White Paper Report

Report ID: 103949

Application Number: HD5143211

Project Director: Jerid Francom (francojc@wfu.edu)

Institution: Wake Forest University

Reporting Period: 8/1/2011-12/31/2012

Report Due: 3/31/2013

Date Submitted: 3/30/2013

White paper

HD-51432-11
ACTIV-ES: a novel Spanish-language corpus for linguistic and cultural comparisons between communities of the Hispanic world
Jerid Francom
Wake Forest University
March 31, 2013

A. Project activities

Summary of the goals

Funding from this NEH Digital Humanities Start-up grant was applied to develop a novel Spanish-language corpus, ACTIV-ES. The aim of this electronic resource was to compile the first Spanish language corpus documenting linguistic and cultural expression from 'everyday' encounters for three linguistically, culturally, and geographically distinct communities of the Spanish-Speaking world— Argentina, Mexico, and Spain. The work conducted during the grant period (including a series of planning sessions, data collection and corpus processing steps, and pedagogic exploration) has made significant gains towards providing scholars, instructors, students, and other interested parties with a unique perspective on the rich cross-linguistic and cross-cultural patterns and themes in the Hispanic world. Grant activities have also highlighted unexpected challenges and breakthroughs that provide insight into further development of this resource in the near future.

In what follows I describe the proposed activities, identify omissions and changes to activities during the grant period and assess the overall progress towards the stated goals of this project.

Description of major activities

The work plan (as stated in the original grant proposal) included three main steps:

A. Data acquisition: to collect language data from the web
B. Corpus evaluation: to evaluate their ecological validity through psycholinguistic testing
C. Pedagogic exploration: to test the value of the archive by exploring practical pedagogical applications and tangible workflows

A. Data acquisition

To acquire and prepare the language data for the subsequent data evaluation and pedagogical implementation two main tasks were planned.

1. To target a number of Spanish-language newspapers and TV/film subtitles sites online and download language data via the existing software WGET (Unix) and SubDownloader.
2. To curate the language data and develop a language corpus by removing text artifacts (non-relevant language and HTML tags from newspaper sites and stripping closed-caption time-scores from TV/film subtitles), augmenting the corpus with key meta-information (i.e. part-of-speech information) and source attributes (i.e. creator, year, genre, *etc*.), and indexing the entirety of the corpus in a relational database (MySQL).

3. To aggregate the corpus into three word frequency lists (corresponding to each of the three Spanish-speaking communities).

(1) In collaboration with Dr. Mans Hulden (U Helsinki), it was decided that online newspapers would not provide a clear window into the 'everyday' speech of native speakers for two primary reasons: 1) the language of journalism is qualitatively distinct from language and linguistic forms used in common speech and 2) articles that appear in any given national newspaper, say in Mexico for example, may not be written by a Mexican, but rather be syndicated piece run across a number of Spanish-language publications making it difficult to establish the authenticity of the language for the current sample. Therefore the focus was restricted to TV/film subtitles.

There are two main practical hurdles to acquiring TV/film subtitle files from the web: 1) what are the relevant titles to acquire and 2) where/how can you acquire them with the least effort. A closer inspection of the subtitle repositories online revealed that opensubtitles.org was the largest and best organized site for systematic downloading as each file was also indexed by an Internet Movie Database ID (IMDbID). It was also noted that the proposed software SubDownloader would not be easily adapted for batch downloading.

Given these two parameters our initial approach (Data acquisition 1) was to leverage the connection between the IMDbIDs on the IMDb and opensubtitles.org by writing an original web-crawler using Python as the scripting language and the IMDbPY and BeautifulSoup packages to manage the downloading and meta-information cataloging of TV/film subtitles. Using plain text files listing all TV/films cataloged by the IMDb (ftp://ftp.fu-berlin.de/pub/misc/movies/database/) to produce a set of seed titles for each of the three countries of interest (Argentina (14,587), Mexico (19,632) and Spain (20,963), some 300+ total subtitle files linked with a rich set of metadata were acquired from opensubtitles.org.

A recent return to the data acquisition step outside the grant period to acquire more data for the corpus revealed that our initial approach (Data acquisition 1) was no longer feasible as the terms and conditions of web-crawling the opensubtitles.org site had changed; limiting downloads to 50 per day. Given this development, the primary investigator contacted the opensubtitles.org administrator with a direct request for all cataloged data and was provided a catalog dump of all TV/film files indexed as Spanish-language subtitles (Data acquisition 2). Using IMDbPY to retrieve TV/film meta-information, a subset of subtitle files corresponding to Argentine, Mexican and Spanish TV/films was created. Spot-checking of the files reveal inconsistencies corresponding to user error (incorrect title/data matches, incorrect language, *inter alia*). After removing erroneous files, this second data acquisition resulted in a total of 468 TV/film subtitles (AR 139, MX 140, and SP 189).

(2) In order to understand how the linguistic/cultural behavior for the acquired language samples matches/mismatches quantitatively, the data, already organized by country/year must be further categorized into parts of speech. This is an important step since many word forms in language are ambiguous between their lexical category: 'house (noun)' vs. '(to) house (verb)'. The same is true in Spanish ('casa (noun)' vs. ('casa (verb)'), and if this ambiguity is not addressed the frequency counts for any given word form are potentially inflated or conflated resulting in spurious word distribution statistics. However part of speech categorization is not a trivial task; on large datasets it is not physically feasible and perfecting automatic, machine-based algorithms is still very much an active area of investigation in computational linguistics. The current state-of-the-art in part of speech algorithms (POS taggers) is based on probabilistic approaches to machine-learning (Hidden Markov Models). This approach can yield very reliable tagging for language data that is well-groomed (i.e. free of non-linguistic artifacts, orthographic anomaly, etc.) and matched well (i.e. congruent in genre, register, etc.) for the training data set (a manually tagged set used to induce the probabilistic associations between words and their lexical categories in natural linguistic production contexts). However, deviation from the ideal situation potentially degrades the performance of the automatic tagger.

In this particular case the subtitle data in its original form is not well-groomed and may not match the training data set utilized in this project, ANCORA (Taulé, M., Martí, & Recasens, 2008). To prepare the subtitle files for part of speech tagging many artifacts inherent to subtitles had to be removed including timestamps and other artifacts (optical recognition software errors, orthographic transcription errors, etc.) that arise given subtitles are often created using OCR and/or created by individual subtitle transcribers. A Perl script was created to remove timestamp codes and related subtitle file syntax as well as to amend consistent orthographic OCR and transcriber errors. However, orthographic errors related to diacritics (meta character glyphs such as ´ and ¨ proved more difficult to correct. For example, it is possible to identify a form that needs diacritic restoration in a case where the word does not exist without the diacritic, such as 'anden'(?) vs. 'andén'(platform) or 'comun'(?) vs. 'común'(common) but difficult to know when a word that appears as 'que'(that), a relative pronoun, should appear as the question word 'qué'(what), or other such cases 'mas'(but) vs. 'más'(more), etc. Using the a finite-state language toolkit (foma) (Hulden, 2009) developed by my colleague and collaborator, Dr. Mans Hulden, we leveraged our existing morphological analyzer (SpanMorph) (Francom, Hulden & Tubino-Blanco, 2010) to identify all words as either possible or impossible words of Spanish (for example 'comun' would be identified as a non-word) and also extended our coverage of possible word cases that were incorrectly marked using simple heuristics (such as a 'que'(that) preceded by an interrogative symbol "¿" should be marked as 'qué'(what)). After applying these conservative measures to improve the orthographic quality of the text, the subtitle data was automatically tagged using the Hunpos trigram tagger (Halácsy, Korean & Oravecz, 2007) trained on the hand-tagged data from the ANCORA corpus. Although there is an open question regarding the overall reliability of the probabilistic tagger given there is a mismatch (genre/register) between the data sets (movie dialogues vs. newspaper and newswire text), little evidence has explored the tangible effects of cross-genre/register mismatches. Our intuition was that the tagging reliability for key lexical categories in this project (nouns, verbs, adjective and adverbs) was good and that those lexical categories most potentially affected (including subordinate markers) were not crucial to the preliminary work carried out in this grant period. However, on inspection of the tagging output consistent errors appeared with the 'voseo' verbal paradigm in the Argentine sub-corpus --a logical consequence given the Hunpos trigram tagger was trained on a dataset in which no instances of the 'voseo' paradigm appear. These inconsistencies were not viewed to be critical for the subsequent stage in this project (data evaluation) as only common, co-occurring words would only be selected (instances of the 'voseo' paradigm do not appear in Mexican nor Spanish dialects). Nevertheless, register and dialect inconsistencies demonstrate the limitations of this approach and warrant more investigation --an area of work that Dr. Hulden and the primary investigator are exploring outside of the grant period. See Section E. Continuation of the project).

After applying the subtitle cleaning, orthographic changes, diacritic restoration and part of speech tagging steps to the corpus data the corpus files were renamed with the most relevant information (i.e. language, country, year, title, type, genre, IMDbID). Although the initial plan was to add the corpus to a relational database, it was decided that given the immediate needs of the project the corpus data could be more easily retrieved and processed using plain text files. A total token count registers: 4,693,439 (AR: 1,432,648; MX: 1,721,970; SP: 1,538,821).

(3) From these plain text files, three wordlists corresponding to Argentina, Mexico and Spain were created (using an original Python script implementing methods in the Natural Language Toolkit (NLTK)) each containing word (word), word frequency (freq), part of speech (pos) and number of files a word appeared (dispersion, docs) information. In this process, a decision was made to remove non-alphabetic entries (numerals, symbols and punctuation) as the primary goal is to focus on word use in this project. The unique words in each list currently total: AR 52,358; MX 56,835; and SP 59,501 with a total unique word count of 112,722. The sum word frequency in each list totals: AR 1,087,771; MX 1,331,509; SP 1,159,907 with a total word count of 3,579,187. This number is less than expected at the onset of the project, but still a sizable sample. The hope was that this sample, given the quality of the sample, would constitute enough data to provide reliable lexical measures for the subsequent behavioral evaluation of the corpus.

B. Corpus evaluation

A corpus is more than a mere collection of texts, rather it is a planned attempt to capture language data that models the linguistic environment of a target community. A unique component of this grant proposal was to assess the extent to which word distributions contained in the three sub-corpora reflect the exposure to language of native speakers from these regions. To this end, we aimed to conduct a series of in-field experiments to gauge the Word Frequency Effect (the well-documented observation that the frequency of exposure to words in the linguistic environment can be detected using a number of tasks gauging lexical processing behavior) of native speakers from these target locales.

Our approach involved two main steps:

1. To merge the wordlists, select a subset of words co-occurring in all three sub-corpora most contrasting for word frequency counts, and to develop a battery of psychological experiments aimed at assessing the correlation between corpus frequency and measured psycholinguistic behavior
2. To conduct in-field testing in Mexico and Spain using the experimental software EPrime and remote testing via Amazon's Mechanical Turk for the Argentine native speakers.
3. To gauge the extent to which the lexical processing behavior of native speakers from Argentina, Mexico, and Spain is predicted by sub-corpora word frequency counts through statistical evaluation.

(1) The wordlists were merged and only the commonly occurring words were selected. I employed an R script (R-project, 2010). This script joins all three lists by the column 'word'. The list of word types combined totaled 18,575 words.

The initial plan for selecting the materials for the experiment proposed to target a set of frequency matched and frequency mismatched words. The idea behind this approach was that matched words would help in the assessment of a primary frequency effect whereas the mismatched materials would highlight target locale exposure and sub-corpus frequency differences --an approach the PI and collaborator Dr. Adam Ussishkin had previously taken in a psychological evaluation of corpus frequency for American and British English corpora with similar goals as the current corpus evaluation.

After some debate an alternative approach was selected which focused on controlling for and filtering words based on a set of known lexical processing factors; word length, neighborhood density, and dispersion (ideally more information would be available to control for imageability and other variables. A new resource, not available at the time of materials selection, EsPal, shows potential for just such needs.). These heuristic filters were applied isolating frequency as the most salient contrastive variable to avoid the potential for unexpected confounding effects. Furthermore, this approach facilitated the use of one master materials list in all three target locales without modification and avoided potential issues with inter-corpus tagging inconsistencies (i.e. the 'voseo' paradigm in Argentine Spanish, *inter alia*). Adopting a matched/mismatched approach, as initially proposed, would have created the necessity to develop and vet new materials for each of the three experiments --a less-than-desirable outcome.

In addition, we selected only nouns, removing all other words and excluded nouns that were homographs with some other part-of-speech form (ex. house (n)/ house (v)). This step, facilitated by the inclusion of the part-of-speech category in the corpus data, ensured that the counts for any given word would not be conflated with the psychological status of other analogous word forms. All and all, these rigorous steps produced a list of 240 words to be used as materials for the behavioral experiment.

(2) The primary investigator had initially envisioned employing the well-known software ePrime to conduct the lexical processing experiments in the field. The downside is that ePrime must be installed on local machines and participants in the experiment must come individually to a laboratory and complete the task. However, the primary investigator became aware of new developments in experimental software that provides access to the experiment via the web (Ibex/ Webspr). This permits the experimenter much freedom in terms of how participants can be recruited and can ultimately complete the task. After some cross-testing using data from a previous experiment, the Ibex proved to be a reliable tool in conducting the proposed lexical processing tasks online (to be described). The results from the online participants (matched for age) achieved significance in the same categories as the laboratory experiment run with EPrime. This development facilitated the process of collecting data in the field in many ways: 1) by liberating the platform (EPrime was Windows only), 2) by making it possible to run many participants on the experiment at one time (with ePrime the experiments would have needed to take place on licensed machines) and 3) Ibex is free and open-source. Furthermore, the need to explore and implement the same tasks in Amazon's Mechanical Turk was avoided as during the grant period a professional contact was made with colleagues at the University of Buenos Aires, Argentina. Therefore, the same methods and experimental approach was extended to round out the three target populations --a welcomed amendment to the original grant proposal.

Using the selected materials, two experimental tasks aimed at gauging native speaker's knowledge of the relative distribution of words in their community were constructed, a lexical decision and subjective word familiarity task. In the lexical decision task, a visual word recognition task, participants are asked to make speeded judgments about the status of a string of letters; "does the string form a word or not". Participants' reaction times (RT) and accuracy (ACC) are both recorded for analysis.

This type of task taps into a speaker's subconscious (implicit) knowledge of the experience with words in their everyday environment. The subjective word familiarity rating task aimed to contrast the implicit knowledge of speakers with their conscious or explicit knowledge of word frequency. In this task participants provide a rating for each word on a scale (1-7) indicating how often they encounter this word in their everyday lives ("Various time a day", "Once a day", "Once a week" *etc.*) using methods adapted from Balota et al. (2001). In addition to these primary language tasks, participants were also asked to fill out a brief language background survey to use in analysis. In all the experiment was designed to take between 15 to 20 minutes to complete.

For each of the three regions, I oversaw the administration of the experiments and collaborated with a local colleague who helped oversee the on-the-ground planning and recruitment for the experiment. In Argentina, Dr. Julieta Fumagalli coordinated the recruitment of 118 participants. In Mexico, Professor Julio Serrano help me recruit 82 participants and in Spain, Dr. Mikel Santesteban facilitated the recruitment of 80 participants. In all locales participants were eager to participate and data collection wrapped up in 2-3 days. Each participant was compensated for their time ($5 USD) and most participants completed the entire experiment in under 15 minutes.

(3) The analysis of the data from these experiments suggests that the sub-corpora do approximate the everyday experience of the corresponding community according to reaction time and rating response measures. The data was submit to standard preparation for parametric tests including participant exclusion, outlier trimming and response transformation. The resulting data set was analyzed for basic correlation strength using Pearson's r and then subsequently submit to more robust statistical assessment in a set of linear mixed-effect models (Baayen & Milin, 2010) with model comparison contrasts.

Findings from both the word recognition and word familiarity tasks show that corpus word frequencies for the Argentine, Mexican, and Spanish sub-corpora generally correlate significantly with the behavior of their respective populations. One notable finding not consistent with this trend is the Subjective word familiarity ratings for Mexican participants. In this condition the Mexican sub-corpus word frequencies do not outperform the Argentine sub-corpus. The most probable explanation for this finding is that the Mexican sub-corpus is somewhat skewed towards TV/films from the 1930 to 1960 (The 'golden age' of Mexican cinema). Looking at corpus word dispersions for the three sub-corpora, all native populations show stronger correlations with their respective sub-corpus, even for the Mexican native speakers.

Overall these data suggest that word frequency and word dispersion information extracted from the ACTIV-ES corpus represent a valid snapshot of the active lexicon of the Spanish language in these three regions of the Hispanic world.

C. Pedagogic exploration

Given the promising results from the experimental evidence presented in the previous section, the next proposed step in this project was to explore the pedagogical applications of the ACTIV-ES corpus. From the PI's professional teaching experience and from the academic literature (Davies & Face, 2006) it is clear that it is too often the case that language used in formal instruction (textbooks, guides, workbooks, *etc*.) is developed without the consultation of empirical language sources. Furthermore, those language resources that do exist tend to be based on formal, mostly written, language and are not built to provide straight forward comparisons between sub-varieties of the Spanish language. The ability to identify language patterns in informal, everyday language and compare these patterns between regions of the Hispanic world would clearly be an asset to the second-language classroom instructor. To this end, two steps were initially proposed to address this issue:

1. To develop modules at the beginner, intermediate, and advanced levels based on the linguistic patterns and cultural themes uniquely found in ACTIV-ES to address typical second-language classroom questions.
2. To pilot-test the teaching modules in Wake Forest University classrooms in Fall 2012 and provide post-course questionnaires to provide student and instructor feedback on their benefits for future enhancement.

(1) In order to facilitate module development, the ACTIV-ES wordlists, previously developed in the data acquisition stage, were statistically evaluated using an R script to identify 'core' and 'dialect-specific' vocabulary. Using knowledge gained in the corpus evaluation stage, word dispersion counts were used as the primary variable for evaluating and identifying 'core' vs. 'specific' vocabulary. A regression model was created which aimed to plot the overall correlation between the ACTIV-ES corpus word dispersion counts and individual sub-corpus counts. Words falling within +/- 1 standard deviation (SD) of the linear trend line were identified as 'core' vocabulary -- those words greater than 1 SD were marked as 'most indicative' of the sub-corpus and those words less than 1 SD were marked as 'least indicative' for an example of the top 'core' and 'dialect-specific' words identified). Given the ACTIV-ES corpus wordlists are also part-of-speech tagged subsequent evaluation could target particular grammatical categories and generate corresponding 'core' and 'dialect-specific' words for nouns, verbs, adjectives, *etc*. An exciting finding from these statistically produced vocabulary lists is they are congruent with native speaker recognition of common and dialect particular forms --providing promising feedback for the methods.

However, one current shortcoming of the 'core' and 'specialized' vocabulary lists is that they target specific word forms (i.e. *soy*, *eres*, *es*, *somos*, *son*) and cannot provide more general information about the behavior of a lexical root or 'lemma' (i.e. *ser*). Future work will aim to include this information in order to be able to tease apart word form frequencies and lemma frequencies.

Other approaches explored to retrieve useful word-level and phrase-level information from the corpus included: 1) word choice searches, 2) Keyword in context searches (KWIC), and 3) association strength measures. Word choice searches can be performed easily on the entire wordlist by using regular expressions ('wild-card searches') for word sets. Given the classification information previously mentioned ('most indicative', 'core', 'least-indicative') was associated with each word, word frequency differences can be interpreted in a straight-forward manner. KWIC searches are useful in identifying the words neighboring a particular word of interest highlighting the real-world usage of a term. To implement a KWIC search a script was written in Python using the Natural Language Toolkit (NLTK). Finally, word association strength measures were explored to identify a target word's most closely connected set of words. For example, in English the words 'white' and 'house' are more closely associated that 'black' and 'house' --the underlying assumption is that when 'house' appears it is most likely to appear with 'white' more than with other words, such as 'black'. This metric can be obtained through the NLTK as well. And a basic Python script was developed to implement this measure for words of interest.

A much valued component of language learning and training depends on multi-word sequences, phrases, idioms, *etc*. of which we can target through KWIC and association measure searches. Although accessible in the ACTIV-ES corpus, a shortcoming of the current project is that these linguistic units have not been directly vetted and are potentially less-than-representative of the target locale's usage given the overall size of the corpus. (The uniqueness of multi-word units increases exponentially, which decreases overall counts inversely --obscuring the difference between typical and non-typical linguistic units.) Future project goals aim to amend this situation by adding more data and conducting a new series of in-field behavioral experiments targeting multi-word sequences.

The aim, then, in collaboration with Dr. Luis González, here at Wake Forest University was to envision how these 'core' and 'specialized' vocabulary could be employed in the classroom setting to enhance language instruction. It was envisioned that 1) the 'core' materials could serve as checklist for need-to-know words at the basic languages level (beginner-intermediate) and that 'variety-specific' materials could be employed in training/preparation for study abroad students going to Argentina, Mexico, or Spain (advanced). 2) Word choice and 3) KWIC searches were identified as useful to spot-check certain grammatical patterns in which known variation between dialects is unclear (such as 'por' vs. 'para', 'quizá vs. 'quizás, *etc*.) and a vocabulary word appears in real-world context (advanced). 4) Finally, to gauge general linguistic and cultural tendencies for word relationships, association strength measures were selected (beginner-intermediate-advanced). Each of these more targeted measures can be aimed at the general corpus, or if desired, applied to a subset of the corpus (i.e. Argentina, or Spain, or Argentina and Mexico, *etc*.). In sum, linguistic and cultural information of the general 'core' sort was in line with more beginner and intermediate level instruction whereas dialect-specific vocabulary lends itself better for courses where linguistic and cultural diversity of the Spanish-speaking world is highlighted.

(2) The next and final proposed step was to pilot the teaching modules in the PI's Grammar and Composition course in the Fall 2012 semester. However this step was not realized for two reasons. 1) The PI's course in the Fall 2012 semester was cancelled due to under-enrollment (this can happen if the time slot conflicts with other courses or is at an inconvenient time for some other reason). Therefore little time was available to line-up another class and adjust the materials. 2) On the inspection of the data and they types of information it could convey it was noted that an upper-level Grammar and Composition course would not be the most appropriate course to pilot the corresponding dialect-specific 'everyday' language material that ACTIV-ES provides given the focus is on written discourse and little attention is placed on inter-dialect variation. In part due to this realization and internal curriculum developments in the Department a new course is being proposed to address oral communication at the advanced level. This type of course will prove ideal to showcase the lexical, phrasal and discourse properties of spoken Spanish and highlight cultural patterns across the Spanish-speaking world.

B. Accomplishments

Project accomplishments

The main goals of this project were accomplished. The PI and project team developed and evaluated a quality corpus resource for the linguistic and cultural investigation for three diverse communities of the Spanish-speaking world, Argentina, Mexico, and Spain. The development of the corpus resource (the largest for dialogue-based language in Spanish), its in-field validation (only multi-variety psychologically validated resource), and instructional evaluation serve as a proof-of-concept for these methods which can be applied to other languages with robust film industries in a similar manner. Furthermore, all of the software used in the development, validation, and exploration is freely available, much of it open-source --including experimental and statistical software.

Shortcomings

There is room for improvement. As far as the corpus is concerned, the size of the corpus is humble compared to other corpus resources for other languages (although this are based primarily on written sources). For word-level investigation the size of ACTIV-ES is not an issue, however, future plans for this resource include the exploration and validation of multi-word units and accurate frequency and dispersion measures for these units will necessitate more data. Furthermore, the data is annotated for part-of-speech (itself showing room for improvement, especially for the Argentine sub-corpus data), which was key to both the validation and instructional components of this project but it was noted that categorizing words by their lemma would be beneficial to enable word group frequencies. Finally, this project did not have the chance to fully employ and evaluate the effectiveness of these materials on instruction. It is hoped that through new curriculum changes that in-class assessment will be possible.

Dissemination

Journal articles

Currently I am drafting a manuscript for publication ("ACTIV-ES: a Spanish language corpus for three linguistically, culturally, and geographically distinct communities of the Hispanic world) to document the development of the corpus and potential for research and teaching. I also plan to draft another article (title TBA) for publication dealing with the particular methodology for evaluation corpora through in-field psychological testing.

Conference proceedings

Activities from this grant have produced one publication in conference proceedings and two more publications are under-review. In the Spring 2012 the PI and collaborator Dr. Mans Hulden presented/published a paper entitled "Boosting Statistical Tagger Accuracy with Simple Rule-Based Grammars" at that year's Language Resource and Evaluation conference in Istanbul, Turkey in which we discuss strategies for improving part-of-speech tagging using the combination of statistical and rule-based methods which shows promise in stemming the problem encountered in this project dealing with training data and target data genre/register mismatches.

Dr. Hulden and the PI are awaiting response from two other computational linguistics conferences. We await confirmation at this year's Statistical Language and Speech Processing conference (SLSP) in Tarragona, Spain. The manuscript for publication entitled "Diacritic restoration via part-of-speech tags" outlines our research on diacritic restoration; restoring orthographically correct accents to latinized text. This paper highlights the importance of considering text genre in diacritic restoration as well as for many other language processing tasks. The second conference, Nordic Conference on Computational Linguistics (Nodalida) 2013, also in collaboration with Dr. Hulden entitled "Spanish diacritic restoration: a JavaScript implementation of hybrid approaches to error detection and correction" involves a software implementation of the diacritic error detection and correction methods explored in the SLSP paper, but with a focus on real-world application and utility in existing technologies. This software can be employed in cross-platform desktop and mobile technologies and converts latinized text into complete and correctly diacritized forms.

Conference talks
I have given talks at two corpus linguistics conferences based off of the work during the grant period focusing design of the corpus and exploring it's pedagogical uses. At last year's Arizona Linguistics Circle the PI presented a talk entitled, "The ACTIV-ES corpus: Moving towards cross-dialectal, ecologically validated language samples, and a more inclusive language science." in Tucson, AZ. Oct 2012, and at this year's American Association for Corpus Linguistics 2013, San Diego, CA. the talk "Word on the street: using the cross-dialectal corpus of 'everyday' language ACTIV-ES to highlight 'core' and 'specialized' language forms of the Spanish-speaking world." was presented. Both of these talks aimed to address more directly the novelty of the corpus creation process elaborated in the ACTIV-ES project and explore the kinds of research that can only be conducted with contemporary, conversation-based, cross-varietal language samples.

Two talks have also been given on the campus of Wake Forest University with the goal to raise awareness about digital humanities and ways that digital investigation can enhance teaching and scholarship. The first talk, organized by the WFU Digital Humanities Initiative, was entitled "What is Digital Humanities: a look at teaching and research applications" and the second, part of a series entitled "New Horizons in the Humanities: Access, Perspectives, and Collaboration at the Intersection of the Humanities and the Digital" was organized by the Wake Forest Humanities Institute. Both talks were well attended and have been promising in encouraging discussion about the digital humanities and how such work plays out on liberal arts campuses.

Social media
I have periodically promoted work and media coverage through Twitter.

C. Audiences
The audiences of activities conducted so far on this project have been academic, primarily scholars and instructors --although some exposure has come through media coverage through Wake Forest University (WFU news "Ready, set, speak — Spanish Film subtitles may be key to localizing language prep"). Through academic publications, professional talks, on-campus seminar meetings, and social media I have started by fanning interest with faculty and colleagues in computational linguistics, cultural studies and language instruction. It is hoped that a subsequent round of

funding will enable more broad-based dissemination of the ACTIV-ES resource to students and the public through a web-based interface.

D. Evaluation

Internal evaluation

(See Project activities. (B)).

Public response

Public response to the project activities has been limited as it is still an internal resource. Feedback from publications, talks and media coverage, however, suggest that there is an inherent interest in the project and the stated goals to enable robust, cross-linguistic and cross-cultural exploration of the Spanish-speaking world.

E. Continuation of the project

I'm currently looking forward to new funding to increasing the size of the corpus, evaluating a larger range of linguistic units and creating a user- friendly web interface for researchers, students and the general public.

I am currently soliciting WFU internal funds for the summer 2013 to meet with my colleague Dr. Julieta Fumagalli over the course of 7 days to run a new round of psycholinguistic experiments to explore the psychological reality of individual word and word phrases (multi-word sequences) extracted from the Argentine component of the ACTIV-ES corpus. The results from these experiments will build on earlier findings (based on individual words) and serve as a proof-of-concept to build a strong empirical foundation for an upcoming set of grant proposals (January 2014) to the National Endowment for the Humanities (Implementation Grant) and the National Science Foundation (Behavioral and Cognitive Sciences).

In these solicitations I plan to request funds to: 1) add more data to the corpus, 2) conduct a wider in-field evaluation in Argentina, Mexico and Spain, 3) to contract a web-developer to facilitate the design and implementation of a robust searching tool for the ACTIV-ES corpus, and 4) to implement teaching modules through through standard instructor-driven pedagogic materials (summary sheets, workbooks, *etc*.) and student-driven web-based interface with the corpus directly.

New collaborative partnerships

The activities conducted during the grant period have spurred new collaborative projects with my collaborator Dr. Mans Hulden. Our work on the ACTIV-ES corpus development and annotation has lead to a number of fruitful spin-off projects in computational linguistics. The opportunity to work with colleagues abroad has also been positive and I look forward to future collaboration with colleagues Dr. Mikel Santesteban at the Universidad del País Vasco (Vitoria-Gasteiz, Spain), Prof. Julio

Serrano at the Universidad Nacional Autónoma de México (Mexico City, Mexico) and Dr. Julieta Fumagalli at the Universidad de Buenos Aires (Buenos Aires, Argentina). Also, this Digital Humanities Start-up grant has lead to my involvement with the Humanities Institute, Digital Humanities and the WISE conference for study abroad efforts on campus here at Wake Forest University.

F. Long term impact

It's difficult to estimate the long term impact of the project activities. However, the feedback from publications, conferences, and informal discussion of the ACTIV-ES project suggest there is wide support and interest for the type of work involved here and the potential to address questions in a number of inter-related fields. Looking forward to continuing work, much of the impact for students and the public will depend on interfaces to the ACTIV-ES corpus that are web-based and intuitive. At this moment manipulating the corpus to provide questions to research, teaching, and general questions about the linguistic and cultural nature of the Spanish-speaking world requires technical knowledge of computing tools. A web-based interface will go a long way to facilitating the involvement students, instructors and researchers that are not experts in technology. Colleagues, students and administrators seem eager to see and interact with the future web-implementation and I am eager to see what questions they will entertain with this future resource.

G. Grant products

Future publication

A manuscript is underway that will document the corpus development and the novel psychological evaluation of the corpus data. I hope to have the manuscript submitted within the semester. In conjunction with the acceptance of the paper I plan to publicly release the ACTIV-ES word lists for download. I also have plans to discuss the pedagogic implementation of the data from ACTIV-ES as a way to supplement current practices in language instruction in order to better prepare students for real-world encounters in study abroad programs (in particular in Argentina, Mexico, and Spain).

Distribution plans

I have begun to list and post software developed during these grant activities on my homepage (http://francojc.wordpress.com/activ/resources/) but I am considering using Google Code or GitHub as a repository to house these resources, and potentially the ACTIV-ES wordlists as well. These services provide more robust version tracking and allows community support and development.

H. References

Baayen, H., & Milin, P. (2010). Analyzing reaction times. *International Journal of Psychological Research*, *3*(2), 12–28.

Balota, D. A., Pilotti, M., & Cortese, M. J. (2001). Subjective frequency estimates for 2,938 monosyllabic words. *Memory and Cognition*, *29*(4), 639–647.

Davies, M., & Face, T. L. (2006). Vocabulary Coverage in Spanish Textbooks: How Representative is It? *Selected Proceedings of the 9th Hispanic Linguistics Symposium* (pp. 132–143).

Duchon, A. Duchon, A., Perea, M., Sebastián-Gallés, N., Martí, A., Carreiras, M.(in press). EsPal: One-stop Shopping for Spanish Word Properties. *Behavior Research Methods.*

Francom, J., Hulden, M. & Tubino-Blanco, M. (2010) "SpanMorph: an open-source morphological grammar for Spanish". Universidad de Sonora. Presentation at the *Eleventh Encuentro Internacional de Lingüística en el Noroeste*.

Halácsy, P., Kornai, A., & Oravecz, C. (2007). HunPos: an open source trigram tagger. *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions* (pp. 209–212). Association for Computational Linguistics.

Hulden, M. (2009). Foma: a Finite-State Compiler and Library. *EACL 2009 Proceedings* (pp. 29-32).

R Development Core Team. (2010). R: A language and environment for statistical computing. *Foundation for Statistical Computing.*

Taulé, M., Martí, M. A., & Recasens, M. (2008). Ancora: Multilevel annotated corpora for catalan and spanish. *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC-2008).*