Atlas Lingüístico-Etnográfico de Colombia Geolinguistic Corpus

Chapter · March 2018

CITATIONS
0

READS
60

5 authors, including:

Johnatan Estiven Bonilla
Caro y Cuervo Institute
7 PUBLICATIONS **3** CITATIONS
SEE PROFILE

Julio Alexander Bernal-Chávez
Caro y Cuervo Institute
21 PUBLICATIONS **8** CITATIONS
SEE PROFILE

Ruth Yanira Rubio López
Caro y Cuervo Institute
5 PUBLICATIONS **3** CITATIONS
SEE PROFILE

Andrea Lizeth Llanos
Caro y Cuervo Institute
4 PUBLICATIONS **3** CITATIONS
SEE PROFILE

LILA '18

CONFERENCE PROCEEDINGS

LILA '18 / V. INTERNATIONAL LINGUISTICS AND LANGUAGE CONFERENCE

Conference Proceedings

ISBN: 978-605-9207-98-0

Özgür Öztürk DAKAM YAYINLARI

March 2018, İstanbul.

www.dakam.org

Firuzağa Mah. Boğazkesen Cad., No:76/8, 34425, Beyoğlu, İstanbul

Cover Design: D/GD (DAKAM Graphic Design)

Print: Metin Copy Plus, Mollafenari Mah., Türkocağı Cad. 3/1, Mahmutpaşa/Istanbul, Turkey

Conference Coordination: DAKAM (Eastern Mediterranean Academic Research Center)

ATLAS LINGÜÍSTICO-ETNOGRÁFICO DE COLOMBIA GEOLINGUISTIC CORPUS

JULIO ALEXANDER BERNAL CHÁVEZ, JOHNATAN ESTIVEN BONILLA, RUTH RUBIO, ANDREA LIZETH LLANOS CHÁVEZ, DANIEL EDUARDO BEJARANO BEJARANO

Julio Alexander Bernal Chávez, Research Director, Instituto Caro y Cuervo; Johnatan Estiven Bonilla, researcher, Instituto Caro y Cuervo; Ruth Rubio, researcher, Instituto Caro y Cuervo; Andrea Lizeth Llanos Chávez, research assistant, Instituto Caro y Cuervo; Daniel Eduardo Bejarano Bejarano, research assistant, Instituto Caro y Cuervo.

Abstract

This paper describes the Atlas Lingüístico-Etnográfico de Colombia (ALEC) geolinguistic corpus. The ALEC corpus is composed of 1523 linguistic, ethnographic and mixed maps, more than 16,000 photographs and 765 audio sessions. Twenty-three researchers collected information in 262 locations of Colombia between 1956 and 1978, interviewing 2234 informants. Since 2015 Instituto Caro y Cuervo's Corpus and Computational Linguistics research line has been working on the digitization of these files by means of structural and descriptive metadata, aiming to present a linguistic conservation material to the academic community that can be reviewed with different quantitative and qualitative methods and at different language levels. Three Information Systems are being developed for the presentation of the ALEC corpus: an ALEC website (ALEC Web); a Geographic Information System of the ALEC (ALEC GIS); and the Oral Corpus of ALEC (Website CLICC)

1. Introduction

Atlas Lingüístico-Etnográfico de Colombia (ALEC) Corpus is a project led by the Instituto Caro y Cuervo's (ICC) research line on Corpus and Computational Linguistics in Bogotá, Colombia. The aim of this project is to create a geolinguistic and oral corpus on the web with the materials collected by the ALEC researchers between 1956 and 1978 at 262 locations across the country. Building this corpus enables systematization, disclosing and exploiting the materials through the new instruments and platforms developed.

Linguistic atlases are long-term research products that allow for the collection of large amounts of data in different locations across a territory. Dialectology and geolinguistics are the basis of their investigation processes. Dialectology is about language variation related to geographic (Montes, 1993). While Coseriu (1956) defines geolinguistics as a comparative and dialectological method that allows the creation of maps from the linguistic information of a specific territory. Traditionally, the researcher considered work over once the maps publishing was done. That is why several authors have mentioned the maginalization of the materials produced on language geography (Fernández, 2010). In this sense, one of the main goals of this project is to facilitate the use and disclosure of geolinguistic data for new research and as a pedagogical tool for teaching Colombian Spanish.

Moreover, the use of technological tools and the possibility of storing large amounts of data make it easier to publish information that was previously limited to printed atlases. As García Mouton (2015) mentions about the Atlas Lingüístico de la Península Ibérica (ALPI) web page designs: "the tool allows you to preserve and offer everything that in a traditional atlas would be buried in a note and not even pass to the margins of a map."

Data represented in atlases are usually homogeneous since researchers trained in the survey application obtained data through an identical questionnaire with informants with concrete and coincident profiles in sociolinguistic characteristics and locations that cover a specific territory (Fernández, 2010; García Mouton, 2015). These conditions consolidate atlas materials as a corpus, as they are "a collection of actually occurring texts (either spoken or written), stored and accessed by means of computers, and useful for investigating language use" (Thornbury, 2006). ALEC materials are a collection of maps, audios, photographs, and images gathered through surveys applied in a location network in Colombian territory. These materials work as a model of Colombian Spanish from 1958 to 1978. The maps, photographs, and images from the ALEC were in paper format and the audios on open reel tapes. Therefore, it has been necessary to carry out several processes for the transition of these formats from analog to digital format. These procedures are crucial in the consolidation of materials in a corpus, since nowadays one of corpus main characteristic is the possibility of managing the data electronically on the web.

Information collected in linguistic atlases has as a central axis the relation of samples and data with their spatial location. Thus, the territory has a relation to the corpus data. In addition, as mentioned above, linguistic atlases are mainly focused on the diatopic dimension, as they are constructed based on geolinguistics methodology. As a result, we argue that geolinguistic corpora can be made of systematized and computerized materials of linguistic atlases. Therefore, we named our corpus "ALEC geolinguistic corpus."

In this article, we will start by describing the most important aspects of the ALEC research. Secondly, we will present the geolinguistic corpus composed of three main systems: the spatial database, the Geographic Information System (GIS) and the Web Atlas. Then we will present the Oral Corpus of the ALEC and its relation with the geolinguistic corpus. Finally, we will discuss future perspectives and suggest some conclusions about our work.

2. Atlas Lingüístico Etnográfico de Colombia (ALEC)

ALEC research's objective was to know firsthand the main characteristics of Colombian Spanish. Then it sought to establish differences and affinities with other varieties of Spanish, such as the peninsular variant and its relation with pre-Columbian languages (Buesa Oliver & Flórez, 1954). The investigation began with a pilot test in 1956 and for twenty-two years interviewers visited 264 towns belonging to twenty-eight of the thirty-two departments of the country. A total of 2234 informants and 23 interviewers participated. Six volumes summarized results, and each is 50 x 35 cm. They gather 1696 sheets, of which 1523 are linguistic, ethnographic or mixed maps. According to Flórez (1983), the linguistic maps register the names that informants gave to the questions in the surveys. The ethnographic maps show the areas or the geographic spread of "things" or objects of popular material life. As a complement, ICC published an alphabetical index, a manual with additional information on the locations, informants and interviewers, and a supplement to Volume III that includes spontaneous speech samples and two vinyl discs with games recordings and funeral songs from Caribbean and Pacific Colombian coasts.

In order to gather information, interviewers made direct surveys to the informants based on a 2000-item linguistic questionnaire on lexical, phonetic and morphosyntactic levels. The questionnaire and the volumes are divided into sixteen semantic fields related to life in the countryside and the daily life of the informants: human body, clothing, housing, food, family and life cycle, institutions and religious life, festivities and distractions, time and space, countryside--farming and vegetables, agriculture-related industries, livestock farming, domestic animals, wild animals, occupations and jobs, transport boats and fishing. Furthermore, there were questions concerning phonetic and grammatical aspects, specifically, onomatopoeia, the variation of vowels and consonants in various word contexts and grammatical phenomena (Bonilla et al., 2017). Due to the ethnographic interest of the research, interviewers collected speech records, photographs, and objects in a non-systematic way. Some of these were published in the supplement as games and funeral songs or as additional material of the maps. However, ICC filed the photographs and recording sessions for future researching.

According to Flórez (1983), selected informants were natives from the location or had lived there for most of their lives. A criterion was that informants were illiterate or had little school instruction. Regarding age, there was a preference for adults between 40 and 60 years old. After systematizing information contained in the ALEC's Manual (Flórez, 1983), we found that 32.9% of the informants were women and 67.1% were men. Informants were between the ages of 16 and 100, distributed as follows:

Ages	Informants percentage	Ages	Informants percentage

16 to 20 years old	0.45	61 to 70 years old	14.80
21 to 30 years old	5.42	71 to 80 years old	6.95
31 to 40 years old	18.64	81 to 90 years old	0.99
41 to 50 years old	28.02	91 to 100 years old	0.50
51 to 60 years old	24.10	More than 100 years old	0.09

Table 1. Age ranges of ALEC informants

In terms of academic training, 21.5% were illiterate and over 70% had not completed primary school. Most worked as farmers (32.7%), domestic work and house makers (27.9%), livestock farming (3.9%), merchants (3.8%), carpenters (2.6%) and fishers (2.3%).

Regarding the locations, although the main criterion was geometric--that is, they were at relatively equal distances from each other and the largest possible proportion of the territory was covered (Flórez, 1983)--Colombian geography forced the interviewers to select the locations as the fieldwork was carried out. Sometimes, taking into account chronological criteria, in relation to the time of the town's foundation; demographic, by the population numbers; or topological, climatic and socioeconomic criteria, looking for representation of geographic and cultural biodiversity.

3. Geolinguistic Corpus

We divide into three phases the design and development of an online tool that allows the transit of ALEC's printed data for subsequent consultation and analysis. The first phase (2016) consisted in the spatial database modeling, based on the linguistic atlas' particularities and the future needs of the project. Besides, we carried out the definition of the system requirements, possible users, and software specifications. It is important to point out that we decided to use free software. We digitized information contained in ALEC's Manual, Volume III, and the supplement in order to have enough data for testing and implementing the systems. In addition, the *Biblioteca Nacional de Colombia* scanned printed materials in high resolution in order to extract symbols, images, illustrations, scores, and texts.

3.1. ALEC Spatial Database

ALEC spatial database contains information on locations in terms of name, department, latitude, longitude, and year of foundation, height above sea level, population, economic activities, access roads, survey date, comments, and website. About informants, database relates the name, surname, age, sex, occupation, educational level, origin or place of birth, travels and parents and spouse's origin. Last, for interviewers, the database stores names, surnames, years working on the project, locations where the interviewers conducted surveys and the total number of surveys.

Initially, we designed spreadsheets for the linguistic maps and used them for the spatial database modeling. Nowadays it is possible to upload map information directly to the spatial database through GIS administration tools in a similar order. On the first sheet, we typed the information concerning the map title, scientific name, map number, sheet number, and references to other linguistic atlases, the semantic field and the name of its scanned image for subsequent location in the server. On the

second sheet, we assigned an identifier number to each variant and related it to a code given to the symbol that represents it on the map. It is important to distinguish that in the ALEC some variants are mapped and some are not. The mapped variants display a symbol on the map to indicate the place of occurrence, while the unmapped variants (which are less frequent) have a unique code from the location that is placed to the left of the map (Figure 1). The first letters of the name of the department and a number compose that location code. The third and fourth sheets linked the identifier number of the mapped or unmapped variant to the identifier of the location. Finally, the fifth page included additional information of the map (audio, text, photography, illustration, score) assigning an identifier that was then geographically referenced in the sixth sheet linking it to the location.

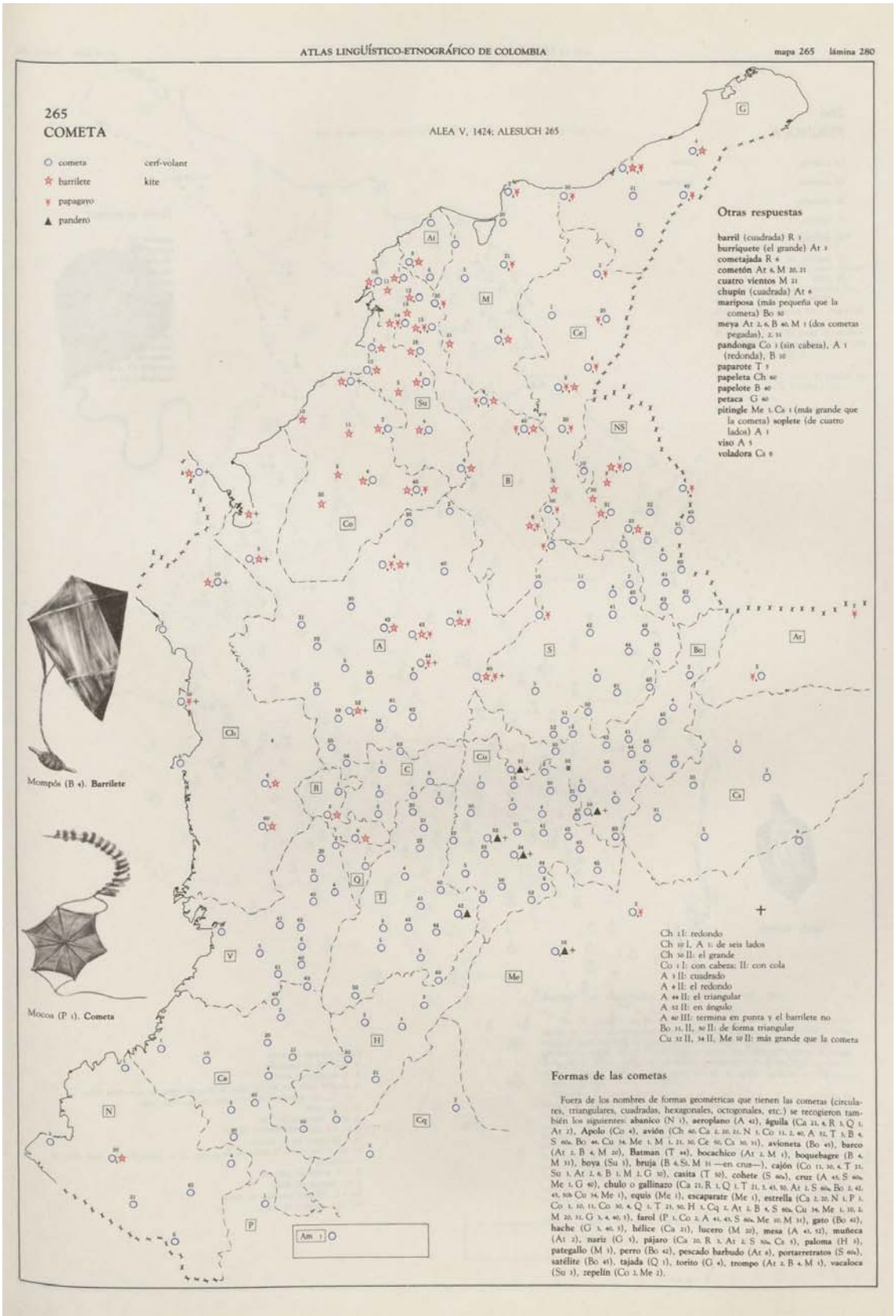

Figure 1. ALEC scanned map - Volume III, Sheet 280, Map 265, *Cometa*

We reproduced and implemented the spatial database model in a PostgreSQL database server with digitized information of the Manual, Volume III, and supplement. This database work as a pilot for the two systems developed during the second stage: First, ALEC GIS[1] with the query and administration tools that allows advanced queries, edition and addition of information, information crossings and statistical and geospatial analysis. Second, the ALEC Web[2] that allows simple queries of ALEC's maps, information contained in the manual regarding informants, locations and interviewers and, finally, access to a limited number of photographs and recordings of the research.

3.2. Geographic Information System (GIS) and Web Atlas

The second phase (2017) consisted of the design and programming tasks of ALEC GIS and ALEC Web administration and user interfaces through the implementation of the map server based on GeoServer. Geoserver works as a bridge between the geographic information of the spatial database and the user interface. The map server uses PostGIS as its spatial database connected to PostgreSQL and allows the creation of geographic services in order to use them through the web application, linking the final user with the information of the GIS and the Web Atlas. In addition, it is important to highlight that the geographic services fulfill the Open Geographic Consortium standards.

The implemented library for the interactive maps in both ALEC GIS and ALEC Web is Leaflet because, although it is light, it has all the elements that we need to represent online maps. Leaflet designing considers simplicity, performance, and usability, and it works efficiently with most desktop and mobile platforms. Also, Leaflet is a tool with many add-ons and a user-friendly API, and it is well documented and has a simple and readable source code.

Broadly speaking, ALEC GIS has different layers of access and we designed for what we consider to be advanced queries. ALEC GIS user interface (Figure 2) has six functions. First, one named "Seleccionar" (select) allows one to call up a linguistic, ethnographic or mixed map following any of the routes: Volume, Semantic Field, Map, or typing a word related to a title or any of the variants. Once the user visualizes a map or variant, he or she gets access to all related information such as audio, images or text, and he or she can draw a point on the specific location where the related information belongs. In the same menu, it is possible to consult the information related to interviewers, informants, and locations. The second functionality is named "Administrar capas" (layers management), where it is possible to call up multiple maps, variants, informants, interviewers and/or locations and add them to the map according to the researcher’s needs. The third function, called "Exportar" (export), allows accessing and downloading data in different formats (shapefile, kml, CSV), which can be implemented and analyzed on a desktop GIS or other spatial data management software. Fourth, we have the "queries" (Consultas) where it is possible to make crossings in database information. It exceeds mere representation on the map and makes it possible to use the metadata related to informants, locations, interviewers, and maps in order to delimit searches, as people might need to do. Regarding

[1] http://alec.caroycuervo.gov.co/siglogin

[2] http://alec.caroycuervo.gov.co/atlasweb

the fifth function, "Análisis" (analysis tools), we have been implementing intersection, buffer and disjoint tools until now. Lastly, we have the administrator function, which only appears when the user has the necessary permissions. It enables the ALEC's spatial database management, i.e. to add, edit, and delete maps or any other component.

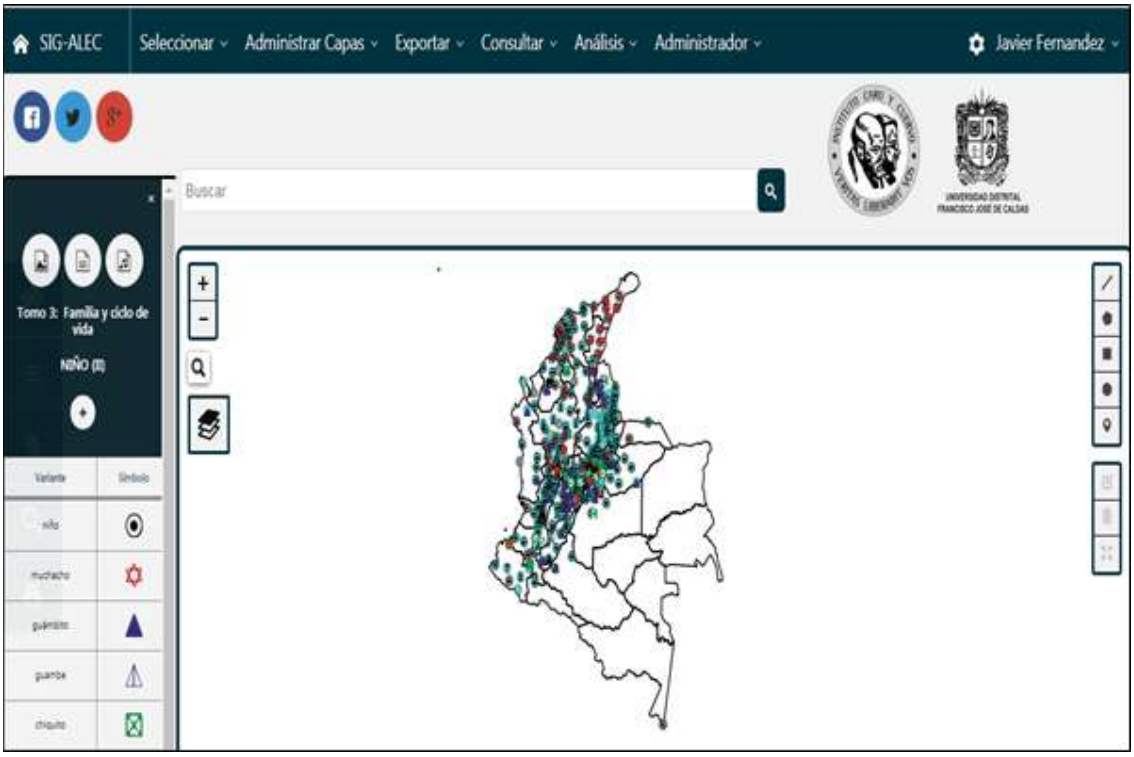

Figure 2. ALEC GIS user interface - Map query *Niño (II)*

On the other hand, ALEC Web (Figure 3) is a simpler tool completely open to the public. Among its capacities, ALEC Web allows map consultation--choosing Volume, Semantic Field, and Map, and simple searches by map and word. Likewise, it is possible to access map-related images, audio or texts. Information concerning interviewers, locations, and informants is consulted separately, each with a link from the side menu panel. Finally, it is possible to access a limited number of unpublished ALEC items, specifically more than 1000 photographs from the photographic registry and a sample of the oral corpus with a simple search by location from the sound registry.

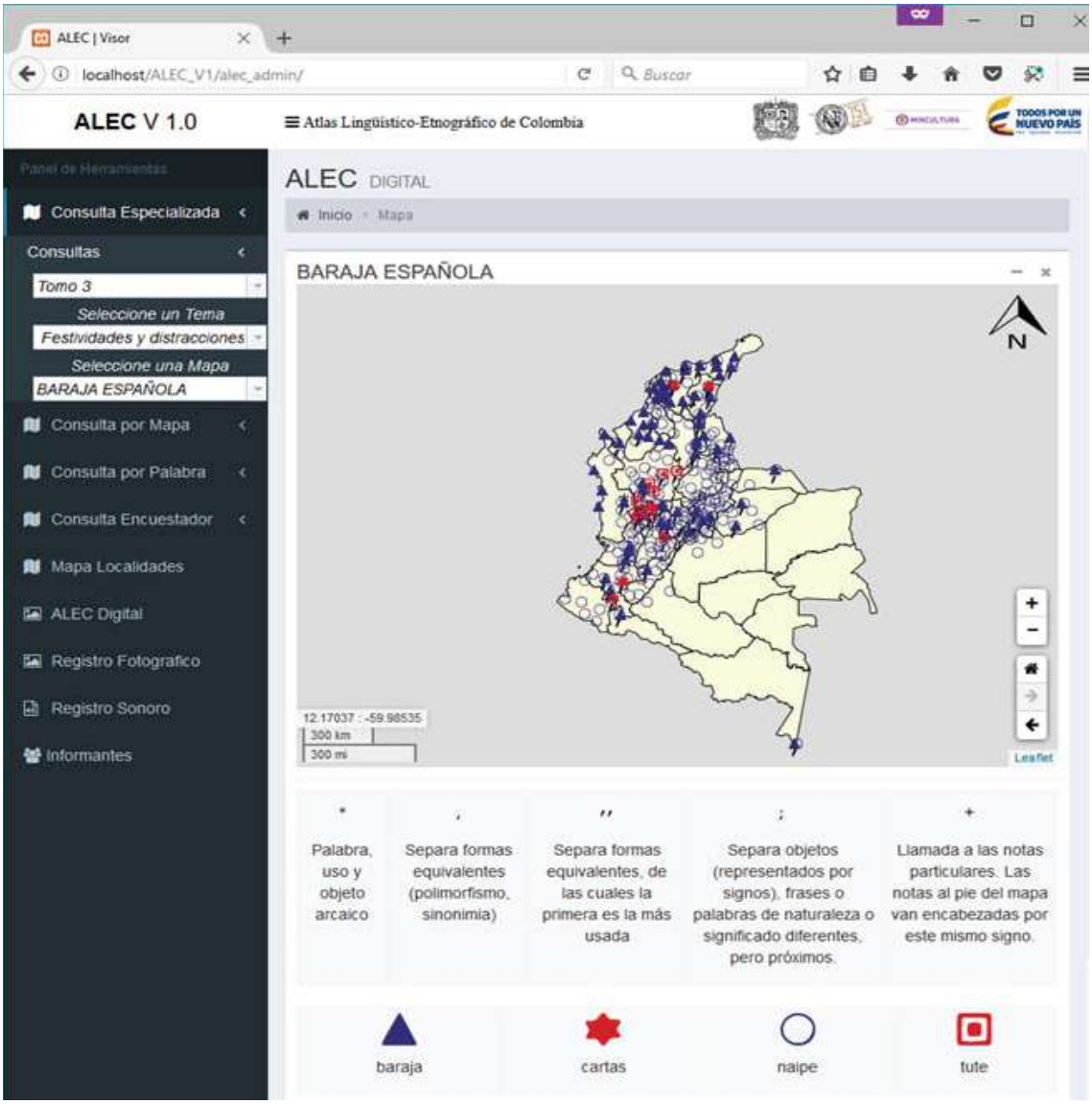

Figure 3. ALEC Web interface, Map query *Baraja Española*

The last phase of the project (2018-2019) consists of the evaluation and documentation of software developments and security tests for online publication. In addition, we are doing usability tests with external users to optimize the systems. Along the same lines, we are digitizing and uploading to the spatial database the rest of the ALEC volumes (I, II, IV, V, VI).

4. Oral Corpus

The ALEC oral corpus is composed of 765 audio sessions. Each session lasts between two and sixty minutes. As Buesa and Flórez mentioned (1954), researchers used tape recorders to record, along with romances, lullabies, carols, and some other traditional songs, sounds to summon or to scare animals, lexically concrete aspects (such as the coffee harvesting process), conversations, short stories

and local tales (p. 166). In this way, the corpus contains audios with phonetic surveys, songs, stories, folkloric samples and interviews related to the ALEC's semantic fields. The informants are not always the same ones who responded to the linguistic questionnaire. There are 1176 informants, normally men and women from 36 to 55 years of age with a few years of primary school or without schooling.

The construction of the ALEC's oral corpus began in 2013 along with the "*Español Hablado en Bogotá"* (EHB) and "*Habla Culta de Bogotá"* oral corpora (HCB). For its creation, we have developed several processes:

The first step (2013) was the conversion of the audio from open reel tapes to digital files. The "*Fundación Patrimonio Fílmico Colombiano*" digitized ALEC audios in .wav format with the same organization that they had before their digitization. That is to say, they are stored in boxes and marked according to the box number and the location where interviewers took the recording. Likewise, we made back-up copies to guarantee the conservation and facilitate work with the audio.

The second step (2014) was to define the general metadata about the samples and the informants. We subdivided metadata into four central areas: sample data (file name or i.d., location, audio quality and comments on quality); informant data (names, age, date and place of birth, educational level, etc.); session data (topics, description, date of the sample, informants, among others) and fragment data for the cut of the audio (start time, location, end time, first part, etc.).

The third step (2015) was to listen to the audio, complete the metadata table already defined, and to verify that the information and the marking coincided with the audio content. That is to say, a session could be located in several tapes and not coincide with the location with which it was marked. For this reason, we are currently figuring out which of the audio fragments that make up a single session are cut and pasted based on the metadata table completed by the researchers. This process will facilitate the consultation and transcription of the recordings. The fourth step (2016) was to review the metadata tables.

The fifth step (2017) is the orthographic transcription and the alignment of the audio snippets. To begin this process, we have transcribed and aligned twenty orthographic transcriptions of short sessions. Researchers are currently building a protocol for the orthographic transcription of the recordings. With this protocol, we will modify the test transcriptions and alignment of the audio will continue with the PRAAT[1] software.

A final step will be to associate the database of the oral corpus with the database of the geographic corpus. For this purpose, we are linking the corpus data to the identification numbers of locations, informants, interviewers, semantic fields and maps from the ALEC's spatial database.

In 2017, the Corpus and Computational research line started the project "*Corpus Lingüísticos del Instituto Caro y Cuervo* (CLICC)."[2] The project has developed a Content Management System (CMS) for

[1] http://www.fon.hum.uva.nl/praat/

[2] http://corpus.caroycuervo.gov.co

the storage, organization, consultation, and exploitation of the ICC's corpus. Corpus CMS is composed of a relational database, and administrative interface and the final user interface. The first corpora we are uploading are ALEC, EHB, and HCB oral corpora. The Corpus CMS allows for the consultation of audio and ALEC transcriptions from the metadata defined for consultation. Currently, CLICC has two types of searches: First, users have simple search possibilities by word, if there is an audio transcription, or by four main metadata (age, gender, date, and place of the survey) (See Figure 4). Second, the advanced search, for registered users only. It allows searches by diverse metadata (subjects, locations, interviewer, age and gender of the informants, etc.) and by word (See, figures 5 and 6).

Figure 4. Display of quick search results by word.

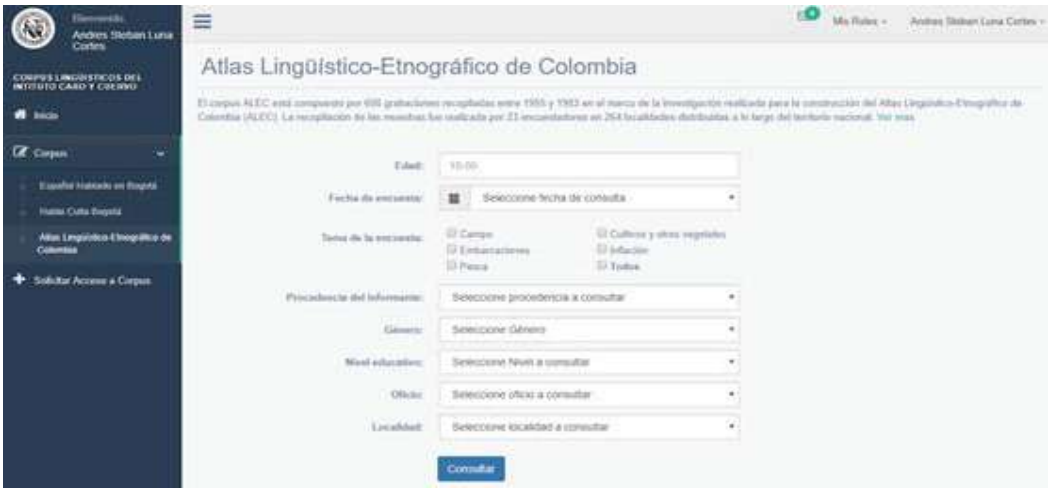

Figure 5. Advanced search display for registered users.

Advanced search enables the user to perform a larger quantity of tasks. It enables the query with several metadata and per word, visualization of audio alignment with its transcription, and authorizes the downloading of audio and transcriptions in different formats and allows the performance of

automatic morphosyntactic analysis with the TreeTagger[1] software when the audio transcriptions are available.

Figure 6. Audio display from the ALEC aligned with its respective transcription.

5. Projection and conclusions

The ALEC Geolinguistic Corpus' vision for the next five years (2018-2023) responds to three main areas related to the objectives of the Instituto Caro y Cuervo. It aims to promote teaching, new research, and the appreciation of the linguistic and cultural heritage of the Nation. 1. Corpus and Computer Linguistics: currently researchers are working on the morphosyntactic labeling of variants and texts of the ALEC in order to install in the ALEC GIS and ALEC Web the possibility of making queries through regular expressions. Regarding the Oral Corpus, the transcription of the audio snippets will continue for its subsequent automatic labeling thanks to the implementation of the TreeTagger tool. Additionally, we are linking spatial database data to the oral corpus metadata in order to integrate the databases. These procedures will facilitate the relationship of the audio with their geographical location.

Dialectology and new methods for statistical and geospatial analysis: to this date, we have made significant progress in dialectometric and statistical analyses with the information from the ALEC spatial database using tools available on the web such as Gabmap[2] and Diatech.[3] We expect to develop and integrate these tools in the GIS, so we can provide the academic community with various possibilities for online geolinguistic data analysis and feedback.

[1] http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/

[2] http://www.gabmap.nl/

[3] http://eudia.ehu.es/diatech/index

Updating of the ALEC data: there have been discussions about the validity period of the ALEC and about the lack of representation of the areas not visited at the time of the research (Guaviare, Guainía, Vichada, Vaupés and vast areas of the Amazon), where Spanish speakers now live. About this situation, the ICC, at the forefront of the Corpus and Computer Linguistics research line, is carrying out the first theoretical and methodological approaches to support the project for updating the ALEC's materials. For example, seeking to implement new techniques for collecting information based on ICTs, we are developing studies on Colombian dialects based on information from a Twitter corpus that allows the creation of lexical and morphosyntactic maps.

Moreover, it is important to mention that data users can approach samples and web tools as pedagogical means to teach Spanish and its diversity. In the coming years, we will hold different workshops at universities and libraries to teach about the management of these tools and their pedagogical possibilities. In addition, there are options for the creation of new multimedia educational applications based on Colombian Spanish for teaching Spanish as a foreign or second language.

Additionally, the advantages that the data provides for building dictionaries must be mentioned. In this regard, we can say that one of the students of the ICC's linguistics Master's program developed his final project on the creation of the template for an ALEC dictionary. On the other hand, the more progress we make with the transcription of the corpus, the more useful it will be for finding examples for dictionaries such as the Colombianisms dictionary that the Institute leads.

The change of formats (analogous to digital), the systematization of the materials, and the design and construction of the platforms has been a complex task of several years of development. It will serve to encourage linguistic researchers to take advantage of the ALEC's materials in different areas of knowledge. We hold that technological tools allow for the development of research from new perspectives and facilitate the automatic analysis of large amounts of data. In addition, the storage capacity, searching tools, and user interfaces facilitate the visualization and searching of data and materials in an agile and user-friendly way, which encourages its use and exploitation. Several ALEC materials that were previously hidden and forgotten are now available for different types of users and from anywhere with an internet connection.

References

Bonilla, J., Bejarano, D., Bernal Chávez, J., Rubio, R., & Llanos, A. 2017. Procesamiento informático de los materiales del Atlas Lingüístico-Etnográfico de Colombia: Sistema de Información Geográfica. En Instituto de Literatura y Lingüística, Estudios Lingüísticos. La Habana: Instituto de Literatura y Lingüística.

Buesa Oliver, T., & Flórez, L. 1954. El Atlas Lingüístico-Etnográfico de Colombia (ALEC) Cuestionario Preliminar. THESAURUS Boletín del Instituto Caro y Cuervo, X (1,2,3), 147-315.

Coseriu, E. 1956. Determinación y entorno. Dos problemas de una lingüística del hablar.

Flórez, L. 1983. Manual del Atlas Lingüístico-Etnográfico de Colombia. Bogotá: Instituto Caro y Cuervo.

Fernández, X. S. (2010). Entre el atlas lingüístico y el diccionario. Un diccionario de léxico tradicional a partir de los materiales del ALPI. In Metalexicografía variacional: diccionarios de regionalismos y diccionarios de especialidad (pp. 237-256). Servicio de Publicaciones.

García Mouton, P. 2015. Lengua y espacio. Revisión metodológica. En Hernández, H. & “Variación y diversidad lingüística. Hacia una teoría convergente.”

Instituto Caro y Cuervo. 1981-1983. Atlas Lingüístico-Etnográfico de Colombia. Bogotá: Instituto Caro y Cuervo.

Thornbury, S., & Slade, D. 2006. Conversation: From description to pedagogy. Cambridge University Press.