One of the things we value most highly about working as interns at LINGUIST List is getting the opportunity to be involved in a number of very different functions. Out of all the projects we've worked on, MultiTree is unique in that it allows students and interns to work with both the research and the technical aspects of creating an online resource of such a significant size.

The Multitree project works on digitizing language relationship hypotheses and making them available in a searchable online database for use by the linguistic community and general public alike. It also maintains language code standards for use on various projects at the LINGUIST List and provides one of the coolest interactive visual representations of language trees currently available online.

Students and interns create each tree individually from original sources, which, more often than not, have never before been digitized or made available online. Some of the challenges we face when working on the MultiTree project include finding credible sources, determining how to classify languages and dialects, and transforming language relationship hypotheses into machine-readable format, all while remaining faithful to the author’s original intent and maintaining consistency across multiple sources.

The first step involved in creating a tree is locating the literature on which the tree will be based. A common starting point for us is the growing electronic treasure trove of myriad papers and books that LINGUIST List maintains on its private server. Oftentimes we will find publications with language relationship hypotheses that have yet to be represented on the MultiTree website. We also consult the citations and references found in these writings for other sources, which we then find online or through the EMU library system. We also obtain sources from corresponding with scholars who specialize in the language families that we work with.

Once we have found a source, we are then able to start the digitization process. Language relationship trees are formed by entering all languages involved into a spreadsheet that tracks the parent-child relations between nodes, as well as any additional information the author provides for the languages, such as alternative names, geographic location, or approximate dates of existence.

Each node must also be assigned a language code. This code is a unique three-letter tag that provides a more efficient way to refer to the language and is better suited for use by a computer. It also disambiguates between different languages with the same name, and...
Calvin Cheng

Much of my career in linguistics has involved not knowing what it’s really about, and even now, my understanding of the field continues to evolve. I originally imagined that it would give me the opportunity to learn lots of Romance languages (an obsession that I had in high school) and to analyze the orthographic correspondences between cognates in Catalan and Italian; little did I know that it would instead entail practicing implosives to keep myself amused while sitting in traffic or raising some guy named John—whose every action seems to be the obsession of syntacticians—to the spec-of-TP position. Then, after learning about language acquisition, I started to think that babies—not John—were the obsession of linguists, and that all the models that they created were ideas that ultimately dealt with how kids could acquire language so quickly and effortlessly.

I suppose that I should be disappointed that I’ve been wrong so many times about the essence of linguistic inquiry. Yet I continue to be inspired by the field, and I remain undiscouraged in my pursuit to deepen my understanding of language. I am therefore thrilled to be interning at LINGUIST List, where my involvement with the MultiTree and LL-MAP projects enable me to not only expand my knowledge of the discipline and of specific languages but also gain a hands-on understanding of some of the technologies that can be used to organize language data.

I am originally from Cerritos, a little suburban city in the outskirts of Los Angeles County, right next to the OC. I graduated magna cum laude from UCLA, completing a major in Linguistics and Spanish and two minors in Asian Languages and Asian Humanities. This fall, I get to keep being wrong about language through the Erasmus Mundus International Masters in Natural Language Processing and Human Language Technology, and I will be dividing my two years of study among three universities: the University of Wolverhampton (one semester), the Universitat Autònoma de Barcelona (one semester), and the Université de Franche-Comté (two semesters). I am very excited about how much I will grow as a linguist as I continue on my journey down the linguistics garden path, both from working at LINGUIST List and from my upcoming master’s program—I haven’t reached a dead end yet!
Jacob Collard

I am Jacob Collard, a rising sophomore at Swarthmore College in eastern Pennsylvania, but I have been studying linguistics for almost five years now, beginning from some introductory textbooks I found at the library. Lately, I find myself extremely interested in language documentation, particularly for endangered languages, and have been working with the ELCat project this summer, learning a lot about the extremely interesting languages of Africa and Australia and having a good time with it. I have also spent time on LL-MAP, with a focus on Africa and Australia there as well. After this summer and after my graduation in three years, I hope to get my PhD in linguistics and go on to study endangered languages, possibly in Africa. I also hope to spend time working with the syntax and semantics of these languages, in particular.

Outside of linguistics, I enjoy spending time outdoors, and I do the majority of my studying in forests. I also enjoy reading extremely old texts, programming, and occasionally attempting to write something, though my writing skills are much sharper at technical writing than at creative. I have an interest in pencils and other writing equipment, grown out of my need for extremely fine tips in order to support my minuscule handwriting. Some nights I also play roleplaying, strategy, or party games, grown out of my need to build worlds and tell stories.

Seng Lee

I am a recent University of Virginia graduate who jumped at the chance to work in my own field during the summer. Having — eventually — gotten over my swoons at the prospect of doing so at an organization as well-established as the LINGUIST List, I then proceeded to thoroughly relish the prospect of using my powers for good (which sounds much catchier than ‘using my education to help facilitate others’ studies’). Some of my too-many linguistic interests are language in contact situations and in interaction with social gender. I also appreciate language as a tool of daily communication and medium of creative endeavor. Outside linguistics, I enjoy SFF, transformative works, and dabbling with various arts.

Caela Northey

As a double major in computer science and linguistics, I am thrilled to be an intern this summer at the LINGUIST List! I just completed my third year at the University of Wisconsin – Madison and plan on graduating in the winter of 2013. My main interests in computer science are software development and natural language processing, and my linguistic interests include phonology and historical linguistics.

I’d like to one day attend graduate school at the beautiful Edinburgh University, but until then the class I am most excited to take next semester is Modern Irish Language. It’s really a privilege to work with such an amazing group of people, and I’m looking forward to seeing what everyone accomplishes in the future.
Monica Lesher

I fell into linguistics quite by accident. After dropping my original major in school I took an interesting-sounding course in English linguistics and it was love at first class. My great-grandparents emigrated to the U.S. from Germany, but unfortunately they did not wish to conserve the language within our family, German not having been a very popular language at the time. So I took it up in high school and have been studying it ever since.

I'm a Michigan native and recent graduate of Eastern Michigan University where I studied English Linguistics, German Language and Literature, and Philosophy. While in school I also participated in an intensive German immersion program through Portland State University in Oregon as well as attended a semester abroad in spring and summer of 2012 at Justus-Liebig Universität in Gießen, Germany.

I am currently an intern, working mostly on the MultiTree team. While I don't have a linguistic specialty yet per se, I haven't yet met a linguistic topic I didn't like and am enjoying the opportunity to expand my knowledge of linguistics at the LINGUIST List this summer. In the near future I'll most likely be pursuing a Master's degree in linguistics as well as certification in teaching English as a foreign language (TEFL) and/or teaching German as a foreign language.

EmemObong Udoh

Hi there! Thanks for watching my picture. You are looking at EmemObong Udoh, a Nigerian in his thirties who is always interested in interesting things. Read on, and you'll get to know a drop of this.

I hold a Master of Arts degree in Computational Language Documentation and a Post Graduate Diploma in Mass Communication, from the University of Uyo, Nigeria. My Bachelor's degree in Linguistics came courtesy of the same University. My interest areas for now include Phonetics/Phonology, Morphology, and Language Technology. My sympathy for endangered languages has, however tended to tilt me towards having more dealings with Language Documentation, and I have been doing this in the past years by way of joining internationally sponsored projects and presenting local and international papers at conferences. Be that as it may, Phonology remains my major calling in the Linguistics paradise.

I joined the teaching profession in July 2010 with an unquenchable longing to contribute in whatever measure towards saving some endangered languages of Nigeria from the steady march to extinction. Following her success stories and sincere interest in creating a central point of access to dispersed linguistic databases for the dissemination of information on accessible digital resources on all the world’s (endangered) languages, and on the discipline of linguistics, I found LINGUIST List to be a must-be place. I knew it would be a place that could contribute to meeting this longing of mine. I am therefore glad to be part of this success-story crew – an awesome crew of Linguists doing an awesome job for the always-very-curious Linguists.

Away from Linguistics, I am animatedly in love with music. I write songs and sing all kinds of music provided they make sense. There is surely an interphase between Linguistics and music, especially in the area of tonology. This is one nursed interest area for me, which will before long, become manifest; so if you have this kind of interest, you could come let's collaborate. Thanks though, for reading my lines.
Andrew Peters

I'm a linguistics major at Carleton College in Northfield, Minnesota, where I recently finished my third year; I grew up about forty miles south of there in Rochester, MN. Studying Latin in high school (and, more than I care to admit, reading a lot of Wikipedia in my free time) was what first got me interested in linguistics, and though I didn't yet know how seriously I would pursue it, my choice to attend Carleton was influenced by its being one of the only liberal arts colleges with a full-blown linguistics department and major.

These days, my favorite of the three traditional branches of linguistics is phonology, though there are of course many linguistic topics I have yet to learn. I'm also particularly interested in Spanish and Japanese linguistics as those are the major two non-English languages I know something about; this spring I studied Japanese linguistics at Doshisha University in Kyoto on a Carleton off-campus program. In June of this year, I joined LINGUIST List as a summer intern, and have been primarily working on the LL-MAP project. This project, as well as simply spending time with a group of talented and interesting linguistics-minded people, has really expanded my knowledge of the discipline – an opportunity that I'm grateful for and that I will be sad to leave behind in the fall.

Svetlana Tchistiakova

Hello! I am Svetlana Tchistiakova, and although I am originally from Russia, I've been living in Southern California for most of my life. My family left Russia and moved to Israel when I was three years old. Three years after that, we moved to the United States. We have been living in the Los Angeles area ever since. Learning a number of different languages in my childhood is what first sparked my interest in linguistics, although today I speak only English, Russian, and some French.

In 2010, I graduated from UCLA with a major in Linguistics and a minor in Russian language. While I am insatiably interested in almost all topics in linguistics, my previous studies have been primarily centered around Slavic languages, Icibemba vowel phonology, and language acquisition. I have recently also found an interest in computational linguistics. I feel very fortunate to have had the chance to study with some best and nicest linguists at UCLA, and am now very excited to put what I have learned into practice alongside the wonderful crew here at the LINGUIST List!

Outside of linguistics, I love to travel, and hope to get the chance to do more of this in the future. My future plans also include continuing my education in linguistics, though I have yet to pick a sub-speciality. I spend most of my leisure time with my boyfriend and my cat, playing video games, reading, and doing various outdoorsy activities.
The life of an LL-MAP map begins with its original source. We identify linguistically relevant maps from a variety of places – atlases, linguistics papers, databases, even Wikipedia articles. If the map is a physically printed one, we scan it in high resolution. The original image will be uploaded along with the digitized version, so that the source of our data is transparent to LL-MAP users.

Armed with images of our map, the first step in its transformation to LL-MAP is rectifying, or georeferencing. With the help of standardized basemaps showing borders and other geographical features, we mark tens to hundreds of points on our scanned image which we match with geographic coordinates. This allows software to precisely distort the image so that it fits the projection used by LL-MAP.

The next step is vectorizing: we trace the relevant shapes from the original image, recreating them in a digital "vector" format. This format is lossless, meaning that unlike a conventional image, it will not become pixelated or grainy no matter how far it is enlarged. It also requires a certain degree of interpretation. Some maps represent data through conventions that work well on paper but not in the LL-MAP interface, and in older maps, authors may not have had the same conception of language relationships as modern scholars. At this (and any) stage of the process, when deciding how to represent the data digitally, we have to balance the goals of faithfulness to the original source and accessibility to the user. We mark each item on the map with a variety of information, including relevant language identification codes – a crucial step because it's necessary for making LL-MAP searchable by language and interoperable with other LINGUIST List projects, primarily MultiTree, a system in which language relationships are represented as graphical “trees” for ease of understanding.

Indeed, one of the highlights of LL-MAP is this interoperability, and its implications for different tasks necessary for many subfields of linguistics. MultiTree is notable for its extensive use of the above-mentioned language codes (an extension of ISO 639-3), which allow for unambiguous reference to specific languages, dialects, and subgroups. The links between LL-MAP features and language codes allow users to easily access further information on the language in question. As both LL-MAP and MultiTree have expanded over the summer, there is a great deal of linking between different maps and trees possible. This also allows for quick reference from a map to typological information about a language.

The interoperability of LL-MAP also makes it easy to use data from other projects to create a map. This is useful not only for external scholars looking to map data from their research, but also lets us rapidly expand LL-MAP’s inventory of LINGUIST List maps. Data that is used for other LL projects has been applied to LL-MAP, meaning the maps also contain information about endangered or little-documented languages and genetic information, all in visual format.

After that information is properly coded into our maps, all that remains is to upload it to the LL-MAP website. Images we've digitized must be exported as vector shapefiles, while point data recorded in spreadsheets can be uploaded directly -- but once they’re run through the Scholar’s Workbench, both become datasets that can be used by any user, in as many online maps as are desired.

Pure data — even in shapefile form — cannot make a true map, however. We must take these datasets and flesh them out with color and texture, adjusting them for optimum readability. Some of the older maps, originally digitized without the modern vector format or high-quality monitors, needed to be reformatted in this stage in order to be brought up to par with the shiny new ones we've been publishing over the past three or four years — but a summer's time was more than enough for us to convert them all.

After a gamut of double-checks by fellow team members, the team leader, and the management, finished maps are ready to be made live for public viewing. We’ve been combing through the database lately for the best of the best, and they’ll be showcased on our front page as new featured maps — check it out!
Zac Smith

Zac Smith is unaware as to if he should write this biography in the first or the third person. He doesn't know if it would sound too conceited to write a whole list of fun facts and accomplishments in the first person, constantly bragging about how “I did this!” and “I am that!” He generally hates it when people talk about themselves and doesn't think that his own story is all that interesting, so if someone else were to write about him, or, if it at least appeared as if someone else wrote about him, the inevitable braggary and smarminess that pervades autobiographical articles could be done away with entirely, and his accomplishments and eccentricities would be able to shine on their own, untainted by first person narration and the discomfort of the reader doomed to read such a piece.

What he would want you to know, if he had any say in this thing, is that he considers himself extremely lucky and privileged to have been accepted into Cornell's PhD program in Linguistics after the summer. He has somehow convinced this esteemed university to not only let him attend, but to in fact pay for his attendance, and even pay him real money to do real research and real teaching. His parents would want you to know that they find it ironic that he refused to apply to any Ivy League school for his undergraduate career, and yet here he is, going to Cornell for graduate school. They would laugh like parents do.

It should also be known that upon entering his freshman year at Eastern Michigan University, he had no idea that he would be interested in linguistics. In fact, he had no idea what he wanted to do. He had some vague desire to be some kind of “scientist” and he also knew that he enjoyed Spanish class in high school. His first year at Eastern, he knew that the only classes he really enjoyed were his Spanish classes. He took all kinds of General Education classes to figure out what he wanted to do, including Astronomy, which was nice and all, but whatever.

However, he also fell in love freshman year, with his current girlfriend Jessica, who will be attending the Ithaca College School of Music, which is right down the street from Cornell. In fact, the main reason he applied to Cornell in the first place was because they wanted to go off to graduate school together. They also applied to some other places but those places weren't really interested in them, and to be honest, these two kids weren't quite interested in those schools anyway, except that Boulder, Colorado is a cool city.

So anyway, eventually, one of the Spanish professors of some higher-level courses at Eastern taught Zac, and this particular professor had studied linguistics when he was younger, and so taught some Spanish with a linguistically empirical flair. That’s what did it. The next semester, Zac enrolled in two introductory Linguistics Courses, and bam! Next thing he knew he was drawing X-Bar projections and analyzing pharyngeals.

He also started working at the LINGUIST List the summer before his senior year, where he got to digitize some maps and work on some websites and stuff. He got to attend the LSA Convention 2012 with a good chunk of the LINGUIST List crew and had a blast in Portland.
Karrie Kuecker

Working at LINGUIST List is really an experience like no other; equal parts intriguing and challenging. It exposed me to many different aspects of the discipline and also helped me identify what my interests are within the field. One of my favorite aspects of the job was editing the Ask-A-Linguist and Ask-An-Expert areas on the LINGUIST site. This provided a snapshot into not only what people were doing within the field, but also into how those issues and aspects that interested non-linguists or novice linguists. As a member and Team Leader of LL-MAP, I got to explore the use of Geographical Information Systems and their role in documenting language data. And, as my other colleagues have stated, one of the best things about working at LINGUIST was getting to work with such amazingly talented people! When looking back at all the things that this organization has accomplished in the last two years, it really is incredible to know that the bulk of the work was done by students! LINGUIST has provided us with all the tools to help not only further our academic pursuits, but to help further the discipline and take it to new and exciting places!

This fall marks the next phase in my academic career, as I will be starting on my PhD in Applied Language & Speech Sciences with a clinical certification in Speech Language Pathology at the University of Louisiana at Lafayette. I would have never known about this program if it had not been for LINGUIST List! I saw an advertisement for ‘Student Support’ about a year ago, e-mailed the professor to find out more information, and learned that the position would allow me to study Clinical Phonetics. Being an ‘Audio Geek’ I was excited to know that there was a program that would allow me to indulge my love of acoustics and also provide a more ‘hands on’ approach than many of the theoretical programs I had also been accepted to. This is sure to bring a new set of challenges but I’m looking forward to where the future brings me!
helps to avoid confusion between similar names. The International Organization for Standardization (ISO) maintains just such a list for all known languages, which is referred to as the ISO 639-3 code list. For example, English is abbreviated to [eng], Mandarin is [cmn], and Russian is [rus]. This provides a means for any two scholars from around the world to talk about the same language across different projects and hypotheses.

The LINGUIST List has generated a few additional codes for in-house use as well, including ones that track subgroups (clusters larger than a single language) and dialects (clusters smaller than a single language). The subgroup codes are four letters long (e.g. Quechua is 'quch', Niger-Congo is 'ncon'), whereas the dialect codes are six: three from the language they are classified under and an additional three-letter suffix (e.g. Eastern Cherokee is 'chr-eas', Sanjabi of Southern Kurdish is 'sdh-san'). These additional codes allow us to encode a finer degree of detail into our trees since we can distinguish between subgroups, languages, and dialects. If we come across a language that is not in the ISO list, we then submit a proposal to ISO to consider including it.

Once the spreadsheet is completed and peer-reviewed, the tree is ready to be uploaded onto the MultiTree website. The uploader builds the tree based on the parent-child relations and adds it to our database. These trees can then be browsed by geographic location and searched by publication, language name, and language code. They display in a customizable interactive viewer that presents the information in an easy-to-understand format. Users can even compare two trees at once, which is extremely useful when contrasting competing hypotheses about the same languages. In addition, scholars can download the XML code of a tree for use in further research.

In working on the MultiTree project, it is sometimes necessary to analyze older works on theories that are no longer accepted as plausible, known as heritage trees. We recently had the opportunity to work on several trees dealing with indigenous languages of the Americas. One theory we've worked, a journal article by Edward Sapir entitled Central and North American Languages, was originally published in the 1929 edition of the Encyclopædia Britannica and dealt partly with a group he called Hokan-Siouan, which is not recognized as a linguistic family today.

You might be wondering, “Why make a heritage tree at all?” We would respond to this question by pointing out that the past sheds light on the present. The development of language relationship theories, especially highly contended ones, often provides insight into other aspects of reconstructive linguistics as a whole. It is in this way that we can see the leaps and bounds that work in reconstructive linguistics has made through the past century. Heritage trees are also important because they act as a record of past linguistic theories.

Organizing a hypothesis such as Sapir’s into the format necessary to upload a tree into our database brings specific difficulties of its own. For one, there is no recognized language group “Hokan-Siouan” in any other available resource. This isn't uncommon; we often have to do some re-analysis of proposed family relationships. In the case of Sapir’s work on languages of North and Central America, the relationships he outlines diverge severely from the currently accepted theories, and a new subgroup code had to be generated. Because original authors of such heritage trees often cannot be contacted, much of the work done on heritage trees is based on personal judgment. For Sapir’s trees, we found ourselves making revisions, sometimes drastic ones, nearly every day.

Lastly, languages change at an alarming rate. Many of the languages Sapir dealt with one hundred years ago may be extinct today, or what Sapir considered a language then may now be considered a dialect by the majority of scholars or vice versa. Choosing the appropriate language and subgroup codes is a challenging but important part of representing the scholar’s work with as much faithfulness as possible.

The final results of our work on the MultiTree project are the product of much deliberation. In the end, we provide digitized theories from renowned scholars in the field of linguistics that might have otherwise been lost. This summer the MultiTree team has uploaded 30+ trees, focusing on language families in Australia, Asia, South America, and Africa. In total, MultiTree now consists of over 1,130 trees. Suffice it to say, the MultiTree project is both an ambitious, and enlightening project to work on at LINGUIST List.