An information-theoretic method for mining regionalisms from Social Media Texts

Anonymous TACL submission

Abstract

The task of detecting regional colloquialisms —expressions or words used in certain regions in informal language— has classically relied on the use of questionnaires and surveys, and has also heavily depended on the expertise and intuition of the lexicographer.

The irruption of Social Media and its microblogging services produced an unprecedented wealth of content, mainly informal text generated by users. This gives an incredible opportunity for linguists to perform studies about the use and variation of languages.

In this work, we present three metrics based on Information Theory to detect regional colloquialisms on Twitter. The metrics proposed take into account both the number of occurrences of the word in a certain region and the number of users of the word. This tool helped lexicographers to discover a number of unregistered words of Argentinian Spanish, and also different meanings assigned to known expressions.

1 Introduction

Lexicography is the art of writing (designing, compiling, editing) dictionaries: that is, the description of the vocabulary used by members of a speech community (Atkins and Rundell, 2008). This work has been aided in the last 30 years by tools coming from Computational Linguistics, mainly in the form of corpora of selected texts. Statistical analysis of corpora results in evidence to support the addition or removal of a word from a dictionary, its marking as dated or unused, as regional, etc., depending on different criteria.

In the process of compiling dictionaries, differences emerge between dialects, where frequently certain words or meanings do not span across all speakers. Since languages are an ideal construct based on the observation of dialects, it is of paramount importance to establish which words are most likely to be shared by an entire linguistic community and which are only used by a smaller group. In this last case, the description profits greatly from information as precise as possible, about geographical extension (region, province, district, city, even neighborhood), about registry (colloquial, neutral, formal), about frequency (actual, past or a combination of both depending on chronological span of the corpus), or any other variable.

Agregar algo de Español acá, no? es donde los dialectos tienen más importancia

Words that are used exclusively or mainly in a particular subregion of the territory occupied by a linguistic community, or that are used there with a different meaning, are called *regionalisms*. For example. the word "che"[1], or "metegol"[2] are terms used more frequently in Argentina than in Spain. These words are commonly detected through surveys (Almeida and Vidal, 1995; Labov et al., 2005), or transcriptions, using methods depending more or less on the intuition and expertise of linguists. The result of this work is of great value for lexicographers, as they need evidence to support either the addition of a word into a regional dictionary or the indication of where the word is used. Information gathered with these traditional methods has been used to computationally calculate similarities in dialects (Kessler, 1995; Nerbonne et al., 1996).

buscar referencia para esto y chequear

The irruption of Social Media and its microblogging services produced an unprecedented wealth of content, mainly informal text generated by users. In particular, Twitter gives a great opportunity to linguists due to the possibility of access-

[1] interjection used to get the interlocutor's attention

[2] mechanic game that emulates footbal (futbolín) (de Letras, 2008).

ing to geotagged tweets, and gathering information about the procedence of users. This has been used to study dialects of languages(Gonçalves and Sánchez, 2014; Huang et al., 2016) and establish "continuous" isoglosses of them.

An valuable framework to navigate in this ocean of data is Information Theory. Tools from this field have been used to tell whether a hashtag is promoted by spammers (Cui et al., 2012; Ghosh et al., 2011) by analyzing its dispersion in time and users; also, to detect valuable features in Sentiment Analysis for statuses in this microblogging platform(Pak and Paroubek, 2010).

In the present work, we introduce a quantitative method based on Information Theory to find these regionalisms in Social Media Texts, particularly on Twitter. This method aided lexicographers in their task, avoiding most of the manual work described, and let them add a number of words into the 2018 version of *Diccionario del Habla de los Argentinos*.

expandir un poco

2 Method and Materials

Data

We looked for users with geographical information available. Information of departments in Argentina (second order division of administration of the country) was collected from the 2010 Census (citation needed). Then, a lookup was made for users with location matching those departaments, balancing the number of users per province. *tweepy* was used to interact with the Twitter API.

Although the retrieval of users was attempted on each department, they mainly concentrate around some cities. This phenomenon is due to limitations on the geographical information that Twitter makes available. As our unit of study is the province, this is not an issue.

From these users, their entire tweetlines were retrieved. Tweets were tokenized using *NLTK*(Bird et al., 2009), which has a special algorithm for this kind of text. Hashtags and mentions to users were removed, and remaining words were downcased, and repeated vowels normalized up to three repetitions ("woaaa" instead of "woaaaaaa").

As already done with respect to users, the number of words per province was balanced. Table 1 lists the figures for the collected dataset.

It is well known that Twitter vocabulary tends to be very noisy (Kaufmann and Kalita, 2010) with lots of words that come from contractions, non-normal spellings (such as vocalizations), typos, etc. Only words occurring more than 40 times, and being used by more than 25 users are taken into account. This removes about 1% of the total words and reduces vocabulary from 2.3 million words to around 135 thousand words.

While this normalization seems insufficient for most analysis, it is acceptable for our study as the phenomenon of locally-used words would emerge in spite of different spellings, typos, and other morphological variations. Indeed, the same word might appear in several different variations or spellings but with one normalized form which would be the more frequent.

Metrics

We can think of a *regionalism* as a word whose use is not uniform across all the studied region, whose concentration is high in a specific part of the country. We are trying, in fact, to measure disorder of the use of a word, and there exists an information-theoretic tool for this: entropy.

It is known that entropy holds information about the semantic role played by a word (Montemurro and Zanette, 2002, 2010). High-entropy words, given a text, are more likely to be pronouns, connectors and other non-informative words, whereas its low-entropy counterparts are usually nouns and adjectives playing a key role in the document.

Taking into account occurrence of words across a geographical region, words with high entropy (high disorder) can be regarded as words whose use is similar all across the region of study. On the other hand, low-entropy words are more used in a region than in the rest of it.

Let $p = (p_1, p_2, \ldots, p_N)$ the relative frequencies of occurrence of the word ω for each geographical region (in our case, provinces), the *word-count entropy* is defined as:

	Total	μ	σ
Words	647M	28.14M	6.64M
Tweets	80.9M	3.51M	0.91M
Users	56.2K	2.44K	0.04K
Vocabulary	7.5M	0.32M	0.04M

Table 1: Dataset summary information. Total figures are provided, along province-level mean and deviation

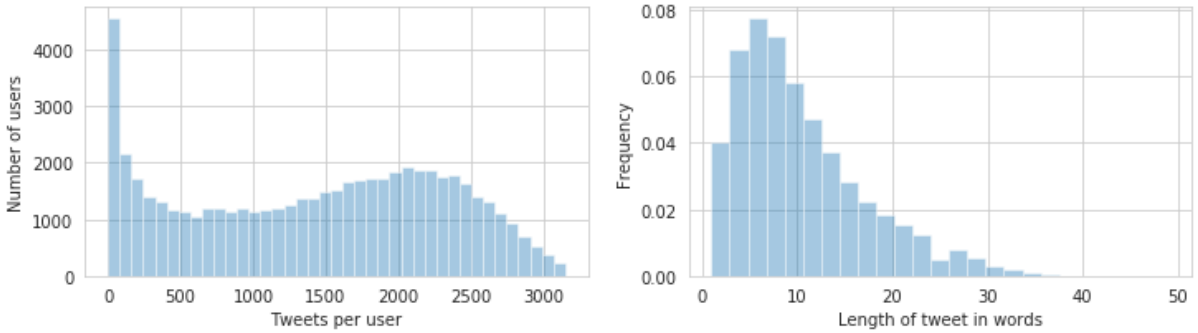

Figure 1: Distributional figures of the dataset. First figure shows the distribution of tweets per user. The second figure plots the distribution of the length (in words) of a tweet

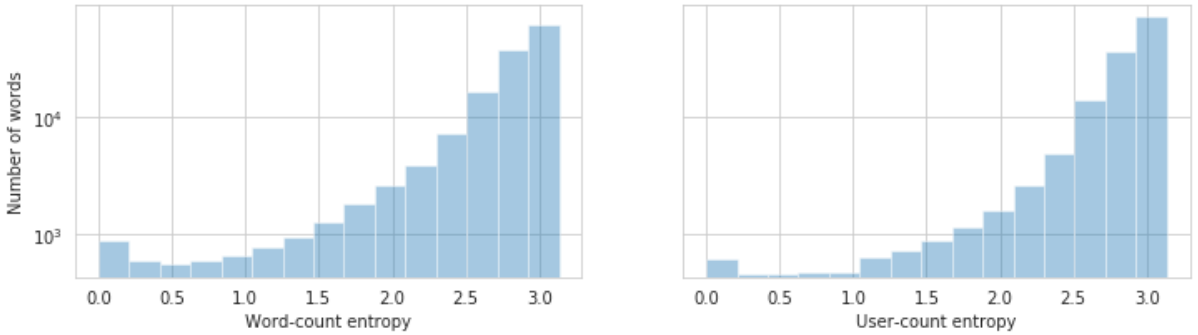

Figure 2: Histogram of word occurrence entropy (H_w). As one would expect, most words are used uniformly across the country

$$H_w(\omega) = -\sum_{k=1}^{N} p_k \log(p_k) \tag{1}$$

This measure does not take in account the frequency of words, which is also important. For instance, a word occurring once in a province has the same entropy as a word occurring millions of times at the same place. To remediate this issue, a measure based on (Montemurro and Zanette, 2010) was developed, taking into account the frequency of use. We define the *Information Value* of w as:

$$I_w(\omega) = p(\omega)\left(\log(n) - H_w(\omega)\right) \tag{2}$$

The $log(n)$ in the formula represents the maximum possible value of $H(\omega)$ (Shannon, 2001); for our work, recall that $n = 23$ (number of provinces in Argentina).

Another factor of information of a word is the amount of people using it on Twitter, something already used in previous works(Cui et al., 2012). Using the same arguments as above, if ω is a word, and if $q_1, q_2, \ldots, q_n$ are the relative probabilities of a user of ω belonging to each province, we define the *user-count entropy* as.

$$H_p(\omega) = -\sum_{k=1}^{N} q_k \log(q_k) \tag{3}$$

Then, we define *User-Count Information Value* of ω as:

$$I_p(\omega) = p(\omega)\left(\log(n) - H_p(\omega)\right) \tag{4}$$

This measures the concentration of the people using this word. In 2 we can observe the distribution of *Word-Count Entropy* and *User-Count Entropy* for the words in our vocabulary. As one would expect, most words are used uniformly across the country —both in occurrences and in users of it.

To take into account both *word-counts* and *user-counts* we define a third measure called *Mixed Information Value* as:

$$I(\omega) = I_w(\omega) I_p(\omega) \tag{5}$$

High values of $I(w)$ are reached when information value on users and on ocurrences is also high.

Due to the Zipfian nature of the distribution of words, the frequency of use of the most-used words is many orders of magnitude higher than others. This phenomenon also occurs with the user count. So the $p(w)$ term in equations (2) and (4) becomes a problem as very frequent words would overcome their low entropy. To alleviate this, we performed a normalization on the word frequency as follows: let

$$M_w = \arg\max_{\omega \in W} \#\omega \tag{6}$$

we define then

$$n_w(\omega) = \frac{\log(\#w)}{\log(M_w)} \tag{7}$$

Due to the logarithm in the formula, terms with a high amount of occurrences have little difference of $n_w(\omega)$ between them.

In an analogous form, we define a normalization for the amount of people using a word, named n_p. Hence, we redefine our information measures as:

$$I_w(\omega) = n_w(\omega)(\log(n) - H_w(\omega)) \tag{8}$$
$$I_p(\omega) = n_p(\omega)(\log(n) - H_p(\omega)) \tag{9}$$

Lexicographic Validation

Based on the list of words identified as having a contrastive use in a region, we performed a linguistic validation of the first thousand words given by each metric. It consisted of a detailed study, word by word, to determine if the word in question is part of the lexical repertoire of a community of speakers.

This excluded, as is traditional in lexicography, proper and local place names. These words occur mostly in their respective regions, having high entropy and also a high number of occurrences, hence resulting in high values for our metrics. To facilitate the detection of regionalisms, we automatically highlighted words suspected to be toponyms so that the team of lexicographers have a first warning about it.

A team of lexicographers analyzed the first thousand words for each metric. Along these lists, they were provided with tables having figures for each word and province: users, occurrences and normalized frequency (ocurrences per million words). Table 4 shows the results of this labeling. Using information of the number of users of a word seems to be capturing the highest number of words marked as regionalisms.

From these words marked as regionalisms, lexicographers performed a characterization of the results according to the linguistic phenomenon they represent. Table 3 displays lexicographic groups among the regionalisms found in the analyzed words with examples. Table 2 displays occurrences of three outstanding examples: words coming from guaranitic region.

.

Word	Guaranitic Region	Litoral Region
angá	32.80	0.33
angaú	8.42	0.03
mitaí	15.06	0.04

Table 2: Guaranitic words and its occurrences per million in two different dialectal regions

3 Results

Figure 3 displays the log-frequency and Entropy of words, along with the value of the respective metric. As we can see (both in the User-Count and Word-Count plots) the words we rank high in our lists are those closer to the upper-left of the plot — that is, highly-concentrated and occurring a consideriable number of times. Mixed Information Value seems to respect the User-Count order, and the last plot shows that although User-Count and Word-Count ranks are similar, they are not exactly the same as shown by the "darker" points close to the upper-left corner.

Regarding the lexicographic validation, Table 4 display the percentage of words marked as being lexicographically interesting. The number of users of a word alone seems to be the most informative feature.

4 Discussion

Although there are no other projects that provide a comparison term to assess the degree of success involved in this relationship, there is no doubt that, at least in the detection of local colloquialisms currently in use, the tool poses a real point of inflection for contrastive lexicography. This area of the lexicon is just the most elusive, since its impact on any printed medium arrives noticeably later and, even more important, in most cases it

Agregamos algo sobre que "exploramos más palabras" que las primeras 1000? o lo dejamos ahí?

Que agregue algo más Santiago sobre esto

Chequear estos números!!!!

Agregar un listado de las 10 primeras palabras de cada listado?

cambiar a "regionalism"

Group	Province	Example
Colloquialisms	Córdoba	"Perdon pero tenes que ser muy **culiado** para ir a mc y pedirte una ensalada"
	Mendoza	"Q **chombi** hacer un chiste y q la otra persona no se ría o no lo entienda"
	Formosa	"Tenía la re expectativa para este sábado y al final **trancó** todo"
Colloquialisms with a verbal or noun base	Buenos Aires	"Me vine a acostar y ya me dicen que parezco de 80 años ME CHUPA UN HUEVO LO QUE PIENSEN, DEJENME **ABUELEAR** "
	Neuquén	"Me calma mucho **mimosear** a mi perro "
	Tierra del Fuego	"Estaría bueno que ari venga aunque sea a saludarme y que no se quede todo el tiempo **pollereando**."
Indigenisms	Formosa	"Te regalo ser **mitaí** y ir a jurar la bandera con el guardapolvo caliente ese y la corbata que te ahorca todo"
	Corrientes	**Angá** mi negrito, esta triste
	Tucumán	Gracias tormenta **ura** por sonar como una pochoclera de chasquibums a las 3 de la mañana en mi ventana
Words alluding regional reality	San Juan	"Quiero a alguien que me diga vamos a comer **piadinas**, un pancho, un chori, una hamburguesa lo que sea y soy feliz"
	Misiones	"**Tareferos** que reclamaban asistencia interzafra en Posadas estarían preparando una protesta para hoy en la Fiesta del Inmigrante en Oberá."
	Jujuy	"Me encantan los bohemios anti sistema que usan vans. Es como que seas ecologista y uses un cuaderno hecho con media **yunga**."
"Leism"	Misiones	"No te olvides de **saludarle** a tu suegro hoy"
	Misiones	"Vine a **visitarle** a mis primas y estan re colgadas, para eso me quedaba en mi casa no maaa "
	Formosa	"A **esperarle** a nahuel, que traiga los teresss "
Fusions and acronyms	Buenos Aires	"Los sueños de la siesta me dejan **patra** "
Words with different meaning in a region	Mendoza	"Mañana que alguien **atine** con parque y porrones"
	San Juan	"**Mansas** ganas de sentarme a tomar un te con semitas"
	Tierra del Fuego	"**Habilítenme** una nueva espaldaa"
	San Juan	"sigo **asada** por cosas que han pasado hace como dos dias, que falla (Mendoza) / Que **asada** estoy, tengo la cabeza echa un lío"
Intejerctions	Formosa	"**Aijué**, encima me decís vieja, re que no pinta esto facundo jaja ya te dije como es la onda, fin "
	Formosa	"**Ains**, una mujer hablando de fútbol."
	Corrientes	"Al fin una buena: hora libreeee! **Yirr** "

Table 3: Characterisation of the regionalisms found in the analysis.

never reaches. Several words that are already included in the Dictionary of the Speech of Argentines (de Letras, 2008) were included as relevant, given that this fact is an additional confirmation of the relevance of the location that assigned the metric.

It is worth mentioning that words coming from *guaraní* —language spoken in Northern Argentina, Paraguay, Bolivia and Southwest of Brazil— coincide with the region delimited by

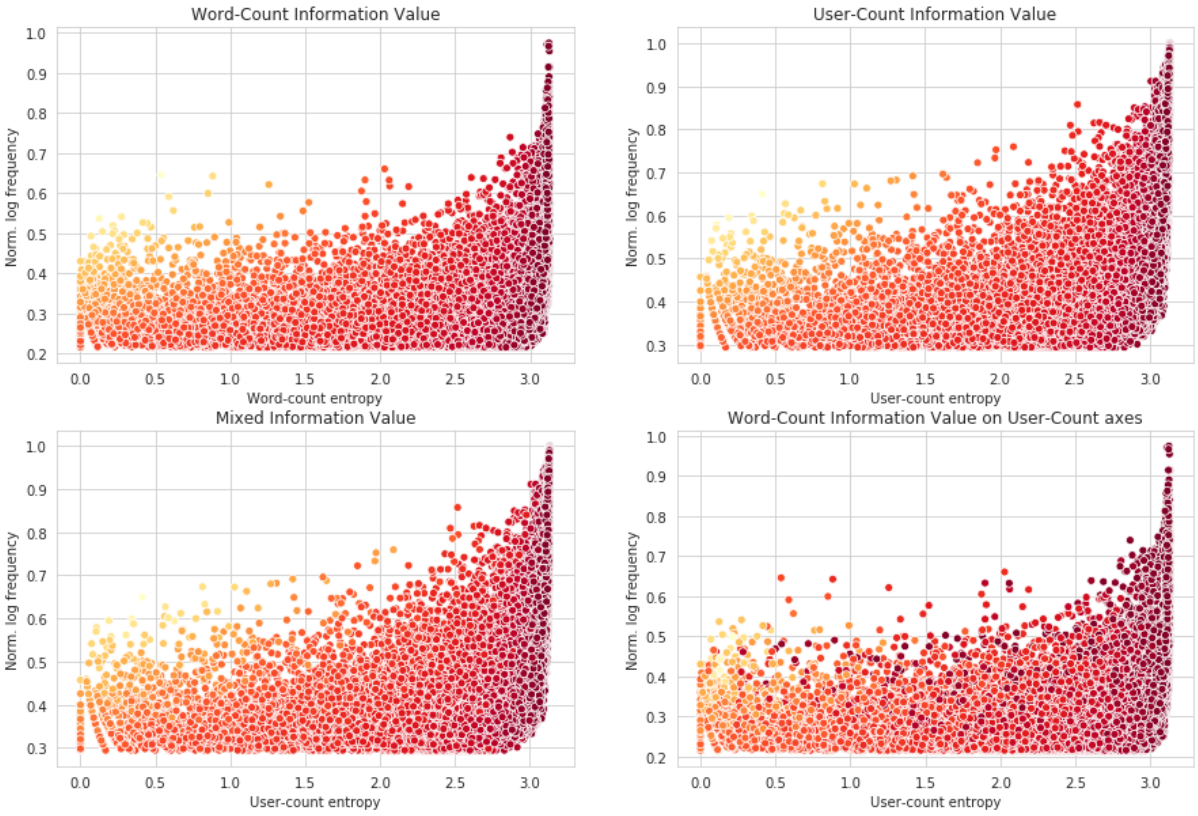

Figure 3: Scatter plot of words. Horizontal axis represents user-count entropy, vertical axis represents the normalized log user-frequency of the word, and colour indicates User-count Information Value: lighter means higher value

Metric	% of interesting words
Word-Count	21.9%
User-Count	30.2%
Mixed	25.3%

Table 4: Results of the metrics. The percentage of detected regionalisms in the first 1000 words is given for each metric: Word Information Value, User Information Value, and Mixed Information Value

(Vidal de Battini, 1964). An example of this phenomenon are the words *angá*, *angaú* y *mitaí*.

- Agregar que esto aplica no sólo a encontrar regionalismos sino temas/términos que sean propios de una región geográfica particular (marketing, ver penetración de una marca como en una de las citas).

5 Conclusions

In the present multidisciplinary work we developed metrics to detect regionalisms, using provinces as unit of measure, and collecting data from Twitter. These metrics used both the count of occurrences of a word as well as the number of users of it, and were based on entropy to measure their "concentration".

The metric we create uses entropy to measure the variation in the number of occurrences and the number of users that used it in the different provinces of the country. The 5000 words with the highest contrastivity value were selected to carry out a linguistic validation by the Argentine Academy of Letters. The validation yielded a result with around 300 words worthy of study of those 5000 words, that is, 1 word every 17. Although there are no other projects that provide a comparison term to assess the degree of success involved in this relationship, there is no doubt that, at least in the detection of local colloquialisms currently in use, the tool poses a real point of inflection for contrastive lexicography. Several of the words detected from the developed metric will be added to the Speech Dictionary of the Argentines.

Regarding statistical validation, we leave as future work the calculation of a statistical analysis applicable to our metric, since it is outside the

scope of this thesis. However, based on the analysis made through the Welch t test, we also have indications of the virtues of the developed metric.

In this paper, the regions formed with a province as a regional unit are analyzed, but this can be changed to replicate the analysis with different granularity. This way you could see the contrastive words in the different Spanish-speaking countries and compare the variations between larger regions or replicate the work within a single province or city.

One of the challenges that this work triggers is that of being able to identify regions / clusters with different dialectal uses. At the same time, it would allow to validate the validity of the regions proposed by Vidal de Battini in 1964 cite vidal1964espanol.

Also, the normalization process could be improved to have a greater precision in the words used. From a better normalization and the lemmatization with metalinguistic information of the corpus, we could go beyond the lexicon to study the syntactic phenomena of Spanish, as its variation in different regions. Continuing with the research line, we could analyze the lexical contrastivity by comparing the distribution of n-grams. On the other hand, it would be useful to add a system of recognition of names of entities to highlight certain proper names, in such a way that the list of words has more alerts about terms without linguistic interest.

It is important to note the advantages of textit Twitter since it allowed us to collect a large volume of text data, written by different people with information about their location. Regarding the disadvantages of this platform we can highlight the orthographic errors of the texts, the intentional modification of the words to generate emphasis or with a reason for faster writing. All this leads to an increase in the difficulty to normalize the text. We believe, despite all this, that the volume of data prevails when deciding a platform to collect them.

6 Contributors to this document

This document was adapted by Lillian Lee and Kristina Toutanova from the instructions and files for ACL 2018, by Shay Cohen, Kevin Gimpel, and Wei Lu. Those files were drawn from earlier *ACL proceedings, including those for ACL 2017 by Dan Gildea and Min-Yen Kan, NAACL 2017 by Margaret Mitchell, ACL 2012 by Maggie Li and Michael White, those from ACL 2010 by Jing-Shing Chang and Philipp Koehn, those for ACL 2008 by Johanna D. Moore, Simone Teufel, James Allan, and Sadaoki Furui, those for ACL 2005 by Hwee Tou Ng and Kemal Oflazer, those for ACL 2002 by Eugene Charniak and Dekang Lin, and earlier ACL and EACL formats, which were written by several people, including John Chen, Henry S. Thompson and Donald Walker. Additional elements were taken from the formatting instructions of the *International Joint Conference on Artificial Intelligence* and the *Conference on Computer Vision and Pattern Recognition*.

References

Manuel Almeida and Carmelo Vidal. 1995. Variación socioestilística del léxico: un estudio contrastivo. *Boletín de filología*, 35(1):Pág–50.

BT Sue Atkins and Michael Rundell. 2008. *The Oxford guide to practical lexicography*. Oxford University Press.

Berta Elena Vidal de Battini. 1964. El español en la argentina. Technical report, Argentina.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Anqi Cui, Min Zhang, Yiqun Liu, Shaoping Ma, and Kuo Zhang. 2012. Discover breaking events with popular hashtags in twitter. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM '12, pages 1794–1798, New York, NY, USA. ACM.

Rumi Ghosh, Tawan Surachawala, and Kristina Lerman. 2011. Entropy-based classification of'retweeting'activity on twitter. *arXiv preprint arXiv:1106.0346*.

Bruno Gonçalves and David Sánchez. 2014. Crowdsourcing dialect characterization through twitter. *PloS one*, 9(11):e112074.

Yuan Huang, Diansheng Guo, Alice Kasakoff, and Jack Grieve. 2016. Understanding us regional linguistic variation with twitter data analysis. *Computers, Environment and Urban Systems*, 59:244–255.

Max Kaufmann and Jugal Kalita. 2010. Syntactic normalization of twitter messages. In *International conference on natural language processing, Kharagpur, India*.

Brett Kessler. 1995. Computational dialectology in irish gaelic. In *Proceedings of the seventh conference on European chapter of the Association for Computational Linguistics*, pages 60–66. Morgan Kaufmann Publishers Inc.

William Labov, Sharon Ash, and Charles Boberg. 2005. *The atlas of North American English: Phonetics, phonology and sound change*. Walter de Gruyter.

Academia Argentina de Letras. 2008. *Diccionario del habla de los argentinos*. Emecé Editores.

Marcelo A Montemurro and Damián H Zanette. 2002. Entropic analysis of the role of words in literary texts. *Advances in complex systems*, 5(01):7–17.

Marcelo A Montemurro and Damián H Zanette. 2010. Towards the quantification of the semantic information encoded in written language. *Advances in Complex Systems*, 13(02):135–153.

John Nerbonne, Wilbert Heeringa, Erik Van den Hout, Peter Van der Kooi, Simone Otten, Willem Van de Vis, et al. 1996. Phonetic distance between dutch dialects. In *CLIN VI: proceedings of the sixth CLIN meeting*, pages 185–202.

Alexander Pak and Patrick Paroubek. 2010. Twitter as a corpus for sentiment analysis and opinion mining. In *LREc*, volume 10, pages 1320–1326.

Claude Elwood Shannon. 2001. A mathematical theory of communication. *ACM SIGMOBILE mobile computing and communications review*, 5(1):3–55.