ERC Starting Grant

Research Proposal (Part B2)

Section 2: Scientific Proposal

1. State-of-the-art

1.1. Native South American languages

Although Spanish and Portuguese are the primary languages for almost the complete human population of South America, there are actually more than 500 languages spoken in this part of the world (Gordon & Grimes 2005, slightly lower estimates of the number of living languages can be found in Campbell 1997: 170 and Fabre 1998). Most of these languages are so-called 'native' languages predating European conquest, most probably direct descendants of the languages of the first human populations entering this continent. Although human populations only entered this part of the world relatively recently (when exactly is controversial), and notwithstanding the very early descriptive and comparative interests by linguists for these languages (the first grammars and dictionaries were published in the 16th century, and the first linguistic families were proposed in the 18th century, cf. Campbell 1997: 29 ff.), still very little is known about the linguistic phylogeny of South American languages. This project aims to unravel the linguistic events that shaped the languages of this continent.

Currently, there is broad consensus about the existence of a few major linguistic families in South America (e.g. Arawakan, Cariban, Tupían, Tucanoan, Chibchan, Quechuan, Panoan), and very many small and shallow groups of related languages, many of which have only one or two member-languages spoken today (Campbell 1997: 170 ff.). Both the details of the internal structure of the accepted groups, and the relation between the many linguistic families, both large and small, are still open questions. Partly, this lack of knowledge is caused by the relatively limited amount of research that has attempted to unravel the linguistic prehistory of South American languages, and by the fact that only relatively recently have broad descriptive efforts started to document more than just the few major native languages of this continent.

The special geographical formation of South America, which is dominated by large mountain ranges and large tropical river basins, has probably also lead to a much more intermingled linguistic situation compared to many other parts of the world. Cross-linguistically, mountain ranges often have an intermingled linguistic situation (cf. the Caucasus or New Guinea), probably caused by valley-wise spread of languages, with older reclusive groups isolated at the end of valleys. Tropical river basins have the special property that populations spread almost exclusively along the rivers, with newer populations settling in between existing groups, or passing them by completely. This leads to the strikingly non-continuous geographic distribution of related languages.

The approach taken by this project to uncover deeper and more detailed genealogical relationships between the languages in South America is to bring together as much as possible of the information available about these languages, and to develop computer-assisted methods to quantify the analysis of all this data. The basis for the quantification will be the historical-comparative method.

1.2. The historical-comparative method

The historical-comparative method to uncover genealogical relationships between languages is an established and highly refined approach, informed by more than two centuries of experience and application. This method is based on the detection of homologous strings of sounds between languages, called 'cognates'. Shortly summarized, the historical development of phonemes is established on the basis of a sizable set of cognates, ultimately leading to a phylogeny of the languages considered (cf. Hoenigswald 1960; Hock & Joseph 1996).

Although the historical-comparative method is strongly codified and linguistically highly technical, in practice it is still considered more of an art than a mechanical procedure. The art-like nature of this method is to a large extent caused by the constitution of the available data. To truly understand and evaluate the available sources according to the historical-comparative method, an in-depth knowledge of the languages in question is needed, and also of the descriptive tradition of these languages. Much of the necessary knowledge about individual languages is not explicitly codified, requiring inference of the implicit assumptions made in the sources, or requiring an experienced teacher to introduce one to the tricks of the trade. Aggravating the situation is the fact that the available sources consist almost completely of printed books (or even unpublished manuscripts). Consequently, the quintessential search for cognates normally involves manually paging through large piles of alphabetically ordered dictionaries. Only a good intuition about the approximate written appearance of a cognate in a particular dictionary will allow one to successfully find any cognates, let alone a sizable collection of them.

Both these problems will be addressed in this project. First, by digitizing published data and, second, by codifying implicit knowledge in a computer-readable format, it will become possible to enhance the traditional historical-comparative method with computer-assisted techniques. The development and deployment of such computer-assisted processes is the second main goal of this project. The quantification of the historical-comparative method will allow more linguistic data to be processed faster and more reliably. As a consequence, more evidence for genealogical relations will be uncovered. The availability of more evidence is essential to push back the time-depth beyond the few thousand years to which historical linguistics is traditionally limited.

1.3. Wordlist comparison

To circumvent the highly successful, but rather tedious historical-comparative method, most quantitative approaches to language history have instead focussed on the wordlist approach, proposed 50 years ago by Morris Swadesh (1952). He suggested restricting the historical comparison of languages to the translations of a (short) list of meanings, preferably to such meanings that would only slowly change their expression. Although originally rather intended to quickly get an impression of promising genealogical links (which then were to be fully investigated by the historical-comparative method), the wordlist approach has often been used in recent decades to proclaim proof for the detailed structure of genealogical trees of languages (Gray & Jordan 2000; Nakhleh et al. 2005), and even to attach dates to putative proto-stages (Gray & Atkinson 2003). In mainstream historical linguistics, this approach has been heavily criticized (though not always for the right reasons), but because of its easy deployment this approach is still very popular.

In its simplest incarnation, the wordlist approach uses a basic measure of (dis)similarity between pairs of wordlists (e.g. a Levinstein distance) to quantify the linguistic proximity between languages. In its more compelling, but also more elaborate interpretation, the wordlist approach quantifies lexical replacement. Given a particular meaning, the fact that one set of languages shares a cognate lexeme for that meaning, while other languages share another cognate lexeme for the same meaning, can be interpreted historically as the result of a lexical replacement. Somewhere in the history of this family of languages, one of the ancestral languages switched from using the one cognate to the other. Given a collection of such replacements (without assuming any knowledge about the direction or probability of replacement), a subgrouping of the languages can be inferred by using any of the character-reconstruction methods from biological phylogenetics, like Maximum Parsimony, or the currently favored Maximum Likelihood models based on Bayesian inference and Markov chain Monte Carlo methods.

This second, more convincing, lexical-replacement interpretation of wordlists has the obvious disadvantage that it is necessary to ensure cognacy before it is possible to assess that replacement has happened. This is highly problematic, because the decision whether two forms are cognate or not is normally far from obvious for languages that are only distantly related. Taken seriously, this interpretation of wordlists presupposes the results of the traditional historical-comparative method, which it originally tried to circumvent. For that reason, this project focusses on the quantification of the historical-comparative method. Only if that method has been applied, lexical replacement will be computed as one possible method to quantify subgrouping of languages.

Another problem with the wordlist approach is that the meanings that form the basis of the wordlists are only very approximately defined. Wordlists are normally collected by looking for the translation of a particular English word (or a word from another widely spoken major language) into any of the other languages studied. However, the extension of word meanings is radically different between languages, so it is often difficult to tell whether there has only been a slight shift in meaning or a complete replacement. Both situations will be equally classified as examples of replacement in a wordlist study. In this project, the conceptual principle to use lexical replacement for the investigation of the phylogeny of languages will be taken up in a more general form. Instead of restricting the relevant data to short and heterogeneous wordlists, this project will quantitatively compare the structure of a collection of lexical fields. A lexical field contains many similar concepts, and the fine-grained semantic shifts of the lexemes within such a field will be used to refine the phylogeny of the languages investigated.

1.4. Grammatical comparison and corpus-based approaches

Almost all available knowledge about genealogical relations between languages is based on the comparison of lexical elements (including both roots and bound morphemes), either by using the historical-comparative method, or by wordlist comparison. In recent years, there have been various endeavors to also use more abstract grammatical characteristics (Dunn et al. 2005) and stretches of running text (Huffman 2003; Lüdeling 2006) to investigate the phylogeny of languages. Both these approaches are highly promising and deserve further research. However, the amount of available (or easily obtainable) data for such comparisons is much lower in comparison to the wealth of lexical information available about the world’s languages.

Texts with translation (or, even better, texts with detailedly aligned glossing) are considered here to be lexical resources insofar as the link between phonological form and meaning is concerned. One aspect of a text-cum-translation is that it provides information about lexical elements and their meaning, and in this sense textual data will be used in the current project. However, there are other characteristics of texts, like frequencies and contextual dependencies, that will not be investigated here. One of the major problems with such characteristics is their strong dependency on style and genre, both of which are difficult to control for the lesser-described languages that will be the focus of this project.

Besides the amount of data, another problem with abstract grammatical and textual characteristics is that it is more difficult, compared to lexical data, to distinguish between horizontal and vertical transfer. Languages are complex objects, shaped both by descent (vertical transfer) and various kinds of contact (horizontal transfer). The historical-comparative method offers a clear model how to separate these two kinds of transfer, based on the fact that loanwords do not share the same phonological history as vertically descended words. The process to separate the two kinds of transfer is a tedious one, but it is possible purely within the analysis of the lexical data. In contrast, for grammatical and textual characteristics it is unclear how a linguistic separation between vertical and horizontal transfer can be achieved, except by using a comparison with non-linguistic factors, like geographical proximity.

1.5. Relation to other disciplines

Knowledge about the history of languages is only one piece of the puzzle to unravel the history of the human species. The ultimate goal of historical linguistics is the comparison of its result with the insights gained from other disciplines, like archeology, genetics, or cultural anthropology (e.g. Cavalli-Sforza et al. 1988). Both matching and non-matching phylogenies from different fields will lead to reconstructions of the forces that shaped the historical events, albeit different forces will be reconstructed for matching and for non-matching situations. Yet, for this comparison to be successful, comparable evidence is needed from all fields independently. This project aims to produce more finely-grained linguistic phylogenies to enliven the cooperation with other disciplines interested in the history of the human species, both with regard to dating, and concerning the interpretation of the forces underlying the historical events.

One of the major deficits of traditional historical linguistics is that it only produces minimally resolved phylogenetic topologies, which are of limited value for interdisciplinary comparison. In almost all available linguistic phylogenies, there are various ancestral nodes that split up into more than two branches ('unresolved branching'), and the branch-lengths are never specified ('only topological structure'). To allow for better comparison with phylogenies from other disciplines, linguistic phylogenies should at least specify whether any case of unresolved branching is due to limitation of the data, or whether there is a substantial claim being made that the historical situation is truly unresolved (truly unresolved branching can arise through the late split-up of a long-lasting active dialect chain, or through subsecutive splits of human populations in rapid succession). Likewise, branch lengths have to be specified, also if they just describe some measure of cumulative change between ancestor and daughter state. The difficult relationship between such a purely linguistic measure of branch length and the amount of historical time that has passed between ancestral and daughter state will never be resolved simply by applying a clock-like mechanism of language change (which represents at best a rough approximation). Only the joint forces of linguistics, genetics, and archeology will allow for a more secure dating of historical events and proper calibration of branch lengths.

2. Objectives

2.1. Uncovering phylogeny of South American languages

The immediate goal of the project is to uncover and clarify phylogenetic relationships between South American languages. Very little is known about the historical development of the linguistic diversity in this part of the world, and any more detailed insights will be of strong interest not only to linguists, but also to geneticists, anthropologists and archeologists.

Although the languages of this part of the world are relatively little studied, still large amounts of data are available about them, though predominantly in printed form. To quantitatively investigate the linguistic relationships, a large effort will be made in the context of this project to digitize as many dictionaries and translated texts of South American languages as possible, starting with the more accepted larger families (e.g. Arawakan, Cariban, Tupían), and working our way through to the many smaller families and language isolates. The resulting large body of digitized data is expected to stimulate historical-comparative linguistic research also after the current project has ended.

Further, specialists working on the relevant languages will be contacted to assist in the assessment of the quality of legacy descriptive data, to obtain unpublished data, and to help with the interpretation of the quantitative investigations. The project strives to incorporate the linguists working on individual languages into the high-level analyses performed in the context of this project, aiming to build a bridge between descriptive linguistics and phylogenetic modeling.

2.2. Quantifying the historical-comparative method

A more general objective of this project is to transform historical-comparative linguistics from a primarily handcrafted scholarly endeavor, performed by individual researchers, into a quantitative and collaborative field of research. This transformation will be based on the experience of two centuries of historical linguistic practice, but will overhaul the traditional methodology by incorporating mathematical modeling and algorithmic approaches. This project explicitly focusses on quantifying the traditional historical-comparative method. The major advantages of quantitatively assisted method for historical-comparative linguistics are that:

- more data can be processed consistently and much quicker, and as a consequence more evidence for genealogical relations can be uncovered. More evidence is essential to push back the time-depth to which historical linguistics is traditionally limited;

- the impact of arguments (and the amount of negative evidence) can be assessed more precisely. For example, the frequency of occurrence of a regular sound change can be quantified, and likewise the frequency of instances where the sound change does not apply;
- interpretative decisions have to be stated explicitly, allowing for the comparison of different interpretations as to their adequacy. For example, the impact of different expectations about preferred directions of sound change can be easily compared;
- recurring historical linguistic processes can be identified and inform reconstruction efforts. For example, the probability of sound and meaning changes can be estimated from the data;
- the resulting quantitative proposals for the phylogeny of a group of languages can be much better compared to phylogenies from genetics or archeology.

The anticipated methodology for doing quantitative historical linguistics will necessarily be collaborative and interdisciplinary, involving linguists, mathematicians and computer scientists alike. A central aim of the project is to build a truly interdisciplinary team, and spend time and effort within the context of the project to improve the interaction between linguistics and mathematics.

2.3. Reframing and extending phylogenetic methods

This project explicitly aims to surpass the simple application of existing phylogenetic methods to linguistic data. Instead, this project strives for a reformulation of available phylogenetic methods in the light of the special requirement of linguistic data. In this way, it is not just historical linguistics that can profit from methods and insights previously developed in mathematics and computer science. It is also expected that the special requirements posed by language variation and language change will reframe and further extend the reach of current phylogenetic methods. Specifically, the following characteristics of linguistic data pose a challenge to available phylogenetic methods; a challenge that will be addressed in the context of this project:

- the complex interplay between vertical and historical transmission in language change, maybe somewhat comparable to the biological situation with bacteria or viruses;
- the absence of a fixed mapping between phonetic sound and phonemic inventories in languages (i.e. strongly similar sounds can have rather different phonemic status in different languages). Because the large majority of linguistic data is in phonemic-like orthographies, the problem of language reconstruction can be compared to a (counterfactual) situation in biology in which codons would be compared across species, and the genetic code linking the codons to the amino-acids is slightly different between different species;
- the lexemes that will be compared are short, but highly informative strings. This is the reverse of the situation in biology, in which there are long strings of amino-acids, each element of which individually only contains limited information (viz. one of four states);
- the role of contextually-bound interpretation (‘meaning’) is straightforwardly accessible in the study of language. In comparison, the role of gene expression in biological phylogeny is extremely hard to tackle. The study of language phylogeny thus opens the possibility to more directly study the interplay between forces that relate to formal changes and those that concern functional changes.

3. Methodology

3.1. Approach

To investigate linguistic phylogeny, this project will focus on lexical material, both in form and meaning. For the context of this project, lexical material covers all language-particular strings of phonemes, including content words, grammatical words, and bound morphemes. Most of the necessary quantitative methodology will have to be newly developed, as not much is available for historical linguistics (but see Bergsma & Kondrak 2007; Kondrak 2002). The methodology to be developed will both be inspired by phylogenetic methods from biology and informed by the qualitative approaches as traditionally used in historical linguistics.

3.2. Mathematical modeling of language change

As a basis for the concrete algorithmic analyses of lexical material, we will develop a dynamic and probabilistic mathematical model of lexical change. At least the following kind of events will have to be captured by the model:

- changes in the phonological inventory (e.g. mergers, splits, push/pull-chains);
- changes in the phonetic expression of phonemes;
- conditions on the applicability of phonological and phonetic changes (i.e. context dependency);
- changes in the meaning of lexemes;
- introduction of new lexemes through 'spontaneous' innovation;
- introduction of complex lexemes through derivation and/or compounding, and the ensuing eroding of such complex forms;
- introduction of new lexemes through borrowing from other languages (either as superstrate, substrate, or adstrate).

The model will operate on three different levels, to be implemented either as parallel sub-models, or (ideally) as different aspects of the same mathematical formulation. The three relevant levels are:

- a physical utterance-based model, i.e. each individual sound-wave of an utterance is the basic entity of the model (cf. a replicator or gene-based model in biology);
- a psychological/neurological speaker-based model, i.e. each human speaker is considered as a basic entity of the model (cf. a vehicle/interactor or organism-based model in biology);
- a sociological language-based model, i.e. a language as a collection of shared codes is the basis of the model (cf. a group or species-based model in biology).

The model should both fit linguists' intuitions about the dynamics and probabilities of language change, but also be of practical value to develop and test algorithmic approaches for the analysis of real data. A basic version of the model should be ready within the first few months of the project, but refinements and changes will be incorporated throughout the project as experience with the practical analysis of concrete data is gathered.

3.3. Algorithmically assisted analysis of lexical comparison

The central methodological innovation of this project will be the development of algorithmic approaches to assist the historical-comparative approach to the reconstruction of the phylogeny of languages. These approaches will both include search and analysis of the basic data, and visual and statistical methods to help in the interpretation. The basic task will be to develop methods to uncover cognate forms between languages. This methodology will roughly look as follows:

- search through complete digitized lexica (not just small wordlists) for similar words. Both similarity in form and similarity in meaning will be evaluated;
- similarity of form can be assessed through computer-readable orthography profiles, to be prepared for each source by the linguists in the project. Further, a model of sound change will be used to find similar words after the application of common, or expected, sound changes;
- similarity of meaning can be assessed through WordNet-like semantic lexica as available for various of the major world's languages. However, typical local semantic similarity (e.g. arising from cultural specific metaphors) can be added by mining the language-specific lexical similarity (viz. meanings that are recurrently expressed by similar forms in set of languages can be interpreted as having similar meanings within the context of these languages);

- the central evidence to establish cognates is to successfully align similar words from different languages phoneme by phoneme (alike to the alignment of amino-acids in biology). Alignments will be informed by a model of sound change, and the structure of attested alignments can be used to refine this model in turn;
- aligned phonemes across languages are known in historical linguistics as 'correspondence sets'. The more often the same (or highly similar) correspondence sets are attested, the better the evidence for lexical cognacy of the aligned words ('regular sound change'). Similar, but not identical, correspondence sets indicate either contextually dependent sound changes, or cognate words based on horizontal transfer ('borrowing');
- The attested cognates and their correspondence sets will be used to quantitatively build a phylogeny of the languages involved, using the general model of language change as described in the previous section.

3.4. Reconstruction of lexical fields

Besides cognacy and regular sound change, this project will also investigate changes in the structure of lexical fields. This interest is inspired by the success of the quantification of lexical replacement as used in recent analyses of wordlists. However, instead of only looking at shared lexical replacements for a limited set of meanings, this project will reconstruct changes in the lexical organization of sets of similar meanings (i.e. 'lexical fields'). The method basically consists of two steps: first, to compile suitable lexical fields and, second, to analyze the differences between the structure of these fields in different languages.

The first step is to sample a suitable lexical field. Starting from a rough conceptual definition (e.g. household objects, body parts, verbs of motion, kin terms, verbs of causation, etc.) words are selected in each language that express meaning in the realm of this definition. Subsequently, meanings will be added to the field depending on language-specific synonymy or derivational patterns. The ideal size of such lexical fields has to be investigated in the course of this project.

The second step is to compare differences and similarities in the lexical structure of such a field between different languages. This comparison can at least be performed in two different ways, either without assuming any knowledge about cognates between languages, or with detailed insights into cognacy. First, without assuming cognacy, the basic approach is to characterize the structure of the lexical field independently for each language as a distance matrix of the meanings selected. The language-particular distance between two meaning is defined by the lexical dissimilarity within the structure of the language. The comparison of the resulting distance matrices can be used to estimate phylogenetic events in the realm of meaning. Second, when knowledge is available about cognacy between words of different languages, then the reconstruction of phylogenetic changes in lexical coding straightforwardly becomes a variant of character-based analysis in biology (viz. each meanings is a character, and cognacy is shared coding).

3.5. Linguistic interpretation and evaluation

The results of all quantitative methods developed in this project will first be tested against accepted linguistic families and relationships. Not only should the known relationships be (re-)discovered by the new methodology, it is also expected that much more evidence for the known relationships will be uncovered by the computer-assisted analysis. The more difficult, but also more interesting, problem is to evaluate new claims about genealogical relationship. The force of the proof of any newly proposed genealogical relationship will depend on the extent to which specialists of the languages in question can be convinced by the evidence unearthed. It is here that the planned quantitative emulation of the historical-comparative method will be extremely useful. By this methodology, the results of the quantitative methods will be such as to also make sense to 'traditional' historical linguists. The kind of evidence that will be produced consists of a large collection of cognates with regular sound correspondences, and shared changes in the meaning of lexemes (or shared lexical replacements).

Given the large and diverse group of languages studied in the course of this project, close relationships to the various specialists working on the primary description of these languages will be established. This project not only depends on their work because it provides the elementary data for our analyses, but also because the specialists are the prime judges to evaluate any new results put forward in this project. To show our involvement in and appreciation for their work, we plan to regularly invite language specialists to collaboratively discuss results from our quantitative studies. Further, and possibly more importantly in the long run, we plan to offer co-authorship to all specialists that have contributed a substantial amount of data, or their informed judgments, to our research. Also when the data has already been published (so a reference to the published work would suffice), we believe that the importance of basic language description should be much more strongly acknowledged in comparative linguistic publications. The simple solution put forward for this project is to move from the traditional single-authored publication in linguistics to multi-authored publication already established in many other fields of investigation.

3.6. Digitization of legacy data

To be able to computationally process large bodies of lexical material (including texts with translation), the data needs to be prepared in a suitable and coherent electronic format. Because the majority of linguists still rely primarily on printed publication of dictionaries and text collections, most of the information will first have to be digitized to be used in this project. The electronic availability of well-organized lexical material is of quintessential importance to transform comparative linguistics into a quantitative and collaborative field of research. To also allow others to take part in this line or research, we will take great care that the substantial digitization effort performed in the context of this project will be made available for reuse after the project has finished.

The digitization is not only a question of automatic scanning, but involves careful processing of all orthographic and formatting details that are central to the full understanding of linguistic data. The goal of the digitization in this project is not to produce approximate ‘quick-and-dirty’ scans of dictionaries as a disposable by-product of high-level interpretative research, but to provide an accurate standards-compliant re-encoding of all the detailed notations and annotations that have become the prime mark of high-quality linguistic description.

Attempts by the present PI to digitize lexica using off-the-shelf optical character recognition (OCR) software have failed miserably. Various developers of OCR technologies were contacted, but none of them was interested in the current project because of the high variability of the legacy originals, and the resulting diminishing returns of any investment. So, to quickly obtain a sizable body of high-quality digitized lexica, this project will outsource manual data entry, using double-keying to control the quality of the data. The digitized versions of legacy material will be encoded in such a way as to resemble the original work as closely as possible. All interpretative conversions that are necessary to obtain coherent data for the automatic processing will be encoded and stored separately. This separation of data source and interpretation will allow for effortless correction of interpretative decisions, for comparing competing interpretations, and for later inspection of the interpretative decision by peers.

To ensure future usability of the data collected, we will adhere to the standards that are emerging in the realm of language technology, like Unicode (ISO 10646) for character encoding, or the various standards developed by ISO’s technical committee 37, like the Lexical Markup Framework (ISO 24613) and Language Identification codes (ISO 639-3). For the encoding of orthographic conventions used in a particular source, we will orient ourselves to the Locale Data Markup Language (Unicode Standard 35), though extended with information that is particularly important to comparative linguistics. Most importantly, we will include an approximate phonemic IPA conversion, and, if possible, a more detailed approximation of the phonetic values underlying these phonemes in context. All conversions are performed basically by using regular expressions, explicitly including exceptional cases. In this way, any source can be converted on the fly to a consistent transcription for search and further processing. Most importantly, the original source is kept verbatim orthographically, so any errors in the conversion can be easily reconsidered, and multiple independent conversions are possible.

The digitized data will primarily be used internally to address the scientific questions as formulated in this project. To also provide outside access to this data, there are two crucial hurdles that have to be addressed: intellectual property and data persistence. These problems are addressed in two separate project initiated by the current PI within the context of the Max Planck Gesellschaft in Germany. The first project ("Linguistics Texts: Enhancing PubMan") is mainly a bibliographic initiative to store and archive photographic scans, contact authors and other IP holders to negotiate terms of republication, and provide a web interface to allow the rightful access to the scans. The second project ("Living Sources in Lexical Description") will provide an online infrastructure for publishing, peer-reviewing and archiving of newly collected structured lexical databases.

3.7. Planning and Execution

Year 1:

- Regular cross-disciplinary meetings of the team (starting with twice a week, later diminishing in frequency) to introduce linguists to mathematical modeling and algorithmic approaches, and to introduce mathematicians to the details of linguistic data and the historical-comparative method.
- Development of mathematical models and implementation of algorithmic approaches to automatically interpret lexical data, focussing on emulating the historical-comparative method. Preliminary testing will be done using computational modeling and by using some limited electronically available data (e.g. Intercontinental Dictionary Series).
- First batch of digitization, focussing on Arawakan, Cariban and Tupían families.
- For the linguists, digitization includes selection and preparation of photographic scans for manual data entry (the actual data entry will be outsourced), interpretation of scripts, codification of sound inventories, and analysis and codification of language-particular conventions in structuring dictionary entries.
- For the infrastructure development, digitization includes building a framework for storage and retrieval of data, including standards for cross-coding of the multitude of orthographies in the sources.

Year 2:

- Applying computational models to the first batch of digitized data: discovering phoneme alignments, regular sound correspondences, and cognate identification; development of methods to interpret these regularities diachronically.
- The linguists evaluate automatically generated proposals, and compare the first results to received opinion on historical relationships between languages.
- Regular team meetings continue, turning their focus to the interpretation of first results. The details of the further direction of the project might be changed depending on these outcomes of these first attempts.
- Second batch of digitization, focussing on genealogical groups of languages with members in close geographical proximity to the languages already digitized. This focus is intended to be able to investigate contact (horizontal transfer).
- Refining computational model and algorithmic approaches, also incorporating contact (e.g. using network models).

Year 3:

- Third batch of digitization: choice of languages will have to be decided upon during the project.
- Applying models and algorithms including contact to the data. Also, a first workshop with invited specialists on the languages in question will be held in this timeframe.
- Discussion and publication of results.
- Refining model and algorithms by incorporating changes in lexical semantics.

Years 4 and 5:

- Applying models and algorithms that incorporate meaning change.
- The second workshop with invited specialists on the languages in question will be held in this timeframe.

- Preparing extension of theoretical models used towards future applications on grammatical or corpus-based characteristics.
- Writing-up of results, including writing and defending of Ph.D. theses.
- Finishing digitization, and transferring data to archive for public availability and secure storage.

4. Resources and project costs

4.1. Team

The team will be constituted such as to be able to cope with the three different, though interrelated, endeavors envisioned for this project:

- A group of linguists with experience in South American languages and in historical-comparative linguistic methodology will select, prepare, and ultimately interpret the data. This group will consist of the PI and two Ph.D.-students. Because the preparation of the data for manual data entry will involve much interpretative work for the linguists, the Ph.D.-students will be offered funding for the complete period of 5 years. As it is expected that the Ph.D.-theses will be finished before the end of the project, an extra year of post-doc salary will be available to keep their experience in the project also after they will have finished their Ph.D.
- A post-doc with a mathematical and/or bioinformatics background will develop, implement and apply algorithmic analyses on the digitized data, taking inspiration for the modeling from traditional linguistic methods. Funding will be provided for the whole period of 5 years. For bare programming tasks, this person will be assisted by a student assistant from computer science and the infrastructure programmer (see below). There will be no mathematical Ph.D.-students, because the PI is not qualified to single-handedly supervise Ph.D.-projects in the field of mathematics.
- For the digitization to be organized and performed concisely and accurately, a dedicated infrastructure programmer will focus on building the necessary infrastructure and organizing the data. This person will oversee the many stages of digitization, and control the accuracy and completeness of the data. Linguistic knowledge is not required, though it would be beneficial. The Ph.D.-students and a student assistant from linguistic will assist in maintaining the database.
- Although these three tasks are separated here to highlight their different roles, the crux of this project will be the integration of these three endeavors into a single coherent undertaking. Because of the interdisciplinary nature of the project, time and effort will be dedicated to learn from each other about each other's fields.

Scientific staff:

- **PI** with institutionalized background in comparative linguistics, and intimate knowledge of mathematical approaches to phylogenetic reconstructions.
- **Post-doc** with background in mathematical modeling, recent development in bio-informatics and phylogenetic algorithms.
- **2 Ph.D. students** with background in historical linguistics and/or descriptive linguistics specializing in the reconstruction of South-American languages.

Digitization group (non-scientific staff):

- **Infrastructure programmer**: dedicated non-scientific programmer to support the data-intensive workflow of the project, implementation of algorithms, and feedback mechanisms that help linguists interpret the results of the automatic analyses.
- **2 Student assistants**: two student assistants, one from computer science and one from linguistics, to support the implementation of the algorithms and the input of the digitization efforts into the data framework.

4.2. Necessary resources

All data to be collected and processed will be completely digital. To prepare the data for quantitative analyses, this project needs:

- some limited funds to obtain, copy and scan resources. Most resources will be obtained through interlibrary loan. However, in some cases, copies of rare literature will have to be especially ordered, or it might even be necessary to send somebody to a library or archive to personally make a photocopy.
- funds to outsource manual data entry. Preliminary contact has been made with a company called Grepect GmbH in Garbsen, Germany, specialized in data entry of ‘unusual’ scripts. Costs estimates as listed below are based on their indications.

4.3. Existing Resources

The books to be digitized have to some extent already been collected and photographically scanned by the present PI while at the MPI-EVA in Leipzig. They are currently managed by the library of that institute. Further, the Intercontinental Dictionary Series (IDS, also hosted at the MPI-EVA) contains already digitized wordlists for many South American languages, which can be used immediately at the start of the project to test quantitative approaches.

References

Bergsma, Shane & Greg Kondrak. 2007. Multilingual Cognate Identification using Integer Linear Programming, RANLP Workshop on Acquisition and Management of Multilingual Lexicons:

Campbell, Lyle. 1997. *American Indian Languages: The Historical Linguistics of Native America.* (Oxford Studies in Anthropological Linguistics, 4). Oxford: Oxford University Press.

Cavalli-Sforza, Luigi Luca, Alberto Piazza, Paolo Menozzi, & Joanna Mountain. 1988. Reconstruction of human evolution: bringing together genetic, archeological, and linguistic data, *Proceedings of the National Academy of Sciences of the United States of America* 85(16): 6002-6006.

Dunn, Michael, Angela Terrill, Ger Reesink, RA Foley, & Steve C. Levinson. 2005. Structural phylogenetics and the reconstruction of ancient language history, *Science* 309(5743): 2072-2075.

Fabre, Alain. 1998. *Manual de las lenguas indígenas sudamericanas.* München: Lincom.

Gordon, Raymond G. and Barbara F. Grimes. 2005. *Ethnologue: Languages of the world.* Dallas, Tx.: SIL International.

Gray, Russell D. & Quentin D. Atkinson. 2003. Language-tree divergence times support the Anatolian theory of Indo-European origin, *Nature* 426: 435-439.

Gray, Russell D. & Fiona M. Jordan. 2000. Language trees support the express-train sequence of Austronesian expansion, *Nature* 405: 1052-1055.

Hock, Hans Henrich and Brian D. Joseph. 1996. *Language History, Language Change, and Language Relationship: An Introduction to Historical and Comparative Linguistics.* Berlin: Mouton de Gruyter.

Hoenigswald, Henry M. 1960. *Language Change and Linguistic Reconstruction.* Chicago: University of Chicago Press.

Huffman, Stephen M. 2003. The genetic classification of languages by N-Gram analysis. Ph.D. Thesis, Georgetown University, Washington DC.

Kondrak, Gregorz. 2002. Algorithms for Language Reconstruction. Ph.D. Thesis, University of Toronto.

Lüdeling, Anke. 2006. Using corpora in the calculation of language relationships, *Zeitschrift fur Anglistik und Amerikanistik* 54(1): 217-228.

Nakhleh, Luay, Donald A. Ringe, & Tandy Warnow. 2005. Perfect Phylogenetic Networks: A New Methodology for Reconstructing the Evolutionary History of Natural Languages, *Language* 81(2): 382-420.

Swadesh, Morris. 1952. Lexico-Statistic Dating of Prehistoric Ethnic Contacts: With Special Reference to North American Indians and Eskimos, *Proceedings of the American Philosophical Society* 96(4): 452-463.