's Languages in Danger and the GISLI project decided to display the locations as point feature classes. The latter two applications differ again in that UNESCO's choice of representation was to map individual points for each language, calculated to represent the relative centre of the region in which a language was used, or to represent the relative centre where the most speakers were situated. GISLI on the other hand maps the language locations at the settlement level, thus providing a more detailed illustration. Representing the occurrence of a certain language as point features is only viable for lesser-used languages at a relatively small scale for reasons of clarity.

A visual comparison of the different approaches in the LL-MAP project and the UNESCO Interactive Atlas can be seen in Figure 10 and Figure 11, showing the contrast of how some of the languages in Iran were mapped using polygons forming a mostly continuous surface, and points respectively. Figure 10 largely generalises the linguistic situation in Iran by only showing the more dominant languages:

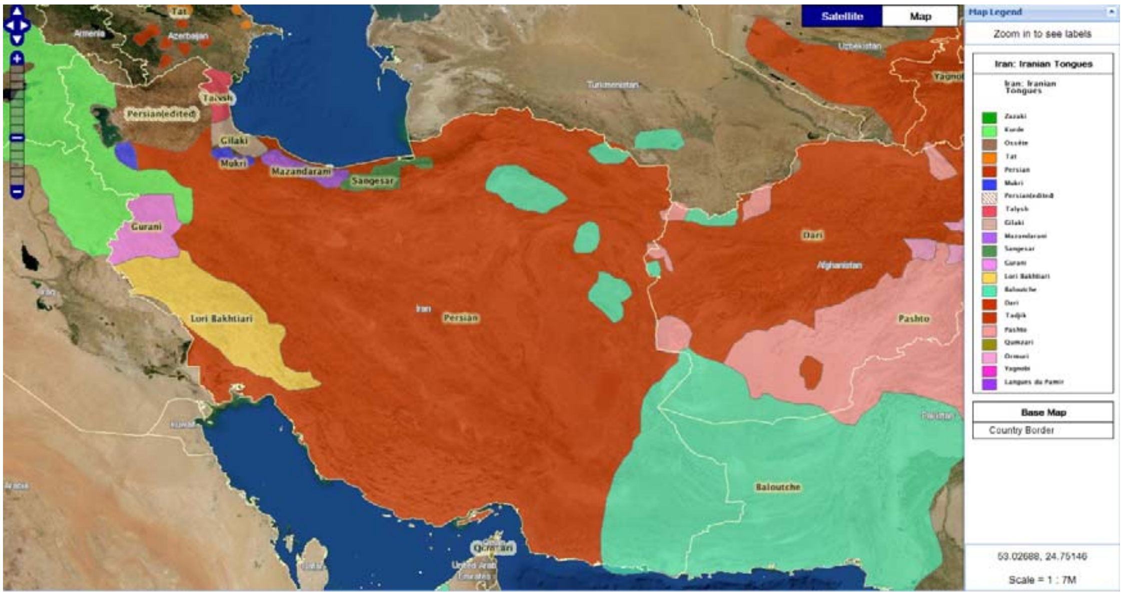
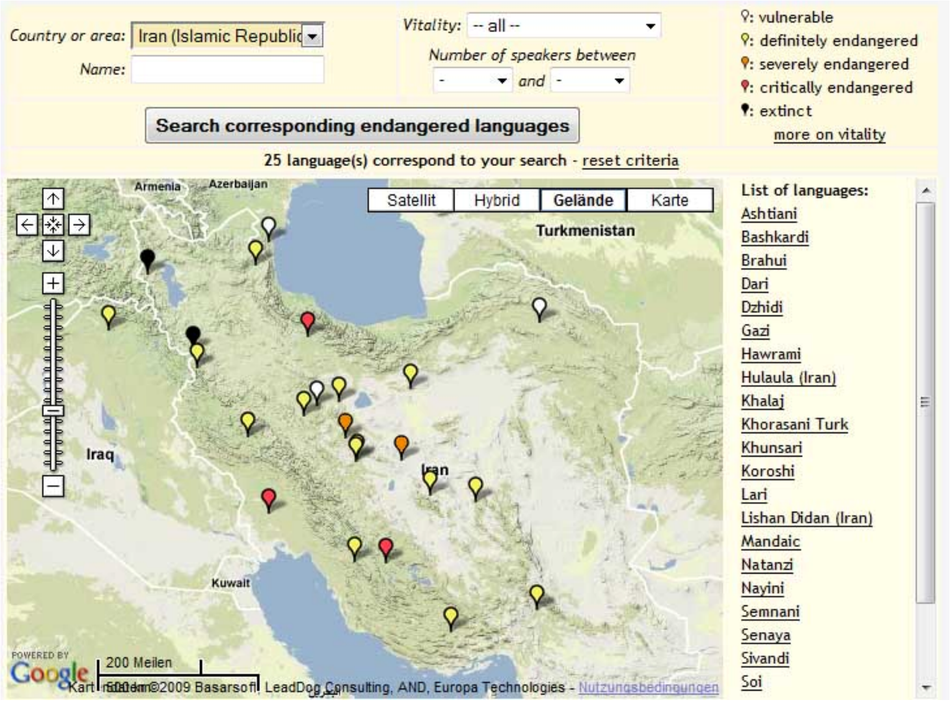


Figure 10 Excerpt of LL-MAP (accessed September 2009)



Country or area: Iran (Islamic Republic

Name:

Vitality: -- all --

Number of speakers between - and -

9: vulnerable
9: definitely endangered
9: severely endangered
9: critically endangered
9: extinct
more on vitality

25 language(s) correspond to your search - reset criteria

List of languages:
Ashtiani
Bashkardi
Brahui
Dari
Dzhidi
Gazi
Hawrami
Hulaula (Iran)
Khalaj
Khorasani Turk
Khunsari
Koroshi
Lari
Lishan Didan (Iran)
Mandaic
Natanzi
Nayini
Semnani
Senaya
Sivandi
Soi

Search corresponding endangered languages
Satellit
Hybrid
Gelände
Karte
Armenia
Azerbaijan
Turkmenistan
Iraq
Iran
Kuwait
POWERED BY Google
200 Meilen
500 km
Kartendaten ©2009 Basarsoft, LeadDog Consulting, AND, Europa Technologies - Nutzungsbedingungen

Figure 11 Excerpt of UNESCO Interactive Atlas of the World's Languages in Danger (accessed September 2009)

The example above shows that the results of the initial conceptual development of a GIS have an immense effect on its output or 'final product', as well as its overall abilities. The extract of the LL-MAP aims to provide an overview of the linguistic situation in Iran and gives the impression that languages are actually used as a continuous surface across all of Iran, uninterrupted by regions which are uninhabited or by locations at which lesser-used languages are spoken. The UNESCO map in Figure 11, on the other hand, has a specialised rather than generalised approach in that the location of endangered languages in the area are plotted, providing no information about languages spoken by the majority of residents.

Displaying language locations as discrete point features at the settlement level is certainly a compromise too, particularly bearing in mind that a language is as mobile as its users. However, not even a polygon representation could accommodate for speakers' potential movement – at this point, uncertainty is an inevitable problem. In cases of lesser-used languages, the boundaries of where a language is used and where not simply cannot be as clearly defined as is commonly done by, for instance, adhering to political and administrative borders. However, the languages are not examined at a very large scale, which in other applications would justify the representation as polygons. Hence, I continue GISLI's approach of representing lesser-used languages as points at the settlement level, rather than as the more widely used polygons.

To sum up, the GIS output needs to meet the following requirements from a conceptual point:

- ✓ Display the location of a set of lesser-used languages at the settlement level as points
- ✓ Provide information about the name of the settlement at which a language is used
- ✓ Provide information about the number of speakers
- ✓ Provide information about dialects and their locations
- ✓ Provide the users with additional information such as topography or political boundaries
- ✓ Serve users with varied (academic) backgrounds
- ✓ Be logical and easily comprehensible to the user
- ✓ Allow users to browse and select which layers to display
- ✓ Be accessible online and free of charge

The data used in the GIS should ideally:

- ✓ Give the location of where a language/dialect is used on a settlement level as point data (coordinates or settlement name from which coordinates can be inferred)
- ✓ Provide information about the language family and potential dialects
- ✓ Be as current as possible
- ✓ Provide information about the number of speakers of each language/dialect
- ✓ Be as independent of any political or ethnical influences or bias as possible
- ✓ Be comprehensively documented

4.2 Logical and physical model

The next step in the design of the database was to list the entity types' attributes:

Table 4 Attribute list of ER-Model for sample application

EthnologueInfo
Language name
Ethnologue code
Language classification
Estimated number of language users

EthnologueDialect
Dialect name
Dialect code

PPL
PPL name
Alternate PPL names
Latitude
Longitude

The resulting Entity-Attribute-Relationship Diagram including primary and foreign keys used to link the entities, as well as the defined data types, is as follows:

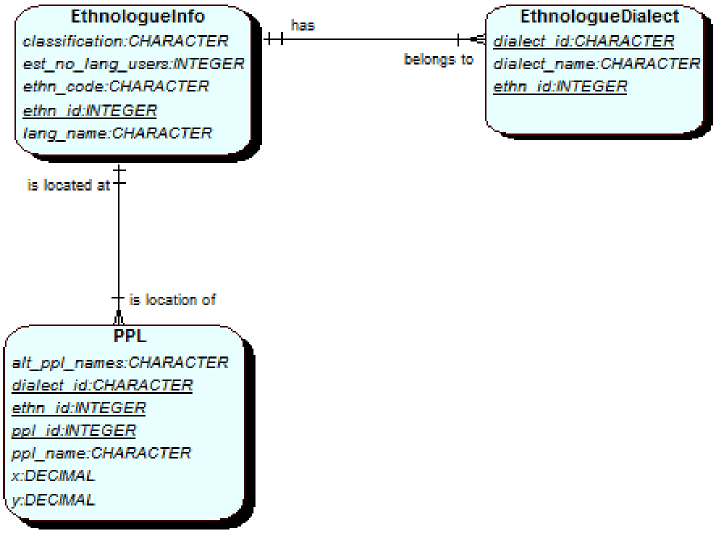


Figure 12 EAR-Diagram of sample application

4.3 Available data and data sources

My main data sources are the SIL International sociolinguistic survey reports on the languages under investigation. The reports provide comprehensive information gathered by trained linguists on site and are commonly used as reference material in the linguistic community. SIL International has formal consultative status with UNESCO and was recommended to me as a data source by Dr Veselinova.

Given that the languages under investigation have fewer than 7,000 speakers each, not much research has been conducted on them apart from the SIL surveys. My search for alternative data sources for the languages under investigation showed that apart from Ethnologue, no other sources seem to have collected data, or have made these data publicly available. I included a search for alternate names by which some of the languages are referred to (e.g. Uyaji, Duduela, Yabatia, Xuyadzitaya and Koki for Uyajitaya, as reported by Lambrecht et al. 2008: 5), but to no avail. At the time of writing (October 2009), there appears to be no current fieldwork being undertaken on Safaliba, Ambakich and Uyajitaya. I was able to identify three researchers working on Chakali and Sam respectively, but am still awaiting their response to my queries. This means that for the application at hand, relying on the SIL data sources has to suffice.

The reports published by SIL are sociolinguistic surveys whose primary purpose was not to determine the accurate location of where a certain language or dialect was spoken. They are indeed characteristic for the materials from which linguists infer locations of languages with very few users. The surveys examined *inter alia* the contextual use of the languages and their vitality. They are similarly structured and mention the locations at which the languages are spoken as well as the dialectal situation in introductory chapters. Table 5 lists the surveys including their year of investigation and the methods by which the data were collected:

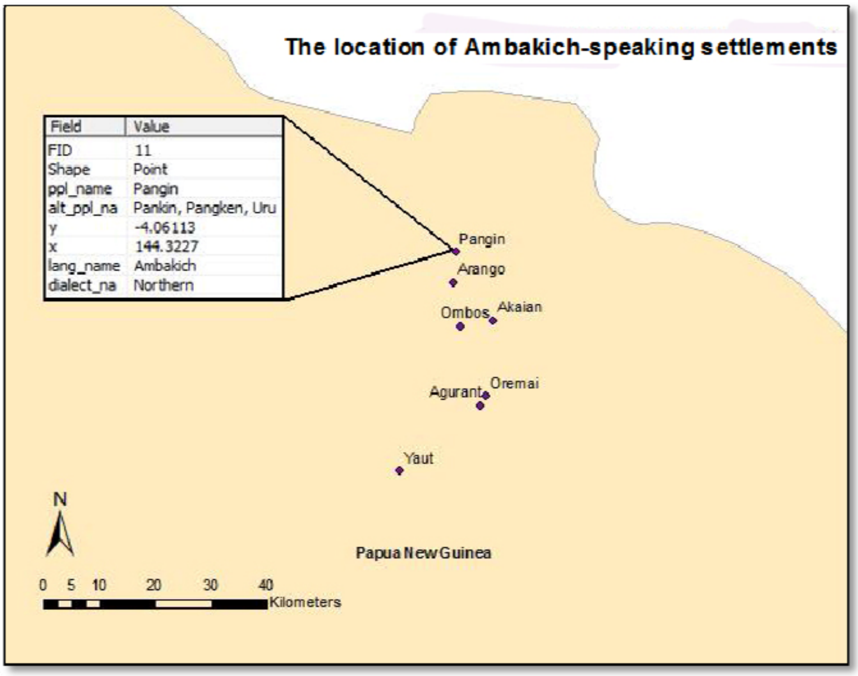
Table 5 Data source documents used in sample application

Language	Year surveyed	Source document	Data collection method	Reference
Chakali	1995	Socioling. survey	Interviews, questionnaires	Tompkins, Hatfield and Kluge 2002
Safaliba	1995	Socioling. survey	Interviews, questionnaires	Kluge and Hatfield 2002
Ambakich	2003	Socioling. survey	Observation, group interviews, questionnaires	Potter, Lambrecht, Alemán and Janzen 2008
Uyajitaya	2003	Socioling. survey	Interviews, questionnaires, observation Some GPS locations recorded during survey in November 2003	Lambrecht, Kassell, Potter and Tucker 2008
Sam	2001	Socioling. survey	Interviews, questionnaires	Rueck and Jore 2003

The information extracted from the SIL survey reports relevant to this application is available for reference in Appendix A. As can be seen from the dates of investigation and publication in Table 5, the data do not reflect the most current status. For instance the Ghana surveys (Chakali and Safaliba), although published in 2002, report findings from 1995.

4.4 Sample application map and assessment of data quality

After creating the database and mapping the data in ArcGIS using the reference coordinate system World Geodetic System 1984 (WGS 84) and a simple base map, the results are as shown in an example in Figure 13:



Field	Value
FID	11
Shape	Point
ppl_name	Pangin
alt_ppl_na	Pankin, Pangken, Uru
y	-4.06113
x	144.3227
lang_name	Ambakich
dialect_na	Northern

Figure 13 Sample of mapping application: The location of Ambakich-speaking settlements

This map does not yet contain additional layers which would illustrate the language occurrence in, for instance, its topographical context, but provides the geocoded locations and information such as the language and dialect.

Regarding the spatial aspect of the data, the general location is given in the data sources both as a verbal description as well as in the visual form of analogue maps. Some of the reports also provide GPS locations recorded by the fieldworkers on site. Other sources include, for instance, the available map displaying the location of Safaliba users shown in Figure 14, which is not satisfactory by any current geolinguistic standards but merely serves for orientation:

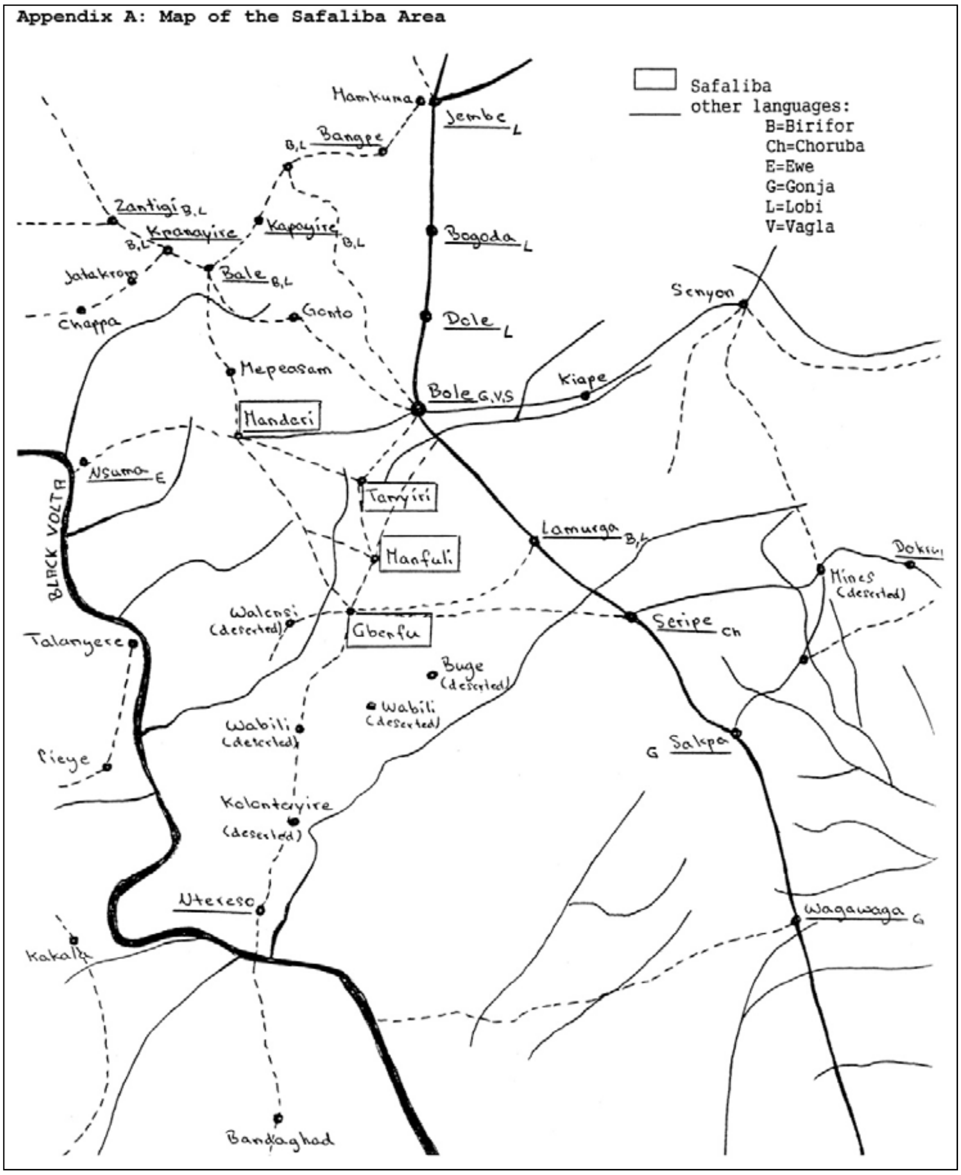


Figure 14 Map of Safaliba Area (source: Kluge and Hatfield 2002, Appendix A)

The assessment of the available data described above against the three – for this application – most essential ISO 19113 data quality elements include the element of *completeness, positional accuracy* and *thematic accuracy*. The elements were described in Table 2 and are illustrated in Figure 15:

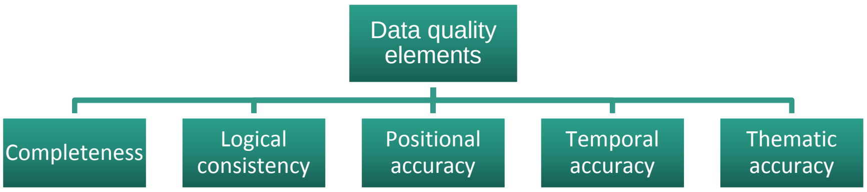


Figure 15 Data quality elements according to ISO 11913 (2009: 50)

- **Completeness**

The assumption that the data from the SIL reports is complete is not realistic. This is not only due to the fact that the purpose of the data collection was different from the use to which the data are put, but lies in the very nature of language and the impossibility of constantly tracking every speaker of a certain language. Moreover, determining the location of all reported settlements proved unfeasible within the scope of this thesis, as the coordinates for the Chakali speaking village of Tissa were not provided in the report and could not be found in any of the most commonly used gazetteers. Enquiries to the Ghanaian Ministry of Lands, Forestry and Mines as well as to staff at the Centre for Remote Sensing & GIS at the University of Ghana have not been answered to date. As such, the data are certainly incomplete with regard to the spatial aspect.

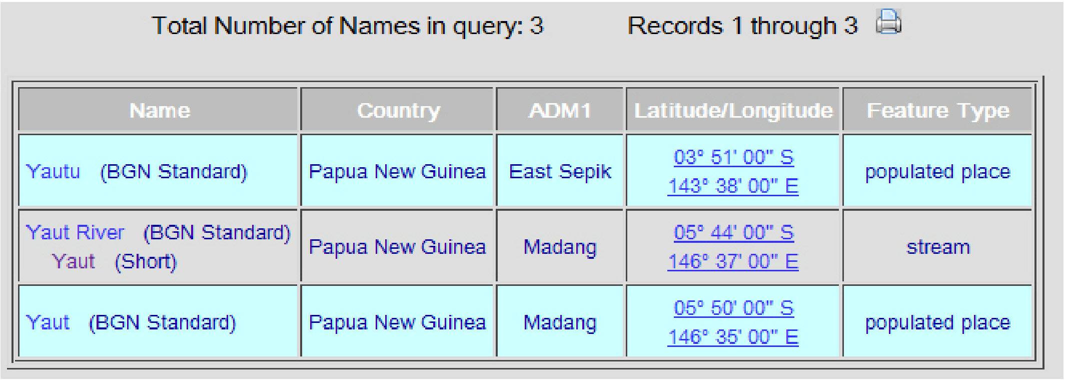
Issues such as thematic completeness cannot be verified due to a lack of appropriate reference materials. I have, for instance, supplemented the “Alternate PPL Names” attribute with the name variants I found on the GNS. Yet as some of the settlements are not contained in this database in the first place, their alternate names – if applicable – will subsequently also be missing from the dataset.

Additionally, there is the problem of unreliable data sources on which the SIL reports themselves are based: Lambrecht et al. (2008: 10), for instance, report being given conflicting information by local residents as to whether the Uyajitaya village of Jio is in fact still inhabited. Representing Jio as a location of Uyajitaya speakers when in fact it is deserted would result in commission, i.e. excess data; conversely, excluding Jio from the dataset although it is still inhabited by Uyajitaya speakers would result in omission, i.e. absent data. Both cases would therefore affect the dataset’s completeness as understood in ISO 11913.

- **Positional accuracy**

The issue of positional accuracy in terms of distance measurement in the application at hand is not as essential as compared to other GIS applications such as cadastral registration or construction. The application is based on reported settlements, providing either their name or the recorded GPS information. The coordinates available from the NGA's GNS are, however, not very precise and only provide the degrees and minutes, but not the seconds of latitude and longitude of any location.

In cases where there is a discrepancy between the GPS locations reported by the SIL fieldworkers and the locations provided by the NGA's GNS, preference was given to the SIL data. This choice was made because some of the locations were not contained in the GNS database. One such example was the Ambakich speaking village of Yaut, situated in Papua New Guinea's East Sepik region. The GNS query set to search for 'Yaut' in Papua New Guinea yielded three results as shown in Figure 16. Two refer to a river and populated place in Madang province; the third 'Yautu' is indeed a populated place in East Sepik province, yet the coordinates given do not correspond to the location of Yaut recorded by the Ambakich fieldworkers (4º24.73'S 144º13.992'E).



Total Number of Names in query: 3   Records 1 through 3

Name	Country	ADM1	Latitude/Longitude	Feature Type
Yautu (BGN Standard)	Papua New Guinea	East Sepik	03° 51' 00" S 143° 38' 00" E	populated place
Yaut River (BGN Standard) Yaut (Short)	Papua New Guinea	Madang	05° 44' 00" S 146° 37' 00" E	stream
Yaut (BGN Standard)	Papua New Guinea	Madang	05° 50' 00" S 146° 35' 00" E	populated place

Figure 16 Result of GNS query "Yaut" in Papua New Guinea, Fuzzy Search

- **Thematic accuracy**

Determining the thematic accuracy of my sample data is difficult because of a scarcity of published reference data, even for attribute data such as dialect classifications. Additionally, not all the locations were in fact visited by the SIL researchers themselves for confirmation, but rather they relied on reports from residents of other villages. This means that initial data capture is not very reliable and likely to be biased, yet it sufficed for SIL's purposes.

The only other data source I was able to locate for reference was Jonathan A. Brindle, who is currently investigating Chakali in his PhD thesis. Asked for his assessment of present day use of Chakali, both with regard to numbers and speakers, he claimed the following: the number of residents in Ducie, reported at an estimated 2,800 by Tompkins et al. (2002: 5), was apparently down to half that number by 2008. In the same year, Katua and Bele (Gulumbele) had twice as many inhabitants than reported in Tompkins et al. (Brindle, Email, October 12, 2009). This discrepancy may be due to Tompkins et al. inferring their estimates from 1984 census data adding an annual growth rate of 2.3% to arrive at the estimated numbers for 1995. I have however been unable to verify Brindle's claims as I am still awaiting information about his data sources and exact numbers, which is why I did not include his data in the application.

5. Results and Analysis

This chapter looks back at the questions I initially set in this thesis and provides the results. The central questions were:

- Does the use of geolinguistic data sources and their data quality pose special challenges to GIS developers and users?
- What are these challenges and problems with regard to mapping lesser-used languages and linguistic GIS applications in general?
- Where are elements of uncertainty introduced into geolinguistic data in GIS applications?

At first glance it may seem rather straightforward to map a certain language occurring at a certain location. Yet although linguistic data sources attempt to grasp and comprise the complex nature of language and its use – as well as to do the perception of language and identity justice – compromises as to the precision, completeness and accuracy of the data from a GIS perceptive are inevitably made. The answer to my first question regarding whether or not geolinguistic data and data sources pose challenges to GIS in terms of their quality is therefore a clear yes. These challenges vary from one application to another, depending on factors such as the desired output, available project funding, or the available data, yet I believe that the following issues – in answer to my second and third question – should be considered in any geolinguistic GIS application.

5.1 Positional and thematic accuracy

Contrary to applications such as for instance mapping each restaurant within a certain part of town, collecting each instance of where a language is used is very difficult – or indeed impossible due to the movement of speakers. Subsequently, the mobility of a language is as inherent as that of its speakers. Mapping each instance of a certain language is therefore impossible and has to be largely generalised.

The application at hand mapped the occurrence of lesser-used languages at the settlement level, meaning that the concept of this application relies on settlement names being reported accurately in order to reference their coordinates. Problems which arise with this is the common practice of villages being referred to by several names, as well as several villages being referred to by the same name (see example of Yaut above) – and subsequently locating the coordinates for the locations in gazetteers or other sources. For instance, the Uyajitaya village of Subalulu is, according to Lambrecht et al. (2008: 8), also known as Tapo, Sari and Balulu. Moreover, the report refers to the village of Buai (alternate name Tagoe), which is more commonly known as Baui.

While the above issue can be resolved with research in gazetteers, factors such as the settings of the GPS device used to record the locations as done for instance for some settlements in Lambrecht et al. (2008) may affect positional accuracy. However, information about which GPS devices were used, what the settings were and if there was any post-processing are not given, thereby making a description of the data's positional accuracy without any ancillary documentation difficult.

From a GIS perspective, this facet concerning the required documentation of the data collection process is essential. Looking at the more recent surveys I used as samples for this investigation, the fieldworkers did actually collect GPS data for the Uyajitaya, Sam and Ambakich surveys[2], but no ancillary documentation is reported. Again it is an issue of re-using data which were originally recorded for a different purpose, understandably without bearing in mind which procedures would be required for a 'textbook' GIS example.

The method of obtaining the coordinates of settlements using gazetteers may also cause such problems, due to the original GPS measurement having been possibly subject to error and unreported methods and settings. However, bearing in mind the application's purpose and scale at which the locations are mapped, I believe that positional accuracy is less of a key issue than in much of the research and discussion of positional accuracy and error in more traditional fields using GIS. While it is certainly expedient to aim at maximum positional accuracy, it is not realistic to map the occurrence of a language to 1 or 10 metres accuracy due to its speaker's mobility.

[2] I was provided with a list of these coordinates recorded during the Ambakich survey in 2003 by SIL Papua New Guinea (LR-Sociolinguistics 2009); this list is not contained in the original survey report published in 2008.

Common ways of determining the thematic accuracy of attribute data in GIS include visual checking, comparison of attributes to other data sources and field checks for independent sampling (Wyoming Geographic Information Science Center, online). For the example at hand, these methods were not always an option for several reasons: Visually checking the plotted phenomena of language use would not have revealed information about their thematic accuracy, particularly as the thematic data are mainly qualitative. Finding other data sources to compare the data with would be viable in the case of comparing language classifications, dialects or numbers of speakers. However, this also proves difficult due to the non-existent documentation beyond the SIL reports and the resulting lack of reference data. Although I did find a data source in Jonathan Brindle, his claims about the Chakali language have not been verified, and were therefore not considered a viable data source. Independent sampling in field checks was not an option for practical and financial reasons.

The options for checking thematic information such as the number of speakers of a certain language in countries such as Ghana or Papua New Guinea are certainly limited. Census data are only partially useful for linguistic research for reasons mentioned above. Some of the estimates in the SIL reports were done using an annual growth rate (between 2.3% for Safaliba and Chakali and 2.7%) to calculate population numbers based on censuses. This method does not take account of predominantly unpredictable complex human behaviour such as migration, or of language death despite increasing population numbers, thereby leading to inconsistent data. It should also be mentioned that the number of residents at a village assigned to a language does not necessarily represent the actual number of speakers of this language. Multilingual scenarios are, according to Brindle (Email, 2009) for instance found in Katua, where other languages such as Pasaale or Wali are also used alongside Chakali. While this does not constitute a problem regarding the fact that Chakali *is* spoken at the location of Katua, it certainly affects the reliability of the estimated numbers of speakers.

Although these problems regarding positional and thematic accuracy result in what may seem rather vague information from a GIS point of view, this kind of data are an established basis of linguistic research.

5.2 Data completeness

Data completeness cannot be linked to real world phenomena, but needs to be seen in relation to the conceptual design of the application, its levels of generalisation and abstraction. Compiling a complete dataset mapping all speakers of the languages under investigation is simply unviable; therefore, the data need to be as complete as possible with regard to the selected and generalised view of the user. This obviously also holds true for other geolinguistic applications, requiring a sound conceptual foundation against which the available data can be measured with regard to their completeness.

Yet determining the completeness of the data used in an application such as the one at hand proves challenging. Reference data and field checks, which are often applied in GIS applications of other sciences, were not an option, and relying on ancillary information given in the data source itself (i.e. the survey reports) was only of limited use. Bearing in mind the conceptual design of the application at hand, the data have to be considered incomplete for reasons such as the failure to determine the location of one of the settlements (Tissa) or the unresolved question of whether Jio is inhabited or not.

There is also another aspect of data completeness which I would like to emphasise and which applies to any linguistic GIS application, namely that of deliberate data manipulation. Language, its use and its documentation are a sensitive issue which is often burdened with expressions of political or economic control and social stigmata. These factors also need to be considered when assessing a dataset's origin and its resulting completeness. A present day example of 'disallowing' a people of their language in documentation and publication is the situation of Kurdish in for instance Turkey or Syria (cf. for instance UNDP 2004: 7; Tejel 2009: 110-113).

5.3 Geolinguistic data sources and the aspect of uncertainty

Referring back to Longley et al.'s (2005) and Brimicombe's (2007) illustration of uncertainty in GIS (Figure 4 and Figure 5), we can see that uncertainty is introduced at several stages in geolinguistic GIS applications. Initially, the concept of language as an observable event with a spatial element varies from user to user and from application to application. However, central problems such as how to define the spatial extent of an interactive phenomenon such as language or its variations will prove a challenge in any application.

The issue of ambiguity as a source of uncertainty described by Longley et al. (2005: 130-132) can undoubtedly also be located in linguistic GIS applications. As an example we may consider semantic and theoretical differences in the conceptual definition of what constitutes a ‘dialect’. Assigning some communication systems the label ‘language’ and classifying others as ‘dialect’ may be done differently among individual data producers and users, thereby leading to inconsistent data. This source of uncertainty was also referred to in Figure 6 and explained by the different universes of discourse for each data producer, user and application. Fisher et al.’s (2009) quote in subsection 2.3 also mentioned this problem of semantics with regard to data quality and uncertainty.

In vector data models, representations of spatial linguistic entities make use of the geometric primitives of points (see application at hand), lines (e.g. isoglosses) and polygons (e.g. in the LL-Map project). The choice of which type of representation works best has to be decided based on the initial concept of each application, yet all three introduce uncertainty. Both the LL-MAP’s polygons as well as the UNESCO’s point feature map shown in Figure 10 and Figure 11 respectively have their drawbacks by introducing uncertainty as to the distribution of a language. The application in this thesis used point features to represent the occurrence of a language at the settlement level, rather than transforming the points into area features or attempting to represent language use as polygons based on verbal descriptions of certain regions. Both approaches treat the occurrence and use of a language as discrete objects, yet point features of settlements do not show a language occurring at locations between these settlements. At the same time, plotting language use as an area means including areas which are in fact uninhabited, thereby adding uncertainty through a vague outline of the area. While it may appear tempting to map certain languages and the rough outlines of their extent as polygons and thereby bypass the problem of imprecise location and distribution of speakers, it still would not solve the problem of uncertainty.

Brimicombe’s first two categories of uncertainty (see Figure 5) address issues of uncertainty at the data level. At the first stage, intrinsic uncertainty is introduced into geolinguistic data during the collection process. Basing one's data collection on a clearly defined conceptual application model with regard to selection and generalisation would certainly help reduce uncertainty. However, the majority of linguistic GIS applications do not include the first-hand capture of relevant data due to reasons of practicability and cost. This frequently results in the re-use of data which were initially recorded for different purposes, reflecting a discrepancy between which dataset is available and which dataset would actually be needed to arrive at a certain output. Although this is a problem common to many GIS applications in various fields, I believe that it is particularly true of linguistic applications. In many cases, the data used are a compromise for a lack of available data – as for instance by Veselinova and Booza (2009). The intrinsic uncertainty of the original data is thereby carried across, while adding uncertainty due to the discrepancy between the conceptual design of the original data collection purpose and that of the present application. Using, for instance, data which were originally collected to investigate the distribution of certain phonological features introduces additional uncertainty and error into an application attempting to examine the boundaries of the whole dialect area.

Linguistic data including a spatial element is mostly recorded in census data and in studies with, for instance, a dialectological, typological or generally sociolinguistic background. Census data not only prove to be a valuable source for synchronic analysis, but also for research into migration and past language use (see Williams 1988: 11-12; Pryce and Williams 1988: 167-237). Yet again they are used as a compromise: they do not present data which are *fittest* for the respective use, but they present the *only* data available. Extracting the data needed to make a GIS application work from such sources presupposes the knowledge and awareness of aspects which need careful consideration. Some of these are a given for researchers with a background in linguistics, but may not be so obvious to GIS experts working in other fields. These aspects include those of social and political issues surrounding language and language use, which turn what at first may appear to be a clear-cut exercise into a far more complex undertaking.

Relying on data sources such as sociolinguistic surveys certainly has drawbacks with regard to data quality and uncertainty too. These weaknesses include the data collection process itself and involve issues of bias of the interviewer/survey compiler as well as the interviewees/test subjects who may influence survey results either subconsciously or deliberately (see also subsection 5.3). Moreover, asking the respondents to reflect on their own linguistic performance by enquiring about their habits of language use will also affect the result of a survey (see for instance Milroy and Gordon 2006: 25, 54). While these aspects may not play such a vital role in mapping language occurrence, they certainly have an influence on applications such as investigating multilingual scenarios or language variation.

To summarise these findings in a nutshell, the following aspects need to be mentioned and should be paid attention to in future geolinguistic GIS applications:

Data quality:

- Bias and influence on data during collection (language as a political/social tool)
- Issues arising from the need to re-use data and use data sources as a compromise
- Lacking or inadequate data collection documentation to establish fitness for use (no tradition of collecting ancillary information which would be useful for GIS applications)
- Lacking resources for comparison to check thematic and positional accuracy as well as completeness

Uncertainty:

- Introduced at conceptual stage due to semantic differences and due to differing universes of discourse when using secondary data
- Intrinsic and inherited uncertainty as most linguistic applications have to rely on secondary data
- Uncertainty arising from choices in representation (point vs. polygon features)

6. Conclusion

6.1 Summary

This thesis dealt with concerns about data quality in geolinguistic GIS applications. It exposed key problems of geolinguistic data being used in GIS in a sample application which led to the following conclusions:

The idea of an unbiased and unselective data collection process which results in a complete and accurate set of data can be dismissed as idealistic in any discipline. Yet in addition to aspects applicable to many more traditional GIS fields, linguistic data always have to be assessed against the background of a larger social and political picture which may be rather complex to infer. It is further hindered by the lack of ancillary information about data collection, which renders the assessment as to whether or not geolinguistic data are fit for a certain use difficult. Methods commonly used in GIS to assess qualitative data such as field checks or comparison to other, more precise datasets are in most geolinguistic cases not an option for reasons of cost and lacking reference data. Hence, other ways of tackling the problem have to be found. Geolinguistic GIS applications would thus greatly benefit from the development of a standardised adoption of data documentation to suit the specific needs of the discipline.

Basing GIS applications on secondary data is in itself of course not an impermissible practice, but rather a common one. However, the issue of introducing inherited uncertainty through the re-use of data and additional error due to different universes of discourse between data producer and user has to be addressed. This acknowledgement and subsequent dealing with the results of the assessment is still missing in current geolinguistic applications.

From the perspective of many traditional GIS fields such as land registration (cadastre), the imprecision, inaccuracy and uncertainty in much of the data that linguists have worked with for decades may seem like an insurmountable problem. This is not to say that the discussions and problems surrounding data quality in the more established fields have already been resolved. Yet it is essential to be aware that forcing geolinguistic data into GISystems without any serious considerations as to the quality of the source material *will* yield distorted results.

6.2 Discussion and future work

This thesis has only scratched the surface of what promises to be a rewarding symbiosis between the disciplines of GIS and linguistics. Unfortunately, the application of GIS in non-traditional fields has in my opinion not yet received the amount of attention it deserves, neither on the behalf of the scientific community, nor as part of GIS education. It yields great potential not only to the application of for instance spatial analyses and interpretation to a variety of subject matters, but also in terms of further theoretical and methodological development of GIS itself.

Geolinguistics shares the fate of insufficient funding with many other disciplines in the humanities and social sciences. During the initial stages of research for this thesis, I had the naive view that overall, language documentation had been done rather comprehensively. Yet I was surprised by the apparent lack of documentation about the world's linguistic heritage. This is certainly not due to scientists not taking any interest in the investigation and documentation of languages, but simply because of insufficient financial support. GISystems and their capabilities for data storage, analysis and manipulation offer extraordinary potential beyond merely acting as a data repository, yet lacking documentation as a consequence of inadequate funding lets this potential go to waste.

I therefore very much appreciate that SIL International so willingly share their published and unpublished data with the public. Fukushima and Heap (2008: 147, 150) as well as Dahl (2006: 2-3) emphasise the importance of data distribution and accessibility as an asset for future geolinguistic research. Advances in technology allowing for WMS to be accessible to anyone with an internet connection offer a great opportunity for knowledge and data distribution. However, as described in this thesis, it needs to be ensured that the data are well described and documented in order to enable users to assess them.

My research for this thesis has convinced me that it is a key issue to raise awareness of the importance and the benefits of ancillary information for geolinguistic data. Starting from my investigation which only represents the tip of an iceberg of unexplored issues in geolinguistics, I believe that there are several paths well worth investigating. These include how current metadata standards can be adopted to suit linguistic research, or the development of effective guidelines on how to gather data documentation at the data collection stage during fieldwork.

Geolinguistics will in future rely even more heavily on GIS as a tool for analysis, policy decisions and other applications. It may at first glance not seem like a worthwhile or ‘glamorous’ task to address issues of data sources and data quality, when the outputs GIS produce seem to spark much greater interest in geolinguists. However, I believe that the investment in research would benefit the discipline greatly and would help construct a sound base from which to effectively exploit the plethora of options offered by GIS.

References

Ambrose, JE & Williams, CH 1991, 'Language Made Visible: Representation in Geolinguistics' in *Linguistic Minorities, Society and Territory,* ed CH Williams, Multilingual Matters, Clevedon, pp. 289–314.

Barni, M 2006, From Statistical to Geolinguistic Data: Mapping and Measuring Linguistic Diversity, EURODIV Conference "Understanding diversity: Mapping and measuring", Siena.

Breton, RJL 1976, Géographie des langues, P.U.F., Paris.

Breton, RJL 1991, Geolinguistics: language dynamics and ethnolinguistic geography, University of Ottawa Press, Ottawa.

Brimicombe, A 1997, 'A universal translator of linguistic hedges for the handling of uncertainty and fitness-for-use in GIS' in *Innovations in GIS 4.* Selected papers from the Fourth National Conference on GIS Research UK (GISRUK), ed Z Kemp, Taylor & Francis, London, pp. 115–126.

Brindle, JA October 12, 2009, Chakali locations. Email.

Britain, D in press (a), 'Conceptualisations of geographic space in linguistics' in *The Handbook of Language Mapping,* eds A Lameli, Kehrein Roland & Rabanus Stefan, de Gruyter, Berlin.

Britain, D in press (b), 'Language and Space: the variationist approach' in *Language and space: an international handbook of linguistic variation,* eds P Auer & J Schmidt, de Gruyter, Berlin.

Britain, D 2004, 'Geolinguistics - Diffusion of Language' in Handbooks of linguistics and communication science, eds U Ammon, HE Wiegand, G Ungeheuer & H Steger, de Gruyter, Berlin, pp. 34–48.

Buckley, DJ 1997, The GIS Primer. An Introduction to Geographic Information Systems, Pacific Meridian Resources, Inc. Available from: http://www.innovativegis.com/basis/primer/primer.html [12 September 2009].

Chen, PP 1976, 'The entity-relationship model—toward a unified view of data', *ACM Transactions on Database Systems*, vol. 1, no. 1, pp. 9–36.

Chrisman, N July 2009, A difference that makes a difference, 6th International Symposium on Spatial Data Quality, St John's. Available from: http://www.mun.ca/issdq2009/ISSDQ2009_Chrisman_Keynote.pdf [5 September 2009].

Codd, EF 1970, 'A Relational Model of Data for Large Shared Data Banks', *Communications of the ACM*, vol. 13, no. 6, pp. 377–378.

Crystal, D 1997, The Cambridge encyclopedia of language, Cambridge Univ. Press, Cambridge.

Dahl, Ö April 19, 2006, GISLI (GIS in Linguistics) - An Interactive Language Atlas, A research proposal submitted ot the Swedish Research Council.

Dahl, Ö & Veselinova, L 2005, 'Language Map Server' in *Proceedings of the 25th ESRI International User Conference.* Available from: http://proceedings.esri.com/library/userconf/proc05/papers/pap2425.pdf [6 September 2009].

De Smith, MJ, Goodchild, MF & Longley, PA 2006-2009, Geospatial Analysis. A Comprehensive Guide to Principles, Techniques and Software Tools. Available from: http://www.spatialanalysisonline.com/output/ [November 5, 2009].

Department of Geography, University of Colorado at Boulder 2000, The Geographer's Craft. Available from: http://www.colorado.edu/geography/gcraft/notes/notes.html [14 September 2009].

Devillers, R, Gervais, M, Bédard, Y & Jeansoulin, R 2002, Spatial Data Quality: From Metadata to Quality Indicators and Contextual End-User Manual, Istanbul, *OEEPE/ISPRS Joint Workshop on Spatial Data Quality Management*, pp. 45–55.

Devillers, R & Jeansoulin, R (eds.) 2006, Fundamentals of Spatial Data Quality, Wiley-ISTE, London.

Devillers, R & Zargar, A 2009, Towards Quality-Aware GIS: Operation-based retrieval of spatial data quality information, Spatial Knowledge and Information (SKI) Conference, Fernie, Canada. Available from: http://www.mun.ca/geog/people/faculty/rdevillers/publications.php [21 July 2009].

Doyle, A & Reed, C 2001, Introduction to OGC Web Services. An OGC® White Paper, Open Geospatial Consortium. Available from: http://www.opengeospatial.org/pressroom/papers [November 6, 2009].

Dublin Core Metadata Initiative 2005, Using Dublin Core. Available from: http://dublincore.org/documents/usageguide/ [15 September 2009].

Dueker, KJ & Kjerne, D 1989, Multi-Purpose Cadastre: Terms and Definitions, American Society for Photogrammetry and Remote Sensing and American Congress on Surveying and Mapping, Falls Church.

ESRI, 2001-2008, What is ArcGIS 9.3? Available from: http://webhelp.esri.com/arcgisdesktop/9.3/pdf/what_is_arcgis.pdf [November 16,2009].

Fisher, P, Comber, L & Wadsworth, R July 2009, What is in a name? Semantics, Uncertainty Standards and Data Quality, 6th International Symposium on Spatial Data Quality, St John's. Available from: http://www.mun.ca/issdq2009/ISSDQ2009_Fisher_et_al_Keynote.pdf [4 September 2009].

Fishman, JA (ed.) 1999, Handbook of language and ethnic identity, Oxford Univ. Press, New York.

Foote, KE & Huebner, DJ 1995, Error, Accuracy, and Precision. Available from: http://www.colorado.edu/geography/gcraft/notes/error/error.html [5 October 2009].

Fukushima, C August 4, 2008, Progress in geolinguistics: what has been made possible using a computer? The Thirteenth International Conference on Methods in Dialectology, Leeds.

Fukushima, C & Heap, D 2008, 'A Report on the International Conference: Geolinguistics Around the World', *Dialectologia*, No. 1, pp. 135–156. Available from: http://www.publicacions.ub.es/revistes/dialectologia1/ [13 September 2009].

Fukushima, C&Y 1993, An Approach to Computer-Assisted Linguistic Geography: SEAL Users' Manual.

Gilliéron, J & Edmont, E 1902-1910, Atlas linguistique de la France, Champion, Paris.

GIS in Linguistics (GISLI). A Language Mapping Project, Department of Linguistics, Stockholm University. Available from: http://ling-map.ling.su.se/website/index.html [September 5, 2009].

Global Mapping International September 14, 2009, World Language Mapping System. Available from: http://www.gmi.org/wlms/ [29 September 2009].

Goodchild, MF 1997, NCGIA Core Curriculum in Geographic Information Science: Unit 002 - What is Geographic Information Science? Available from: http://www.ncgia.ucsb.edu/giscc/units/u002/ [14 September 2009].

Goodchild, MF 1992, 'Geographical information science', *International Journal of Geographical Information Systems*, vol. 6, no. 1, pp. 31–45.

Gubbins, P & Holt, M (eds.) 2002, Beyond boundaries. Language and identity in contemporary Europe, Multilingual Matters, Clevedon. Available from: http://www.gbv.de/dms/sub-hamburg/337531234.pdf.

Heuvelink, GBM 1997, Uncertainty Propagation in GIS. Available from: http://www.ncgia.ucsb.edu/giscc/units/u098/u098_f.html [15 October 2009].

International Organization for Standardization 2001, Geographic Information - Quality principles. Draft International Standard ISO/DIS 19113, Geneva.

International Organization for Standardization June 1, 2009, Standards Guide ISO/TC 211 Geographic Information/Geomatics. Available from: http://www.isotc211.org/Outreach/ISO_TC%20_211_Standards_Guide.pdf [4 September 2009].

Jakobsson, A & Giversen, J (eds.) 2007, Guidelines for Implementing the ISO 19100 geographic Information Quality Standards in National Mapping and Cadastral Agencies. Available from: http://www.eurogeographics.org/documents/Guidelines_ISO19100_Quality.pdf [July 19, 2009].

Joseph, JE 2004, Language and identity. National, ethnic, religious, Palgrave Macmillan, Basingstoke, Hampshire.

Kent, M 2003, 'Space: Making Room for Space in Physical Geography' in *Key Concepts in Geography,* eds SL Holloway, SP Rice & G Valentine, Sage Publ., London, pp. 109–130.

Kluge, A & Hatfield, DH 2002, 'Sociolinguistic Survey of the Safaliba Language Area' in *SIL Electronic Survey Reports*, ed SIL International, Dallas. Available from: http://www.sil.org/silesr/2002/SILESR2002-041.pdf [25 April 2009].

Korte, GB 2001, The GIS Book. How to implement, manage, and assess the value of Geographic Information Systems, OnWord Press, Albany, NY.

Lambrecht, P, Kassell, A, Potter, M & Tucker, S 2008, 'The Sociolinguistic Situation of the Uyajitaya [duk] Language' in *SIL Electronic Survey Reports*, ed SIL International, Dallas. Available from: http://www.sil.org/silesr/2008/silesr2008-019.pdf [26 April 2009].

Language and Location: A Map Annotation Project. Available from: http://llmap.org/ [5 May 2009].

Lewis, MP 2009, Ethnologue: Languages of the World, Dallas. Available from: http://www.ethnologue.com/.

Longley, PA, Goodchild, MF, Maguire David J. & Rhind, DW 2005, Geographical Information Systems and Science, Wiley, Chichester.

LR Sociolinguistics October 19, 2009, Coordinates Pangin. Email.

Luo, W, Hartmann, J, Li, J & Sysamouth, V 2000, 'GIS Mapping and Analysis of Tai Linguistic and Settlement Patterns in Southern China', *Geographic Information Sciences*, Vol. 6, No. 2, pp. 129–136.

Mark, DM 1999, 'Spatial Representation: A Cognitive View. In v. 1, pp. 81-89' in *Geographical Information Systems: Principles and Applications*, eds DJ Maguire, MF Goodchild, DW Rhind & P Longley, Wiley & sons, pp. 81–89. Available from: http://www.geog.buffalo.edu/~dmark/BB2chapter.html [November 5, 2009].

Milroy, L & Gordon, M 2006, Sociolinguistics. Method and interpretation, Blackwell, Malden, Mass.

National Center for Geographic Information and Analysis, no date, GIS Development Guide. Available from: http://www.ncgia.buffalo.edu/sara/volumei.pdf [19 September 2009].

National Geospatial-Intelligence Agency, NGA GEOnet Names Server (GNS). Available from: http://earth-info.nga.mil/gns/html/index.html [October 16, 2009].

Ogino, T 1980, 'Computational Dialectology Using GLAPS. Automated Processing of Field Survey Data', *in: Proceedings of the 8th conference on Computational linguistics*, pp. 605–613.

Parker, D & Cool, B 2008, Computational Approaches to Mapping and Visualizing Language Data, Michigan State University. Available from: http://www.llmap.org/llmap-mclc08.pdf [21 July 2009].

Peeters, YJD 1993, 'Introduction - The Political Importance of the Visualisation of Language Contact' in The Cartographic Representation of Linguistic Data, eds YJD Peeters & CH Williams, pp. 6–8.

Pienemann, M & Keßler, J 2007, 'Measuring bilingualism' in *Handbook of multilingualism and multilingual communication,* eds P Auer, L Wei, K Knapp & G Antos, Mouton de Gruyter, Berlin, pp. 247-275.

Poehali 2002-2009. Available from: http://poehali.org/ [November 16, 2009].

Potter, M, Lambrecht, P, Alemán Laura & Janzen Correna 2008, 'The Sociolinguistic Situation of the Ambakich Language' in *SIL Electronic Survey Reports*, ed SIL International, Dallas. Available from: http://www.sil.org/silesr/2008/silesr2008-012.pdf [26 April 2009].

Pryce, WTR & Williams, CH 1988, 'Sources and Methods in the Study of Language Areas: A Case Study of Wales' in *Language in geographic context,* ed CH Williams, Multilingual Matters, Clevedon, pp. 167–237.

Rueck, M & Jore, T 2003, 'The Sociolinguistic Situation of the Sam People' in *SIL Electronic Survey Reports,* ed SIL International, Dallas. Available from: http://www.sil.org/silesr/2003/silesr2003-019.pdf [26 April 2009].

Shamsi, UM 2005, GIS applications for water, wastewater, and stormwater systems, Taylor & Francis, Boca Raton, Fla.

Shi, W, Fisher, PF & Goodchild, MF (eds.) 2002, Spatial Data Quality, Taylor & Francis, London.

Sumathi, S & Esakkirajan, S 2007, Fundamentals of Relational Database Management Systems, Springer, Heidelberg.

Sun Microsystems, Inc. 2008-2009. MySQL 5.1 Reference Manual. Available from: http://dev.mysql.com/doc/refman/5.1/en/ [16 November, 2009].

Tejel, J 2009, Syria's Kurds. History, Politics and Society, Routledge, Abingdon.

The Open University, Open Learn Learning Space. Using Numbers and Handling Data. Available from: http://openlearn.open.ac.uk/mod/resource/view.php?id=289651 [5 October 2009].

Thrift, N 2003, 'Space: The Fundamental Stuff of Geography' in *Key Concepts in Geography,* eds SL Holloway, SP Rice & G Valentine, Sage Publ., London, pp. 95–107.

Tompkins, B, Hatfield, D & Kluge, A 2002, 'Sociolinguistic Survey of the Chakali Language Area' in *SIL Electronic Survey Reports,* ed SIL International, Dallas. Available from: http://www.sil.org/SILESR/2002/SILESR2002-035.pdf [26 April 2009].

United Nations Development Programme (UNDP) 2004, Human Development Report. Cultural liberty in today's diverse world, New York. Available from: http://hdr.undp.org/en/media/hdr04_complete.pdf [7 November 2009].

UNESCO Section of Intangible Cultural Heritage 2009, Interactive Atlas of the World's Languages in Danger. Available from: http://www.unesco.org/culture/ich/index.php?pg=00206 [25 September 2009].

Van der Merve, IJ 1993, 'A Conceptual Home for Geolinguistics: Implications for Language Mapping in South Africa' in *The Cartographic Representation of Linguistic Data,* eds YJD Peeters & CH Williams, pp. 33–49.

Veselinova, L & Booza, JC 2009, 'Studying the Multilingual City: a GIS-based approach' in *Journal of Multilingual and Multicultural Development,* No. 30, pp. 145–165.

Wang, F, Hartmann, J, Luo, W & Huang Pingwen 2006, 'GIS-Based Spatial Analysis of Tai Place Names in Southern China: An Exploratory Study of Methodology', *Geographic Information Sciences,* Vol. 12, No. 1, pp. 1–9.

Wenker, G unpublished, Sprachatlas des Deutschen Reichs. (Survey period 1876-1887).

Williams, CH (ed.) 1988, Language in Geographic Context, Multilingual Matters, Clevedon.

Williams, CH 1996, '‘Geography and Contact Linguistics’' in *Handbooks of linguistics and communication science,* eds H Goebl, A Burkhardt, G Ungeheuer, HE Wiegand, H Steger & K Brinker, de Gruyter, Berlin, pp. 63–75.

Wyoming Geographic Information Science Center, Metadata Education Project. Available from: http://www.uwyo.edu/wygisc/metadata/quality.html#technique [October 21, 2009].

Yeung, AK 1998, NCGIA Core Curriculum in Geographic Information Science: Unit 051 – Information Organization and Data Structure. Available from: http://www.ncgia.ucsb.edu/giscc/units/u051/ [November 4, 2009].

Zargar, A & Devillers, R 2009, An Operation-based Communication of Spatial Data Quality, GEOWS'09 Conference, Cancun, Mexico. Available from: http://www.mun.ca/geog/people/faculty/rdevillers/zargar-QualityCommunication.pdf.

Appendix A

This table lists the data extracted from the SIL survey reports which are relevant to my sample application.

Language	Year surveyed	ISO 639-3 code	Classification	Dialects	Locations (settlements) and estimated numbers	Total estimated number of speakers
Chakali	1995	cli	Niger-Congo, Atlantic-Congo, Volta-Congo, North, Gur, Central, Southern, Grusi, Western	None (acc. to village elders)	- Tosa: 850 - Tissa: 320 - Sogla: 215 - Motigu: 480 - Ducie: 2,800 - Gulumbele: 370 - Katua: 300	6,000
Safaliba	1995	saf	Niger-Congo, Atlantic-Congo, Volta-Congo, North, Gur, Central, Northern, Oti-Volta, Western, Northwest	None (acc. to village elders)	- Mandari: 1,770 - Gbenfu: 380 - Manfuli: 220 - Tanyiri: 55	4,000
Ambakich	2003	aew	Sepik-Ramu,Ramu, Ramu Proper, Grass, Grass Proper	- Northern (Pangin and Arango) - Southern (Akaian, Ombos, Oremai, Agurant, Yaut)	- Pangin - Arango - Akaian - Ombos - Oremai - Agurant - Yaut	770
Uyajitaya	2003	duk	Trans-New Guinea, Madang-Adelbert Range, Madang, Rai Coast, Nuru	- Uyajitaya (Didiwala, Uya, Bauri) - Amowe (Buai=Tagoe=Baui)	- Didiwala:160 - Tagoe: 351 - Uya: 342 - Bauri: 191	1,044
Sam	2001	snx	Trans-New Guinea, Madang, Rai Coast, Mindjim	None	- Songum - Buan - Wongbe	600-700