

# ACTIV-ES: a comparable, cross-dialect corpus of ‘everyday’ Spanish from Argentina, Mexico, and Spain

Jerid Francom, Mans Hulden, Adam Ussishkin

Wake Forest University, University of Helsinki, University of Arizona,  
Winston-Salem, NC. USA, Helsinki, Finland. Tucson, AZ. USA  
francojc@wfu.edu, mhulden@u.arizona.edu, ussishki@u.arizona.edu

## Abstract

Corpus resources for Spanish have proved invaluable for a number of applications in a wide variety of fields. However, a majority of resources are based on formal, written language and/or are not built to model language variation between varieties of the Spanish language, despite the fact that most language in ‘everyday’ use is informal/ dialogue-based and shows rich regional variation. This paper outlines the development and evaluation of the ACTIV-ES corpus, a first-step to produce a comparable, cross-dialect corpus representative of the ‘everyday’ language of various regions of the Spanish-speaking world.

**Keywords:** Spanish, corpora, dialects

## 1. Introduction

Corpora have been key to the continuing success of research in natural language processing (Manning, 1999) and have been exploited for work in the social (Biber, 1998) and behavioral (Burgess and Livesay, 1998) sciences, and more recently the humanities (Google Books (Michel, 2010)). However, the majority of corpus resources in general, and in particular for Spanish, are by and large based on formal language (both written and spoken) and are not built to capture language variation between dialects despite the fact that the majority of a typical speaker’s linguistic environment is informal language and demonstrates rich regional variation. Thus, many of the widely available corpora and other language resources are potentially register/dialect incongruent with a motivating body of research aimed at investigating language in ‘everyday’ environments.

This is, of course, in large part based on the ease of acquiring born-digital resources or widely transcribed spoken language (e.g. parliamentary speeches) and the corresponding difficulty in acquiring spontaneous, informal language — both written and spoken. Pioneering research by Brysbaert and New (2009) has demonstrated that informal, dialogue-based language data from TV/film subtitles can be easily extracted from the web and, through psycholinguistic evaluation, proves to be more in line with the linguistic experience of the typical speaker than comparable language resources based on formal language (Brysbaert et al., 2011). This approach provides a potential avenue for stemming issues in a growing number of areas of language research where representative samples of informal, dialogue-based language are in high demand.

The current paper discusses the creation and evaluation of the ACTIV-ES corpus: a cross-dialectal record of the ‘everyday’ language use of Spanish speakers from three regions of the Spanish-speaking world. Building on previous

methods, this corpus is based on TV/film subtitles provided from an online repository but extends previous resources in two ways: 1) we leverage meta-data from the Internet Movie Database (IMDb) to identify and extract original-version Spanish language transcripts for Argentine, Mexican, and Spanish productions and 2) we apply a series of techniques to normalize, part-of-speech annotate, and aggregate the transcript texts into 1:5-word n-gram lists.

In what follows, we outline the methods for identifying and extracting relevant data sources, challenges in normalizing orthographic forms and part-of-speech tagging the corpus, and assessments of the lexical distributions of the corpus and sub-corpora as well as an initial evaluation of the representativeness of the sub-corpora for their respective populations through in-field psycholinguistic testing.

## 2. Corpus creation

As a starting point, the regions of Argentina, Mexico, and Spain were identified as the largest potential sources of TV/film data as they are large media-producing countries. Conveniently, these regions also represent a well-dispersed geographic sample of the Spanish-speaking world, providing a high probability of reflecting variation in linguistic usage patterns.

### 2.1. Acquisition of TV/film transcripts from the web

A number of TV/film subtitle repositories were reviewed as potential sources of language data. Of those repositories, Opensubtitles.org appeared most promising, given that it is one of the largest community repositories on the web and due to the fact that primary key for TV/film subtitle in this repository include the same ID that is used by the Internet Movie Database (IMDb). Linking the data from the IMDb provided key meta-information for the development of this corpus including `country` of the production, the primary language of the film, and the `year` it was produced, as well as an array of information of potential use in the future. On request, Opensubtitles.org provided all Spanish subtitles uploaded by their repository community totaling over

---

Thanks to our collaborators Julieta Fumagalli (University of Buenos Aires), Mikel Santesteban (University of the Basque Country), Julio Serrano (National Autonomous University of Mexico)

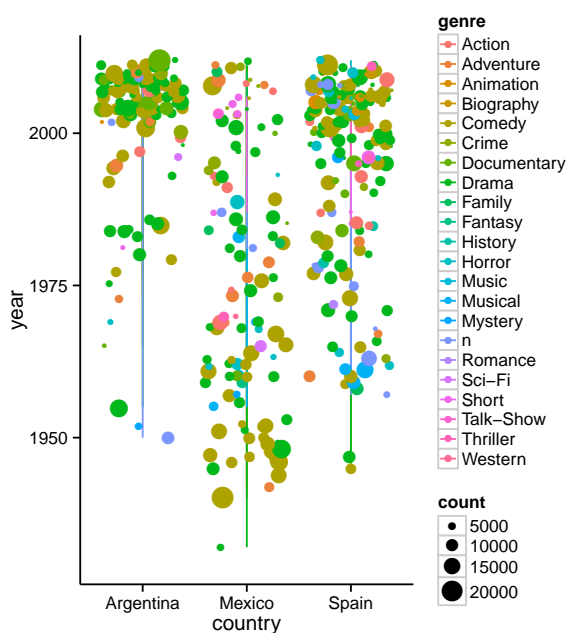


Figure 1: Overview of ACTIV-ES data: year, country, genres, and size of individual transcripts

120k subtitle files.<sup>1</sup> In order to identify only TV/film original-version transcripts from the three target regions, meta-information from the IMDb was accessed through the PyIMDb package for Python. Files were extracted in which the primary language was Spanish *es* and were produced by any of the target regions Argentina *ar*, Mexico *mx*, or Spain *es* and were labeled for *country*, *year*, *title*, *type*, *genre*, and *IMDbID*. Subsequent inspection revealed a number of false positives (including product/locale mismatches and user-upload errors: primarily language mis-assignments) which were identified and re-labeled and/or eliminated by semi-automatic and manual methods.

After applying all filters, the results included 430 total transcript files (Argentina 128, Mexico 119, and Spain 183) from wide range of genres and time periods, as seen in Figure 1. From this random sample of productions on the web, some time-skewing between the sub-corpora appeared, most notably for Mexican productions which most probably reflect the Mexican Golden Age of film (1930-1959). However, subsequent ecological evaluation of the corpus through psychological testing conducted in the field (discussed in Section 3.3.) suggests that Mexican native speaker experience is still predicted by the word frequencies in the Mexican sub-corpus better than either the Argentine or Spanish sub-corpora.

## 2.2. Text normalization

In order to prepare the transcript files for further text processing, two steps were taken: 1) removing subtitle time-codes and 2) orthographic normalization. The original transcript files were packaged in SubRip (*.srt*) format, as seen in Figure 2. To extract the language data from the SubRip format lines associated with time-codes, lines 1, 2, 5, 7, and 8,

```

1: 5
2: 00:03:06,898 --> 00:03:08,570
3: Marilia, esta ahi?
4: 6
5: 00:03:12,298 --> 00:03:13,367
6: Si.
7: 7
8: 00:03:13,367 --> 00:03:16,291
9: Mandame Ia salpillera
10: y el tapete de doble cara.

```

Figure 2: Raw transcript file example

were removed with a Perl script leaving the language data in lines 3, 6, 9, and 10, see Figure 4.

However, given the transcripts added to the Opensubtitles.org repository are user-submitted, the remaining text was not always found to be orthographically correct. First, some transcripts appeared to be automatically generated by OCR screen-reading software, which depending on the quality of the images, appeared to lead to errors such as ‘*la*’ in line 9 which should correctly appear as ‘*la*’. To address orthographic issues of this type, a Perl script was created with rules to find and replace errors that were not orthographically legal in Spanish.

```

# io/ios => lo/los
s/\s([Ii]o|[Ii]os)\s/ $1 /g;

# ia/ias => la/las
s/\s([Ii]a|[Ii]as)\s/ $1 /g;

```

Figure 3: Example Perl rules to correct orthographic errors

Second, some transcripts also lacked correct diacritic markings. An important component of the orthography of many of the world’s languages and key to subsequent natural language processing reliability, diacritic markings are often stripped or appear inconsistent due to technical errors (such as OCR errors) or human error (native speaker errors related to the level of formality and/or orthographic knowledge (Paredes, 1999)).

Diacritic marking errors in Spanish produce either ‘non-word’ errors such as ‘\**ahi/ahí*’ (line 3) or ‘(\**Mandame/Mándame*’) (line 9), which are not words, or ‘real-word’ errors such as ‘*esta/está*’ (line 3) or ‘*si/sí*’ (line 6) which are context-dependent errors (syntactically or semantically ambiguous).

‘Non-word’ errors are readily resolved with access to a large lexicon and/or morphological analysis resources –in this case a Spanish morphological analyzer (based on *foma* (Hulden, 2009)) was employed. However, ‘real-word’ errors are much more difficult to reliably resolve. Based on an evaluation of various methods in the literature for identifying and restoring ‘real-word’ errors (Francom and Hulden, 2013), a two-step process was adopted: 1) apply a series of expert-based hand-written rules to capture ‘safe’ cases, ‘\*¿*Como/¿Cómo*’, ‘\**para mi/para mí*’, etc., and 2) compile and apply collocational information for each ambiguous word in a orthographically correct corpus (Spanish

<sup>1</sup>Thanks to Opensubtitles.org for sharing these resources.

Gigaword (Mendonca et al., 2009)) using pre-specified collocation contexts.

In this approach, based on a decision list method outlined by Yarowsky (1994), lists were rigged so that the most reliable collocations were applied first. The types of collocations considered in the learning task were the following:

- Word to the left (-1w)
- Word to the right (+1w)
- The previous two words (-2w,-1w)
- The following two words (+1w,+2w)
- Any word in a  $\pm 20$  word window (+-20w)

Once the collocation counts were collected, the decision list was sorted by log-likelihood ratio so that more reliable rules were applied before less reliable ones. In the absence of any applicable rule, a ‘default’ rule chose the most frequent accent marking for ambiguous words.

```
3: Marilia, está ahí?
6: Sí.
9: Mándame la salpillera
10: y el tapete de doble cara.
```

Figure 4: Orthographically restored file example

### 2.3. Part-of-Speech annotation

The normalized text was then annotated for part-of-speech information with a standard trigram HMM tagger — HunPos (Halácsy et al., 2007)— trained on the AnCora corpus (Taulé and Recasens, 2008).<sup>2</sup> Visual inspection of the tagging results revealed that the statistical predictions based on newswire text appeared to systematically mislabel a number of key word forms either generally common in dialogue or common for a specific region— but in either case under-represented in newswire text.

Most problematic across all three regions were morphological forms related to the dialogue register: first person (‘yo’-singular and ‘nosotros’-plural), second person informal (‘tú’-singular), and imperative (‘formal and informal’). Regional variation, specifically the ‘voseo’ paradigm of Argentina and to lesser extent the ‘vosotros’ paradigm of Spain proved to be an issue for the AnCora-trained tagger. By and large these forms were labeled as nouns, suggesting training data error rather than tagger error.

Given the qualitative assessment of the tagging errors based on the AnCora corpus, a random sample of sentences was drawn from each sub-corpus (totaling 3,079 words), tagged with the base AnCora data, and then manually corrected and added to the training data in attempt to stem the gaps in the statistical tagger’s estimations.

### 2.4. Resulting data

The resulting ACTIV-ES corpus (version 0.1) has a total token count of 3,897,234 (Argentina: 1,232,656; Mexico:

<sup>2</sup>The widely-adopted EAGLES tagset for Spanish is used (see Freeling: <http://nlp.lsi.upc.edu/freeling/doc/tagsets/tagset-es.html>)

1,107,057; Spain: 1,557,521) and contains two main running text formats: 1) plain/normalized text and 2) PoS annotated text (an EAGLES tagset corpus and a simplified human-readable tagset corpus).<sup>3</sup> Corpus files are labeled for country, year, title, type, genre, and IMDbID.<sup>4</sup> In addition to the running text versions of the normalized and PoS annotated transcripts of the a corresponding set of 1:5 n-gram lists (words only) were created using the NLTK package for Python. These lists in .csv format, contain n-gram, relative frequency (per 100,000 tokens), and relative dispersion (per 10 transcripts) measures for each of the three sub-corpora and the total corpus.<sup>5</sup>

## 3. Corpus evaluation

To provide measures of reliability, the ACTIV-ES corpus was evaluated on three dimensions: 1) orthographic and annotation accuracy, 2) sub-corpus lexical distribution, and 3) lexical representativeness of the sub-corpora for their respective populations.

### 3.1. Text normalization & POS accuracy

First, to address text normalization steps in Section 2.2. and the part-of-speech annotation discussed in Section 2.3. a random sample of sentences (totalling 3,297 tokens) was drawn in roughly equal proportions (standard deviation 5 words) for each country by decade (1950-present). The first author, a linguist and professor of Spanish, manually checked the spelling, diacritic, and part-of-speech accuracy of the sample. Accuracy scores were calculated for word tokens only (2,546 words). The results from this assessment, reported in Table 1 suggest that the orthographic and diacritic normalization measures were highly effective. Yet, part-of-speech annotation accuracy was significantly lower than expected.

	Accuracy
Orthographic	99.8%
Diacritic	99.6%
Part-of-Speech	86.5%

Table 1: Accuracy results for text normalization and part-of-speech tagging

To gauge the effectiveness of efforts to augment the AnCora-based training set with manually-corrected sentences aimed at filling in potential register mismatch gaps, tagging results were compared with and without the manually-corrected sentences in the training sample. The augmented training set did improve tagging when compared to the tagger trained exclusively on the AnCora dataset, but only by a small margin (2.5%).

Reasons for the rather low overall tagging accuracy are not entirely clear. On the one hand, particular part-of-

<sup>3</sup>The simplified tagset is a reduced form of the Penn tagset and includes the following tags: /adv /adpos /cs /cc /det /pron /v /n /adj /interjection /number /date

<sup>4</sup>The IMDbID is key to future investigation aimed at more fine-grained meta-information.

<sup>5</sup>ACTIV-ES n-gram lists are available at: <https://github.com/francojc/activ-es/>

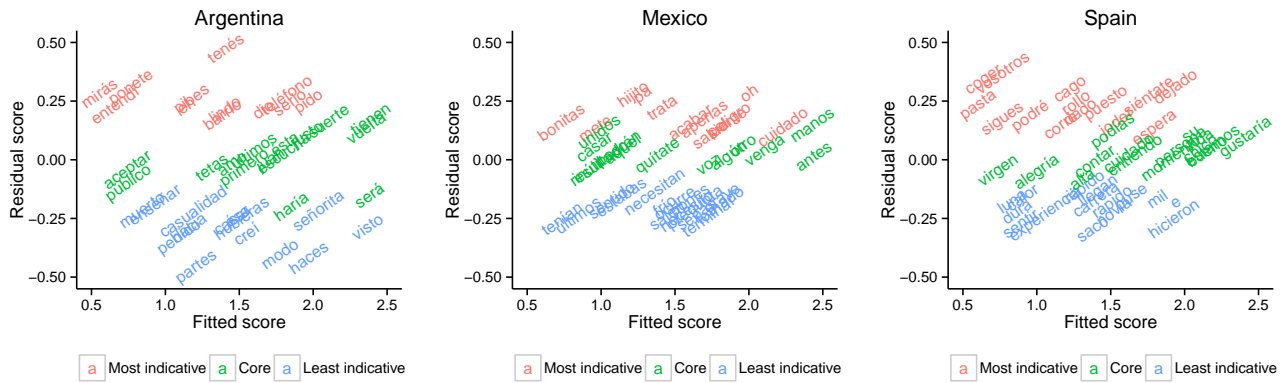


Figure 5: Sample linear regression plots indicating sub-corpus variation from the overall corpus distribution.

speech tagging conventions in the AnCora corpus may be a partial source of error. In particular, the AnCora team opted to associate some word sets with a single tag (ex. ‘al\_alcance\_de\_la\_mano’ as an ‘ADV’, ‘a\_lo\_largo\_de’ as an ‘ADPOS’, or ‘cerraban\_en\_banda’ as a ‘V’). Tagging of the ACTIV-ES corpus proceeded on a word-by-word basis, where word was delimited by whitespace. Conflating words into phrases raises the potential that key word-tag associations will not be seen in the training step; leading to inaccurate word-tag predictions. On the other hand another, residual issues with the register mismatch may also play a hand. As mentioned, AnCora and ACTIV-ES diverge in terms of register. Informal registers, such as found in the ACTIV-ES corpus, are characterized by morphological and lexical variation not typically found in formal registers. To overcome these issues, future work aims to develop a hand-tagged, gold-standard corpus from a subset of the ACTIV-ES corpus which adopts a word-by-word tagging approach.

### 3.2. Sub-corpus distribution

We turn now to an evaluation of the distinctiveness of the ACTIV-ES sub-corpora. The question asked here is to what extent are the sub-corpora, defined here by country, readily distinguishable by lexical distribution. To assess the potential similarity/dissimilarity between the sub-corpora two approaches were taken. In the first, unigram frequency (word occurrence) and dispersion (number of files a word appears in) scores were submitted to a linear regression which evaluated the deviation of frequency and dispersion scores for each sub-corpus from the overall ACTIV-ES corpus scores. As expected, lexical distributions differ between the sub-corpora, often to a great extent—see Figure 5. In a second approach, a text classifier was built to assess the predictiveness of the lexical distributions per country. The ‘e1071’ package for R (R Core Team, 2013) was used to run a Bernoulli implementation of a Naïve Bayes classifier.<sup>6</sup> Words were used as features and the priors were left at the default setting. A 75/25 split was used for training/classifying. Words occurring in less than 5 documents in the training stage and unknown in the classifying stage

<sup>6</sup>If a word appeared in a document it was given the value “yes”, otherwise, “No”

were dropped to reduce the size of the model and simplify the analysis, respectively. Predicted and actual values are reported in Table 2.

Predicted	Actual			Row Total
	Argentina	Mexico	Spain	
Argentina	18	0	0	18
Mexico	0.667	0.000	0.000	25
Spain	1	19	5	25
	0.037	0.679	0.094	65
Column Total	8	9	48	65
	0.296	0.321	0.906	108
	27	28	53	108
	0.250	0.259	0.491	

Table 2: Confusion matrix for Naïve Bayes classifier

These results suggest that the sub-corpora capture distinct lexical distributions, with some overlap most notably for the Argentine and Mexican corpus.

### 3.3. Sub-corpus representativeness

Given the aim of this corpus is to create a comparable corpus of ‘everyday’ Spanish, an important measure to explore is the extent to which the inter-corpus variation is in fact reflective of the experience of the average speaker of the respective population. The current evaluation explores the apparent sub-corpus variation and assess the degree to which the inter-corpus variation observed in the text classification task is reflected in the mental entrenchment of varied linguistic distributions in the native population. To this end, a series of preliminary lexical processing experiments assessing ‘Frequency Effects’ (Scarborough and Scarborough, 1977) were conducted in Argentina, Mexico, and Spain. Recently employed by psycholinguists to vet corpora for reliable frequency norms, co-authors of the current paper have demonstrated that the robustness of the frequency effect can be leveraged as a metric to assess the current ecological validity of a sample and objectively guide the development of corpora aimed at representing contemporary language usage and exposure to language in native populations (Francom et al., 2010; Francom and Ussishkin, accepted).

Results from visual word recognition task conducted in Ar-

gentina (N=101), Spain (N=81), and Mexico (N=84) for 240 co-occurring words (matched for other well-known language processing correlates) show that the frequency and dispersion of words for the sub-corpora correlate significantly with the lexical behavior of their respective populations, and not with the distributions from the other sub-corpora. These findings suggest that, at the word level, lexical variation found in the ACTIV-ES corpus approximates particular usage patterns of the respective native populations and provide support for the representativeness of these sub-corpora.<sup>7</sup>

Follow up tests comparing frequency and dispersion measures drawn from the relatively small ACTIV-ES corpus with corresponding measures from a large, general-purpose Spanish corpus ‘Corpus del español’ (100 million words) (Davies, 2002) to the lexical processing measures of natives demonstrates that the dialogue-based language from ACTIV-ES are more in-line with the ‘everyday’ linguistic environment of Argentine, Mexican, and Spanish populations.

#### 4. Conclusion

In this paper we have outlined the methods and evaluated the results of creating a novel Spanish language resource based on TV/film transcripts for three regions of the Spanish-speaking world. This resource aims to address a gap in current resources which are biased towards formal, often written, language and are not build to capture intra-dialect variation by presenting steps to acquire, curate, and evaluate informal, dialogue-based language data drawn from the web.

Presenting a promising methodology for acquiring informal, dialogue-based language for target dialects from the web, we have also highlighted the challenges of normalizing and annotating text which is dialect and/or register incongruent with available resources. Furthermore, initial evaluation using psycholinguistic methods provide evidence for the representativeness of this cross-dialect resource.

#### 5. Acknowledgements

The first author has received partial research funding for this project from the National Endowment for the Humanities (NEH) grant number HD-51432-11.

#### 6. References

Douglas Biber. 1998. Corpus linguistics: Investigating language structure and use.

Marc Brysbaert and Boris New. 2009. Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior research methods*, 41(4):977.

Marc Brysbaert, Emmanuel Keuleers, and Boris New. 2011. Assessing the usefulness of Google Books’ word frequencies for psycholinguistic research on word processing. *Language Sciences*, 2(March):1–8.

Curt Burgess and Kay Livesay. 1998. The effect of corpus size in predicting reaction time in a basic word recognition task: Moving on from Kucera and Francis. *Behavior Research Methods Instruments and Computers*, 30:272–277.

Mark Davies. 2002. Un corpus anotado de 100.000.000 palabras del español histórico y moderno. *SEPLN 2002 (Sociedad Española para el Procesamiento del Lenguaje Natural)*, pages 21–27.

Jerid Francom and Mans Hulden. 2013. Diacritic error detection and restoration via part-of-speech tags. In *Proceedings of the 6th Language and Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics*, Poznan, Poland.

Jerid Francom and Adam Ussishkin. accepted. Assessing corpus representativeness through psycholinguistic measures. *Corpus Linguistics and Linguistic Theory*.

Jerid Francom, Amy LaCross, and Adam Ussishkin. 2010. How specialized are specialized corpora? Behavioral evaluation of corpus representativeness for Maltese. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC ‘10)*, Valletta, Malta, May. European Language Resources Association (ELRA).

Péter. Halácsy, András. Kornai, and Csaba Oravecz. 2007. HunPos: an open source trigram tagger. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, pages 209–212. Association for Computational Linguistics.

Mans Hulden. 2009. Foma: a Finite-State Compiler and Library. In *EACL 2009 Proceedings*, pages 29–32.

Christopher Manning. 1999. *Foundations of statistical natural language processing*.

Angelo Mendonca, David Andrew Graff, and Denise DiPersio. 2009. *Spanish gigaword second edition*. Linguistic Data Consortium.

Yuan Kui Shen Aviva Presser Aiden Adrian Veres Matthew K. Gray Joseph P. Pickett Dale Hoiberg et al. Michel, Jean-Baptiste. 2010. Quantitative Analysis of Culture Using Millions of Digitized Books. *Science*.

Florentino Paredes. 1999. La ortografía en las encuestas de disponibilidad léxica. *Reale*, 11:75–97.

R Core Team, 2013. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Charles Cortese Scarborough, Don L. and Hollis S. Scarborough. 1977. Frequency and repetition effects in lexical memory. *Journal of Experimental Psychology: Human Perception and Performance*, 3(1):1–17.

Maria Antònia Martí Taulé, Mariona and Marta Recasens. 2008. AnCora: Multilevel annotated corpora for Catalan and Spanish. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC-2008)*.

David Yarowsky. 1994. Decision lists for lexical ambiguity resolution: Application to accent restoration in Spanish and French. *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, pages 88–95.

<sup>7</sup>Plans are in the works to expand the evaluation to include other linguistic units and other tasks and modalities.