ALGORITHMIC EXPLORATION OF AMERICAN ENGLISH DIALECTS

Alëna Aksënova *

alena.aksenova@stonybrook.edu
Stony Brook University
Stony Brook, NY

Antoine Bruguier[a], *Amanda Ritchart-Scott*[a], *Uri Mendlovic*[b]

{tonybruguier,aritchart,urimend}@google.com
[a]Google LLC, Mountain View, CA
[b]Google LLC, Tel Aviv, Israel

ABSTRACT

In this paper, we use a novel algorithmic approach to explore dialectal variation in American English speech. Without the need for human phonemic annotations, we are able to use an existing corpus transcribed in text form only. Our results show that, in general, American English dialects can be divided into two larger groups: dialects of the South (Texas to North Carolina except for peninsular Florida), and the rest of the country. Our results confirm some well-known results from dialectology, such as the *pin-pen* merger, but show that some other ones, such as the *cot-caught* merger, may be losing their isogloss boundaries. Moreover, we demonstrate that our algorithm can extend to dialectal features in other languages.

Index Terms— dialectology, isogloss, phonology, pronunciation learning

1. INTRODUCTION

Speech is varied and diverse; and it largely depends on the geographical and social dialectal features of the speaker. Speech recognition systems are mostly trained using mainstream or standard language data, and therefore they frequently show decreased accuracy when recognizing utterances that diverge from the standard form [1]. A growing body of research has begun to investigate this problem, such as vowel variation in French-Algerian Arabic [2], lenition of voiced stops and coda /s/ in Spanish [3], and regional variation of /r/ in Swiss German [4].

In this paper, we continue this line of research by exploring dialectal varieties of American English speech. American English and its sociolinguistic characteristics have been systematically studied since the 1960s (e.g., [5, 6, 7]). However, most of this research has been done on small amounts of speaker data because it requires expensive phonemic transcription. In this study, we use an audio corpus transcribed only in orthographic form. We investigate well-known dialectal features of American English phonology, including the *pin-pen* merger, *cot-caught* merger, R-dropping, monophthongization of /ai/, and consonant cluster simplification [8]. Additionally, we investigate two well-known lexical isoglosses (geographical boundaries of linguistic features): *you guys-y'all* and *coke-pop-soda*.

In general, our results suggest that there are two major phonological dialects in the US that divide the country in two parts: the South (excluding Florida) and the rest of the country. Moreover, this geographical boundary seems to correlate with the distribution of the lexical items *you guys* and *y'all*, where dialects with Southern phonological features prefer the form *y'all*. Therefore, we suggest that the *you guys-y'all* lexical isogloss outlines a larger dialectal distinction between mainstream and Southern US English dialects.

In the rest of the paper, we explain the experimental setup and how we targeted certain dialectal features, discuss the results of the experiments, illustrate the obtained phonological isoglosses, and briefly discuss extending our approach to other languages.

2. EXPERIMENTAL SETUP

We ran experiments using the same corpus of United States voice traffic as Li et al. [1]. We had corresponding human-annotated orthographic transcriptions and city-level geolocation markers. We did *not* have any phonemic transcription of the audio.

2.1. Algorithm for the computation of lexical isoglosses

To investigate lexical isoglosses, we used regular expressions [9] matching patterns in text to search for the intended lexical items in the audio transcriptions. We then produced a heat map representing the frequency of one word versus another: orange represented areas containing a higher occurrence of the feature, while blue represented areas where the feature was less frequent. These shades of blue and orange are appropriate for being interpreted by colorblind readers. We applied a threshold to the map to reduce noise such that a minimum number of data points was required in order to project a point on the map.

2.2. Algorithm for the computation of phonological isoglosses

The phonological features investigation followed a complex pipeline. In order to transcribe the spoken data, we modified a previous pronunciation learning algorithm [10]. This allowed us to obtain phonologically annotated data without the need of human transcribers.

The first step of the algorithm is to build a force-align finite state transducer (FST). For every utterance, we construct an FST that represents a phonemic sequence corresponding to the orthographic transcription of the audio data. For example, if the text transcript of the data is *pen*, we construct an FST that accepts the phonemic sequence /pɛn/ (Figure 1) because this is the way *pen* is pronounced in standard American English.

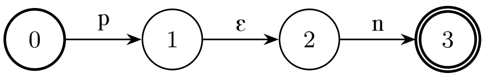

Fig. 1. Utterance "pen" before a phonological substitution.

*Work performed while interning at Google.

978-1-5090-6631-5/20/$31.00 ©2020 IEEE

To illustrate the algorithm further, let us consider the *pin-pen* merger, a dialectal feature where vowels /ɛ/ and/ɪ/ become indistinguishable before nasal consonants. First, we modify the FST to allow for a phoneme change. For every arc where the /ɛ/ phoneme is accepted as an input, we also add a second arc that accepts the /ɪ/ phoneme as input. We do not assign a weight to these arcs so as not to favor one phoneme over the other. For our *pen* example, the new FST is shown in Figure 2.

Once the additional arcs are added, we force-align the audio files using an unmodified acoustic model (AM) [10]. In a traditional force-align mode, the AM assigns probabilities for each phoneme at each transition, but these probabilities are ignored because only one arc is available to transition from one state to another.

However with our modified decoding graph, the AM information is used to determine which arc to take given the assigned probabilities to every transition of the given transducer. These probabilities help determine if the sound under consideration is acoustically more similar to one phoneme or another, such as /ɛ/ or /ɪ/ in the *pen* example. Crucially, this step shows if the pronunciation of the word is prototypical, or if, on the contrary, we are observing a phonological variant. Continuing the *pen* example, if the AM assigns a higher probability to the /ɛ/ phoneme, we are observing a realization of the word *pen* from a dialect without the *pin-pen* merger. However, if the AM detects /ɪ/ as the most likely choice, this indicates the presence of the merger. We did not retrain the AM, but rather re-used the one currently operating in the production model.

Once the forced alignment is complete, we keep track of the most likely sequence of phonemes. If the sequence of phonemes is the same as the canonical pronunciation, we assign the utterance to one color on the map, and if the sequence is different from the canonical pronunciation, we assign the other color. In areas where a dialectal merger does not occur, the forced alignment using the modified decoder should output the canonical pronunciation. In areas where a dialectal merger does occur, we hypothesize that the AM randomly picks the canonical or modified pronunciation.

We also added the option to examine the context of the phoneme substitution. For all the utterances where variation occurs before or after a certain phonological context, we first identify if the utterance has the phoneme being affected by the change. If so, we then need to identify whether the phoneme is in the correct environment. In the *pen* example, we find the desired environment: the phoneme /ɛ/ is indeed followed by a nasal /n/. Therefore, we add one more arc to the current transducer denoting a potential substitution (Figure 2).

Each outcome (merger vs. no merger) can be represented by a dot on a map corresponding to the nearest city of the individual audio file. Different colored dots represent the presence and absence of a certain feature. The final result is a heat map illustrating the distribution of the dialectal feature. For coordinates where there was too few examples, we did not assign any color.

2.3. Unknown accuracy of the phonological isogloss algorithm

While we felt that the proposed algorithm was reasonable, we did not assume correctness *a-priori*. The recognition system is highly complex and uses neural networks (such as the AM) whose behavior is opaque [11]. Regardless of the algorithm, a heat map can always be generated but the reason for one region to exhibit dialectal variation is not clear. As an example, rural areas might be more likely to have audio recorded from speakers' cars, thus making the audio more noisy and in turn increasing the number of substitutions.

We thus ran two sets of experiments. In the first set of experiments, we took independently known isoglosses and confirmed that our algorithm could replicate the results. Once we had confidence that the new algorithm was performing adequately, we ran a second set of experiments that investigated less well-studied dialectal variation of American English.

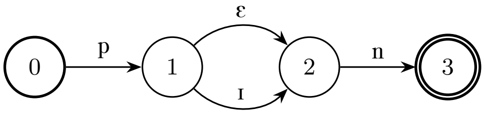

Fig. 2. Utterance *pen* after the substitution.

3. TARGETED DIALECTAL FEATURES

In order to conduct a dialectal study on American English, we prepared a collection of isogloss features based on the following principles. First, the features needed to be relatively well studied so that we could compare our results and therefore verify that our experimental setup was working as intended. Moreover, studying well-known dialectal features also allows us to trace how the features or geographical boundaries may have changed over time (if at all). Second, the dialectal features needed to be regional in nature since we only had associated approximate geolocation information available (and not, for example, gender or socioeconomic information). Linguistically, the features we targeted can be divided into two categories: phonological and lexical.

We mostly focused on phonological features, detecting how different phonemes are realized in different parts of the United States. The phonological features we targeted included the *pin-pen* and *cot-caught* mergers, R-dropping, G-dropping, consonant cluster simplification, merger of /ɪ/ and /i/ before /l/, and monophthongization of /aɪ/, all of which are defined and explained in sections 4 and 5.

We also targeted several lexical features. However, we found that detecting the intended use of content words was a problematic task. For example, when plotting the distribution of the word *pop* for the *coke-soda-pop* experiment, it was difficult to only target uses of *pop* that referred to carbonated beverages and not to the contracted form of the word *popular*; or to make sure that *coke* was used as a generic term and not referring to a particular brand of drink. In contrast to content words, generating isogloss maps of functional words, such as *you guys* and *y'all*, was comparatively easy. Therefore, we only present limited results on lexical isoglosses.

4. SETUP VERIFICATION

To confirm that the algorithm produced reliable results, we chose several well-studied dialectal features (as described in section 3) and compared the results produced by the algorithm against reference maps from previous research. We considered close similarity between the maps as evidence for the validity of our results.

4.1. Confirming the lexical algorithm: *you guys* vs. *y'all*

The form *y'all* is commonly associated with the South, whereas *you guys* is used predominantly in other parts of the country [12]. The heat map on Figure 3 verifies this well-known isogloss. As can be seen, there is a wide area from Texas to North Carolina (except southern Florida) where preference is given to the *y'all* form, while the rest of the country prefers the *you guys* form. In the following results we show that the majority of phonological features we investigate are distributed in a similar way to the *you guys*-*y'all* isogloss.

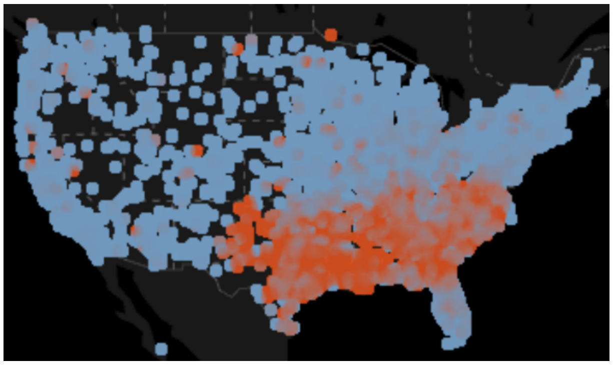

Fig. 3. Distribution of *y'all* (orange) and *you guys* (blue). A dot is plotted when at least 5 matches were detected for that location.

4.2. Confirming the phonological algorithm: *pin-pen* merger

The *pin-pen* merger is a process affecting realization of vowels /ɛ/ and /ɪ/ before /m/, /n/, and /ŋ/. This process has been documented in the South, although it is also claimed to be present in some parts of California [8]. This merger results in the two vowels becoming indistinct from one another, therefore making word pairs such as *pin-pen* or *windy-Wendy* sound the same [8, 13]. Our results mostly confirm this distribution, as illustrated in Figure 4. Interestingly, we see examples of the *pin-pen* merger occurring in some parts of Minnesota as well. The maximum observed level[1] of this merger was 13% in the South, and on average it affects 4.1% of /ɪ/ and /ɛ/ vowels immediately preceding a nasal in all of the American English data.

5. AMERICAN ENGLISH ISOGLOSSES

5.1. Consonant cluster simplification

Consonant cluster (CC) simplification, or deletion of a consonant in the presence of other consonants, is a prominent feature of African American Vernacular English [14]. In dialects exhibiting this feature, *computer* can be pronounced as *coputer*, or *friend* as *frien*. Our results suggest that this feature is widespread in the South, where up to 14% of total consonant clusters are simplified in Mississippi, Louisiana, and Arkansas, as illustrated in Figure 5.

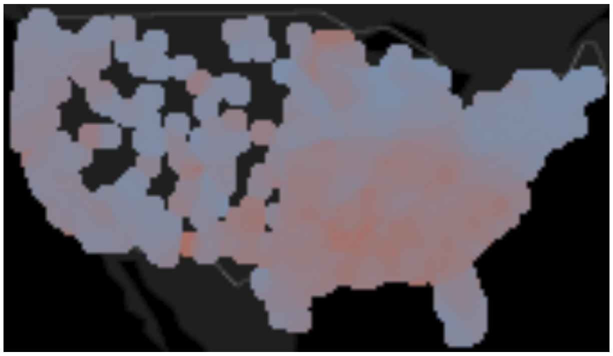

Fig. 4. Presence (orange) and absence (blue) of the *pin-pen* merger (colors are scaled). The maximum observed level of the merger is 13% and the minimum is 0.8%.

5.2. R-dropping

R-dropping occurs when the final /r/ sound is dropped at the end of words, therefore making words such as *car* sound like *cah*. The general assumption is that this feature is mostly widespread in the South and on the East Coast [8]. Nevertheless, our results suggest that even though there is a slight increase in R-dropping rates on the East Coast, this is primarily a Southern feature where up to 5% of word-final /r/ sounds are being dropped, as illustrated in Figure 6.

5.3. Other isoglosses

One of the most unexpected results we obtained was the absence of a *cot-caught* merger (also known as *low-back merger*) isogloss. In the dialects exhibiting this merger, phonemes /ɔ/ and /ɒ/ are merged, which makes word pairs such as *cot-caught* or *Don-dawn* impossible to tell apart. We expected to find an absence of this merger in the South, New England, and around Michigan, and a presence of the merger in the rest of the US [8]. However, our results did now show any area where the presence or absence of this merger would be particularly high or low; on average, the merger appeared in 2.6% of cases where phonemes /ɔ/ or /ɒ/ were present.

We also investigated several other phonological features: G-dropping, /ɪ/ and /i/ merger before /l/, and /aɪ/ monophthongization. G-dropping occurs when /g/ is dropped at the end of words (e.g., *running* vs. *runnin'*), and it is expected to occur more frequently in the South [15]. Our results also support that G-dropping primarily occurs in the South, where it occurs in up to 10% of relevant utterances in Louisiana. We also found that the merger of /ɪ/ and /i/ before /l/ is especially frequent in the area between Texas and North Carolina with a maximum frequency of 8.2% in Louisiana, which aligns with previous results in the literature [8]. This same geographical area shows evidence of /aɪ/ monophthongization, with the most frequent occurrence of the phenomenon in Tennessee. In 6.3% of relevant utterances in Tennessee, the monophthong /aɪ/ is realized as a long /ɒ/ sound. This result aligns with past research, which suggests a higher occurrence of this feature in states with higher numbers of African American Vernacular English speakers [14, 16].

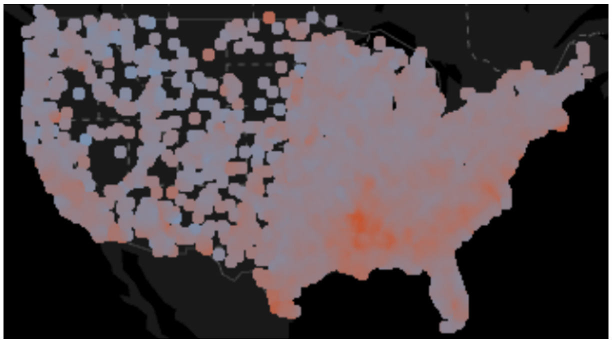

Fig. 5. Frequent (orange) and rare (blue) CC simplifications (colors are scaled). The maximum observed level of the simplification is 13.9% and the minimum is 2%.

[1] Values of the *maximum observed level* with respect to a certain location are obtained by counting all the utterances where the substitution was performed, and dividing it by the total number of utterances where the phonological environment would allow for that substitution. In both cases, we only include utterances produced near the location under consideration.

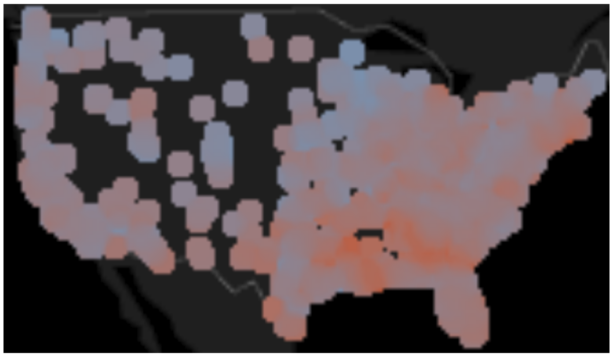

Fig. 6. Frequent (orange) and rare (blue) R-dropping (colors are scaled). The maximum observed level of R-dropping is 5% and the minimum is 0.3%.

In addition to the phonological features listed above, we also looked at the distribution of lexemes *tennis shoes* and *sneakers* across the US. We found a preference for the form *sneakers* in the Northeast and free variation of both forms in the rest of the country. While the strong preference for *sneakers* in the Northeast was expected, we expected to find a stronger preference for *tennis shoes* in the rest of the US [17]. Lastly, we investigated the distribution of lexemes *pop-soda-coke*, but the results did not show any distinctive pattern despite what is usually outlined in the literature [18].

6. EXTENDING TO OTHER LANGUAGES

It should be emphasized that our pronunciation learning algorithm is language-independent and can be extended to other languages. Therefore, we have begun to extend our results to Latin American Spanish, which exhibits dialectal variation that is less well studied (though see [19, 20]).

In an initial experiment, we investigated the existence of *sheísmo* in Argentina compared to the rest of Latin America. It is expected that in and around Buenos Aires the phoneme /ʃ/ is used instead of /j/ by younger speakers [21]. For example, the word *yerba* (English *herb*) would be pronounced as /ʃerba/. Our results confirm this hypothesis; we found up to 4.2% of total /ʃ/ phoneme substitutions in the Buenos Aires region, as illustrated in Figure 7.

7. DISCUSSION

This study suggests that pronunciation learning is a useful and reliable way to conduct dialectal research when spoken and orthographically transcribed data are available. The advantage of this algorithm is that it can automatically transcribe spoken data and detect the presence or absence of non-mainstream linguistic features, therefore allowing cheaper research to be done without phonemic annotations.

We were able to replicate several well-known dialectal maps, such as *you guys-y'all* and consonant cluster simplification, which suggests that our experimental pipeline is correct and can be used to study less well-known or agreed upon regional variation. Moreover, this experimental method is much cheaper and scalable as it can re-use existing sets already transcribed in text form. Additionally, pronunciation learning is language independent and can be easily extended to languages other than English, such as Latin American Spanish.

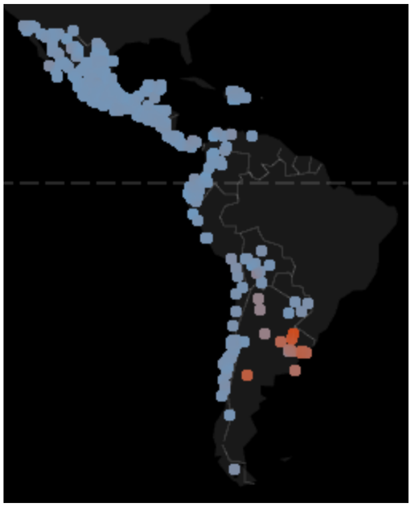

Fig. 7. Frequent (orange) and rare (blue) *sheísmo* (colors are scaled). The maximum observed level of *sheísmo* is 4.2% and the minimum is 0%.

However, there are limitations to this method. Results are based on outputs from the acoustic model; therefore we need to ensure that it correctly recognizes the input audio. For example, to further examine our surprising result of the lack of the *cot-caught* merger isogloss, we could pass a data set exhibiting the absence and presence of the *cot-caught* merger through the acoustic model and see how it processes the variations of /ɔ/ and /ɒ/. If the model correctly determines the presence or absence of the merger, we could confirm that our results regarding the merger are indeed reliable, which would suggest that the merger is now ubiquitous such that there is no clear isogloss boundary any longer. Otherwise, poor performance on the data set would be an indicator that the acoustic model needs further improvements.

Lastly, speech recognition systems can benefit from experiments like these where the frequency and distribution of non-standard forms of a language are analyzed. For example, results from these experiments can be used to create more balanced data sets that include many different dialectal varieties.

8. CONCLUSION

In this paper, we presented a novel algorithm to explore phonological variations in speech. This algorithm does not require data to be transcribed phonemically. Running on an existing data set, we explored lexical and phonological variation in American English speech. In general, the dialectal features we examined revealed one major isogloss across the country: the area between Texas and North Carolina (the Southern dialect) and the rest of the country. The Southern dialect exhibits the presence of the *pin-pen* and /ɪ/-/i/mergers, and also involves higher levels of R-dropping, among other features. Interestingly, this area approximately corresponds to the isogloss outlined by the choice between the lexical forms *you guys* and *y'all*, where the form *y'all* is preferred in the geographical area with the highest concentration of Southern phonological features. Lastly, we demonstrated how this algorithm can extend to other languages such as Latin American Spanish, where we presented an initial result on the presence of *sheísmo* in Argentina.

9. REFERENCES

[1] B. Li, T.N. Sainath, K.C. Sim, M. Bacchiani, E. Weinstein, P. Nguyen, Z. Chen, Y. Wu, and K. Rao, “Multi-dialect speech recognition with a single sequence-to-sequence model,” *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4749–4753, April 2018.

[2] J. Wottawa, A. Djegdjiga, M. Adda-Decker, and L. Lamel, “Studying vowel variation in French-Algerian Arabic code-switched speech,” in *Proc. Interspeech*, 2018, pp. 2753–2757.

[3] I. Vasilescu, N. Hernandez, B. Vieru, and L. Lamel, “Exploring temporal reduction in dialectal Spanish: A large-scale study of lenition of voiced stops and coda-*s*,” in *Proc. Interspeech*, 2018, pp. 2728–2732.

[4] A. Leemann, S. Schmid, D. Studer-Joho, and Marie-José Kolly, “Regional variation of /r/ in Swiss German dialects,” in *Proc. Interspeech*, 2018, pp. 2738–2742.

[5] B. Bernstein, “Language and social class: A research note,” *British Journal of Sociology*, vol. 3, no. 11, pp. 271–276, 1960.

[6] J. Gumperz, “Linguistic and social interaction in two communities,” *American Anthropologist*, vol. 6, no. 66, pp. 137–153, 1964.

[7] W.A. Stewart, “A sociolinguistic typology for describing national multilingualism,” in *Readings in the Sociology of Language*, J.A. Fishman, Ed. The Hague, Paris: Mouton, 1968.

[8] W. Labov, S. Ash, and C. Boberg, *The Atlas of North American English: Phonetics, Phonology, and Sound Change: a Multimedia Reference Tool*, Mouton de Gruyter, 2006.

[9] Jeffrey E. F. Friedl, *Mastering Regular Expressions*, O’Reilly Media, 206.

[10] A. Bruguier, D. Gnanapragasam, L. Johnson, K. Rao, and F. Beaufays, “Pronunciation learning with RNN-transducers,” in *Proc. Interspeech*, 2017, pp. 2556–2560.

[11] Jenna Burrell, “How the machine ‘thinks’: Understanding opacity in machine learning algorithms,” *Big Data & Society*, vol. 3, no. 1, 2016.

[12] N. Maynor, “Battle of the pronouns: Y’all versus you-guys,” *American Speech*, vol. 75, no. 4, pp. 416–418, 2000.

[13] V.R. Brown, *The social and linguistic history of a merger: /i/ and /e/ before nasals in Southern American English*, Ph.D. thesis, Texas A&M University, 1990.

[14] E. Thomas, “Phonological and phonetic characteristics of African American Vernacular English,” *Language and Linguistics Compass*, vol. 5, no. 1, pp. 450–475, 2007.

[15] J. Yuan and M. Liberman, “Automatic detection of “g-dropping” in American English using forced alignment,” in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2011, pp. 490–493.

[16] V. Fridland, “‘Tie, tied and tight’: The expansion of /ai/ monophthongization in African-American and European-American speech in Memphis, Tennessee,” *Journal of Sociolinguistics*, vol. 3, no. 7, pp. 279–298, 2003.

[17] D. Coye, “The sneakers/tennis shoes boundary,” *American Speech*, vol. 61, no. 4, pp. 366–369, 1986.

[18] B. Vaux and S. Golder, “Harvard dialect survey,” online resource, 2003, Cambridge, MA: Harvard University Linguistics Department.

[19] J. Lipski, “Geographical and social varieties of Spanish: An overview,” in *Readings in the Sociology of Language*, E. O’Rourke J.I. Hualde, A. Olarrea, Ed. The Hague, Blackwell Publishing Ltd., 2012.

[20] G. Guy, “Variation and change in Latin American Spanish and Portuguese,” *Issues in Hispanic and Lusophone Linguistics*, vol. 1, pp. 443–464, 2014.

[21] J. Lipski, “Latin American Spanish,” *Language*, vol. 72, no. 4, pp. 821–825, 1994.