What is on Social Media that is not in WordNet? A Preliminary Analysis on the TwitterAAE Corpus

Cecilia Domingo*, Tatiana Gonzalez-Ferrero*, and Itziar Gonzalez-Dios**

*University of the Basque Country (UPV/EHU)
**Ixa group, HiTZ center, University of the Basque Country (UPV/EHU)
[cdomingo003, tgonzalez023]@ikasle.ehu.eus, itziar.gonzalezd@ehu.eus

Abstract

Natural Language Processing tools and resources have been so far mainly created and trained for standard varieties of language. Nowadays, with the use of large amounts of data gathered from social media, other varieties and registers need to be processed, which may present other challenges and difficulties. In this work, we focus on English and we present a preliminary analysis by comparing the TwitterAAE corpus, which is annotated for ethnicity, and WordNet by quantifying and explaining the online language that WordNet misses.

1 Introduction

Natural Language Processing (NLP) tools and resources have usually been developed for major and standard varieties of language. Maintaining and updating these resources is expensive (Aldezabal et al., 2018), but recently open-source methodologies are being used to update the English WordNet (McCrae et al., 2020). However, well-known state-of-the-art tools in data processing that are used in industry and offer semantic analysis, such as NLTK (Loper and Bird, 2002) or knowledge-based word sense disambiguation tools like UKB, still rely on Princeton WordNet (Miller, 1995), which has not been updated for a long time.

On the other hand, NLP tools are being used nowadays in many industrial, marketing and social analysis projects and mainly social media text is being used. It is well known that social media may present several challenges when it comes to data processing, since, among others, non-standard varieties of languages or slang are used. Other problems related to this kind of texts are the length of the produced texts (sentences are rather short) and the difficulty of identifying useful information in order to separate it from what is not useful, such as hashtags, mentions or user-names (Farzindar and Inkpen, 2016).

In this paper, we explore the coverage of sociolects in WordNet to see what information state-of-the-art knowledge-based NLP tools may miss when analysing social media. In this preliminary analysis we use the TwitterAAE corpus (Blodgett et al., 2016), which contains geographic and census information. Exactly, we follow this methodology: a) we select a corpus with geographical/sociological information; b) we extract a sample of each group and we preprocess it; c) we compare it against an NLP resource, in our case, WordNet, and carry out a quantitative and qualitative analysis of the differences.

As a contribution of this preliminary study, we want to raise awareness of, on the one hand, how much linguistic diversity can be found in common data sources and, on the other hand, how risky it may be to use generic NLP tools and resources to process these diverse linguistic registers (colloquial, informal, internet language...) and sociolects for which the tools were not designed.

This paper is structured as follows: in Section 2 we present the corpus we have used and its preprocessing, in Section 3 we compare quantitatively and qualitatively the most frequent lemmas in the corpus to WordNet, in Section 4 we discuss our findings and expand upon some issues that concern the whole project. Finally, in Section 5 we conclude and outline the future work.

2 Corpus selection and preprocessing

The dataset used in this project was the publicly available TwitterAAE corpus (Blodgett et al., 2016), which consisted of 59.2 million publicly posted geolocated tweets (F. Morstatter and Carley, 2013). These data were collected in the United States in 2013. Each message was annotated considering the U.S. Census block-group geographic

area from which it was posted, meaning that ethnicity and race information associated with that district was taken into account. Four different covariants were established for the annotation: (non-Hispanic) Black, Hispanic (of any race), (non-Hispanic) Asian, and (non-Hispanic) White. This grouping reflects the main categories observed in the US census data, removing some smaller categories like "Native Hawaiian and Other Pacific Islander", and thus naturally as limited in nuance as the census categories may be (e.g. the census report groups all Black identities together, whether the respondents are African-American, African, Caribbean...). The terminology used in this paper will therefore reflect this simplified classification. For every user, the demographic values of all their tweets within the dataset were then averaged resulting in a length-four vector. The demographic data associated to each user and their corresponding tweets were then used by Blodgett and O'Connor (2017) to develop a mixed-membership probabilistic model that linked linguistic features with ethnicity. This model, the TwitterAAE model, assigned ethnicity values to the messages according to census data, giving each ethnicity a different proportion based on the "weight" of that ethnicity in the area where the tweet was written. If a tweet was assigned a proportion of more than 0.8, that meant that the tweet had a strong association with one of the previously mentioned demographic groups. A small sample of example messages belonging to each of the four covariants can be found in Table 1.

2.1 Sample selection

NLP tools are mainly trained to handle the standard variety of languages. Given that 2017 US census data reports 72.3 % of the population as exclusively white,[1] we assumed that tweets more strongly associated with the White demographic would be the most representative of standard English, although, as will be discussed after presenting our results, tweets even by the majority demographic present much non-standard language. To extract this sample of assumed standard English, we collected a subcorpus of tweets with the highest possible value of association with the White demographic, which was a proportion of 0.99 or higher. The new subcorpus consisted of around a million tweets.

In order to make a comparison, we needed to create other subcorpora of approximately the same size containing the messages of the other three demographic groups accounted for in the dataset. To create these subcorpora it was necessary to reduce the association value slightly, to a proportion of 0.9 or more, since otherwise the 'very' Black, Hispanic and Asian subcorpora size would have been significantly smaller than that of the 'very' White one.

2.2 Preprocessing

Prior to performing the analysis it was necessary to not only preprocess the subcorpora, but also the lemmas included in WordNet.

The Natural Language Toolkit (NLTK) provided us with the list of all lemmas included in WordNet. We used this version because it is the one that is used in NLP pipelines and applications. Since our goal was to identify the textual information that cannot be processed by WordNet, we needed to extract all the information that is in fact included in WordNet. We first filtered all the Multi-Word Expressions (MWE). It was necessary to work firstly with MWE and then with single-word lemmas separately in order to avoid overlap between the two (e.g. a MWE like "around the bend" has to be extracted first to avoid extracting the single-word lemmas "around" and "bend" separately and incorrectly). At the end of this process, we obtained two lists of lemma types to compare our corpora against.

Regarding the subcorpora, we selected 25 000 tweets for this preliminary analysis. We removed hashtags, mentions, emoticons, numbers, punctuation marks, white spaces, and other elements that did not provide useful information (Blodgett et al., 2018). Afterward, we removed all MWEs from the subcorpora by making use of the previously created file that listed all MWEs in WordNet. The next step was to extract Named-Entities with spaCy.[2] The remaining words were then lowercased, tokenized, and lemmatized, again with the aid of spaCy. Finally, we extracted all the single-word lemmas that were in the WordNet list, leaving only the tokens that could not be recognized.

Moreover, the Asian and Hispanic subcorpora contained a large number of tweets in languages other than English. The tweets in Spanish and Por-

[1] data.census.gov/cedsci/table

[2] https://spacy.io/

Tweet	Black	Hispanic	Asian	White
One min itss dhis, dhen another min itss dhat	**0.902**	0.0	0.0	0.098
Wont look out fa dis punk no mo	**0.91**	0.057	0.002	0.03
Well truth be told, I do not speak great Mexican	0.01	**0.936**	0.008	0.045
Okay, since I got no one to text, a dormir putos!	0.001	**0.93**	0.031	0.037
Y.O.L.O =[Y]ou [O]bviously [L]ove [O]reos	0.008	0.01	**0.956**	0.026
First person to bring me a midol at work wins best friend card for life. GO!	0.0	0.0	**1.0**	0.0
Spongebob will get his license before Taylor Swift finds love	0.0	0.005	0.001	**0.992**
I need to not be an old lady and learn to stay up past 8:30 #idontknowwhy #ihaveaproblem	0.0	0.0	0.0	**1.0**

Table 1: Examples from the Black, Hispanic, Asian and White subcorpora.

tuguese, due to their large number and the availability of language models, were also processed with spaCy to obtain data for the qualitative analysis. However, as Wordnet is an English resource, only the tweets in English were compared against the lists of Wordnet lemmas and used in the quantitative analysis. To detect the language of each tweet, the langid library[3] was used, as it showed the best combination of detail in output and accuracy of classification among the tools tested. The threshold to classify a tweet as English was set as 40; we arrived at this figure after several tests, to achieve optimal precision without much loss in recall.

3 Comparison to WordNet

In this section, we present a quantitative analysis of the lemmas found in each of the subcorpora and in WordNet. On the one hand, we have analyzed the number and percentage of lemmas and unique lemmas not found in WordNet. On the other hand, we have calculated the intersection of the subcorpora with the White subcorpus. Moreover, in the qualitative analysis we present the commonalities and the specifics of each subcorpus.

3.1 Quantitative Analysis

In Table 2, we show the number and percentage of lemmas[4] (repeated and unique) not found in WordNet for each subcorpus.

When we compare the two corpora that were almost fully in English, we observe that the White corpus contained 3501 more words that were not found in WordNet (not counting the words removed in preprocessing). However, removing the repetitions and counting each unique lemma only once reveals the opposite: there are 1994 more unique lemmas in the list of not-found lemmas from the Black corpus. This can be partly explained by the fact that the White corpus contained 8224 more pronouns. When we separate the lemmas found in both the Black and White corpora and look at the lemmas that are different, the list of unique not-found lemmas remains longer for the Black corpus than for the White one.

Looking at the data for the Hispanic and Asian corpora, it again seems that the White corpus posed the biggest challenge for WordNet, but this conclusion can again be discarded: if we include the tweets that were only classified as English with low confidence or that were classified as another language, the number of not-found lemmas rises to 79678 for the Asian corpus, and 31983 for the Hispanic. With regard to the unique lemmas, the number also rises significantly. The majority of these lemmas are in languages other than English. In the Hispanic corpus, however, there is a more balanced mix of Spanish and English lemmas.

When looking at the total amount of not-found lemmas in WordNet, there are 9542 fewer lemmas in the Hispanic subcorpus compared to the White one. Moreover, although the completely opposite happened with the Black corpus, the count of unique lemmas not found in WordNet for the White subcorpus was again considerably higher than those for the Hispanic one, more specifically 1160 lemmas of difference between them.

[3] https://github.com/saffsd/langid.py

[4] The (cleaned) lemma lists are available at ixa2.si.ehu.eus/notinwordnetwordlists.

Subcorpus	Total words (without tags, etc.)	Lemmas not found in WN	% of total tokens	Unique lemmas not found in WN	% of total tokens
Asian (only English tweets)	7290	916	12.706	218	3.023
Hispanic (only English tweets)	138222	20790	15.041	2061	1.491
Black	163549	26831	16.405	5215	3.188
White	228794	30332	13.257	3221	1.407

Table 2: Lemmas not found in WordNet, in absolute terms and relative to the size of each subcorpus

If we take a look at the rates of repeated lemmas, the Black and Hispanic corpora had the highest rate of not-found lemmas; for unique lemmas, it was the Asian and again the Hispanic corpora which had the highest rate. These data suggest that, even when people tagged as Asian and Hispanic users tweet in English, their language deviates more than that of the users tagged as Black and White from the standard English vocabulary recorded in WordNet. Users tagged as Black also seem to employ words not present in WordNet very frequently, but with less variety than the people tagged as members of the Asian group, who use more non-standard words, though with lower frequency. Overall, the users tagged in the Hispanic group proved the most problematic for an analysis reliant upon WordNet.

With regard to the Asian subcorpus, it must again be noted that its linguistic heterogeneity impedes any reliable quantitative comparisons. We will only mention that, even when we express the comparisons in relative terms to compensate for the small size of the English-language tweets of the Asian corpus, the Asian corpus has the lowest rate of unique lemmas in common with the White corpus. This suggests that the English written by the Asian and Black population according to the corpus may be the most different from the variant of the people tagged as White.

In Table 3, we present the intersection between the subcorpora and some illustrative examples. As we are comparing corpora of very different sizes, though we provide some quantitative data, we will focus on the qualitative analysis, which we believe will be of more value and which can be found below in Section 3.2.

3.2 Qualitative Analysis

As can be seen in Table 2, there is a large number of unique lemmas not found in WordNet that appear on one corpus but not on the one it is compared with. The only exception would be the Asian corpus, but this is easily explained by the small number of tweets in this corpus that were classified as English. The overall numbers seem indicative of a significant difference in the lexicon used by speakers of the sociolects reflected in each corpus. This difference can be corroborated by looking at some of the most common lemmas exclusive to each corpus. Due to the large number of lemmas to analyze, we only comment on the most frequent ones since lemmas ranked outside the top 30 already show very low frequencies.

3.2.1 Commonalities of all corpora

As was mentioned in the quantitative analysis, the corpora are not perfectly comparable, as the Asian and Hispanic corpora contain a large proportion of tweets in a language other than English. Still, a general look at all the corpora allows us to see some general characteristics of internet speech that are challenging for NLP tools, regardless of the user's dialect, or even language. It is important to bear in mind, though, that functional parts of speech are not included in WordNet, so understandably the list of common lemmas includes standard pronouns, prepositions or conjunctions. However, there are also many non-standard English (and Spanish and Portuguese) words in the list, and those are the kinds of words that seem to be characteristic of online writing:

- Onomatopoeia and forms of laughter: awww, hahahaha, lmao, kkkk (in Portuguese), jajaja (in Spanish)...
- Words with additional letters at the end: yess,yesss,yessss,yesssss...
- Acronyms: lbs, omg, smh, wtf...

Subcorpora	Exclusive lemmas	Examples
{BLACK (not WHITE)}	4787	*anotha, fckmuhlife, smoken*
{HISPANIC (not WHITE)}	1680	*definitly, samething, burritos*
{ASIAN (not WHITE)}	205	*twittrr, veryone, oleelo*
{WHITE (not BLACK)}	2793	*accidently, cheez, forsureeee*
{WHITE (not HISPANIC)}	2840	*memoryin, sweeet, hdache*
{WHITE (not ASIAN)}	3208	*bdtime, finaalllly, hunny*
{BOTH BLACK and WHITE}	428	*badass, freakin, gurl*
{BOTH HISPANIC and WHITE}	381	*yike, pendeja, hungover*
{BOTH ASIAN and WHITE}	13	*anything, boooo, skype*

Table 3: Count of unique lemmas not found in WordNet that exist in only one of the two corpora compared or that exist in both

- Joint words: bestfriend, forreal, goodnight, lemme, wassup...
- Shortened words: bday, dnt, prolly, txt...
- Words related to technology: retweet, Facebook, whatsapp...

Aside from these types of words and standard words in languages other than English, all the lists of lemmas not found in WordNet contained errors related to preprocessing:

- Named entities that were not recognized as such, possibly due to miscapitalization, and sometimes perhaps because they did not have the typical form of a named entity (e.g. the TV show "Buckwild", mentioned in several tweets, could be mistaken for an adjective or adverb).
- Lemmatization issues in English text, for example, "to poop" was incorreclty lemmatized despite being a well-established word, used in the currently most common sense since at least 1903, according to Merriam Webster's dictionary.[5] We also encountered something similar with the verb "to text", lemmatized as "texte". This error is more understandable, as "to text" has only existed for two decades; still, though perhaps this verb was not so much in vogue when WordNet was created, a modern lemmatizer should be able to deal with such a common verb.
- Lemmatization issues with other languages. Even though the focus of this project was on English-language processing, as spaCy also included models for Spanish and Portuguese, we tried its lemmatizer for the tweets in those languages and encountered more lemmatization problems. These were of a different nature: when a word could be an inflected form of more than one lemma, the lemmatizer tended to select the less frequent one (e.g. the Spanish and Portuguese preposition "para" was interpreted as a form of the verb "parir, meaning "to give birth").

3.2.2 The Black corpus

The meager length of the list of not-found lemmas common to the Black and White corpora strongly suggests a big difference between the sociolects reflected in each corpus. In the analysis of the most frequent lemmas of the Black corpus that were not found in WordNet, firstly, whereas among the lemmas from the White corpus we barely saw any mild profanity (e.g. "douchebag"), here we find several acronyms with "f" and two alternative spellings of the word "shit". All this is not to say that there is no actual strong profanity in either corpus: both corpora feature numerous instances of derivations and inflections of "fuck", but this is a standard word that is included in WordNet. Still, it is interesting to see that a search for forms of "fuck" returns almost twice as many hits for the Black corpus than for the White corpus. Though in online speech we see many acronyms and alternative spellings overall, in the case of profanity these transformations of words might actually serve a purpose: escaping

[5] https://www.merriam-webster.com/ (Accessed on 2020-06-16)

filters so that posts are not removed by moderators. Alternative spellings are overall very common in the Black subcorpus, as reflected by our list of frequent lemmas (e.g. “bruh”, “nomore”, etc.), sometimes reflecting non-standard pronunciations (e.g. “thang”), that are known to be characteristic of African-American English (AAE) (Kohler et al., 2007; Patton-Terry and Connor, 2010). Even though the list of lemmas from the White corpus was sparser, the alternative spellings in the top 28 most frequent lemmas from the Black corpus not found in WordNet had relatively high frequencies, which would justify more efforts to adapt NLP tools to accommodate them, at least if those tools are to process colloquial English.

3.2.3 The Asian corpus

For this corpus, given the small number of tweets written in English, the comparison between the Asian and the White corpus is of little relevance (less than 2 % of tweets in the Asian corpus were classified confidently as English). The English part of this subcorpus contained a large number of tweets from a traffic channel, which distorted the results and took most positions in the top-frequency words. Other frequent lemmas were laughter onomatopoeia in English and Portuguese. Nonetheless, tweets in English were a minority in this corpus (no more than 15 %, if we add the ones classified less confidently as English), so the majority of lemmas not found in WordNet were classified as Spanish and Portuguese. As WordNet is a resource for English, these lemmas were nothing exceptional, but rather ordinary Spanish and Portuguese words (e.g. in both languages the most frequent lemma that was not a preposition or adverb was the equivalent of ”make” or ”do”). Something less ordinary were the 135 variations of the ”jaja” laughter onomatopoeia in the Spanish file - illustrative of the wide variety of laughing expressions used online.

3.2.4 The Hispanic corpus

Although it also applies to the previously described subcorpora, it is surprising that, along with the acronyms and the most varied representations of laughter (“lmao”, “xd”, “hahah”) and agreement (“yeahh”, “yess”), joint words have a strong presence in the Hispanic subcorpus. This may well be due to the appearance of hashtags that have not been recognized as such in the pre-processing, and therefore have not been removed (e.g. “onedayilllooklike”, “whataburger”, “wordsyouneverwanttohear”), or because the user has intentionally got rid of the spaces between words since there is a character limit in the Twitter platform when writing messages. Whatever the reason, the employed NLP tools have not been able to recognize this phenomenon, and even though the lemmas that make up the joint words might be easily found in WordNet, they have remained unrecognized. However, and as one could have expected, the most characteristic feature of this subcorpus is the presence of Spanish words, even if the analysed tweets have been mostly written in English. Evidently, these terms are not found in WordNet. Lastly, it is worth mentioning that the Hispanic subcorpus contains several misspellings. One could say that the type of the observed typos are made quite recurrently by Spanish native speakers (“seriuosly”, “pasword”, “ecspecially”).

3.2.5 The White corpus

As in the Hispanic subcorpus, a noticeable characteristic of the list of lemmas of the White corpus is the variety of expressions denoting laughter (e.g. “hahah”, “hahahah”, “lolol”). Despite the variability, the most frequent onomatopoeias seem to be the shortest (two or three syllables) with a regular pattern of “(ha)*h”. Though very few onomatopoeia exist in WordNet (e.g. onomatopoeia that also function as verbs, like “moo”), the frequency of appearance of these laughter onomatopoeia would justify their inclusion in any NLP tool that could be considered suitable for handling tweets. As has been described, with a few exceptions, there seems to be a regular pattern in the formation of the different forms of laughter, so lemmatizers could be adapted to tackle the most frequent forms. Other frequent lemmas refer to technology (e.g. “snapchat”, “ipad”), understandably too modern to be processed by resources that are not updated regularly.

It is also interesting how some named entities escaped the NER filter applied during preprocessing. This highlights how named entities may adopt different forms in online discussions. For instance, the name of the Canadian singer Justin Bieber, though often spelled full and thus correctly spotted through NER, may also appear as simply “Bieber”, and something similar might happen with other celebrities. Also, we see an example of the popular internet trend of referring to TV shows or other popular sagas/bands/etc. by an acronym;

the American TV show Pretty Little Liars thus becomes "PLL". When the acronym is capitalized, it is recognized by our NER tool (only as an organization, though), but users online often do not care much for capitalization, and "pll" cannot be recognized as a named entity. Lastly, we must mention that several of the lemmas in the list of "white" lemmas were introduced by a single user, a weather channel ("wx", "lotemp", "hitemp").

4 Discussion

This preliminary experiment has allowed us to see that general NLP tools' performance on online, colloquial speech is suboptimal, especially with texts written by users outside the White demographic according the annotations of the corpus. We used the spaCy lemmatizer and NER tool, which are very popular nowadays, but even these modern tools had issues with some phenomena of internet speech: new terms, alternative spellings, new named entities and disregard for capitalization.

We have seen that WordNet was developed with standard English in mind and has not been updated for many years, so it fails to account for "modern" terms (we are considering tweets from 2013 relatively modern), online slang and diverse dialects. Interestingly, we saw that WordNet includes many multiword expressions (over sixty thousand), but the trend online seems to go in the opposite direction: expressions are shortened into acronyms (e.g. "lol", "omg"), and even single words are shortened (e.g. "bday, "txt").

As vast amounts of text are produced online daily, and this is of interest to businesses and researchers, there are initiatives that try to better deal with the type of language used online. For instance, Colloquial WordNet (McCrae et al., 2017) aims to be a more modern version of WordNet, one that includes popular, colloquial terms used online and SlangNet (Dhuliawala et al., 2016) gathers slang words and neologisms from the internet structured like in WordNet. It would certainly be worthy of study whether these resources recognize Twitter lexicon better; in our study, however, we did not perform any analysis using Colloquial WordNet, due to the difficulty in extracting its list of lemmas, at least in comparison with the easy method available in WordNet (a line of Python code suffices and returns text with no inconvenient tags) and SlangNet is not available.

WordNet does not include certain parts of speech, such as prepositions; it only includes "open class words". Nonetheless, as we have seen, internet users create new versions of "closed class words" (e.g. "eht" as a synonym of "at") or create words that merge words from both classes (e.g. "lemme", meaning "let me"). A deeper analysis of the words from our corpus belonging to or containing PoS not present in WordNet would be valuable, to consider whether such words should be added to semantic databases, or whether lemmatizers should be adapted to extract the standard form when processing new variants.

Though the focus of this project was on English-language text, it is important to emphasize the large number of tweets written in languages other than English, especially in the case of the Asian subcorpus. Therefore, any toolkit employed to process tweets from the US will need to include language detection and analysis tools for languages other than English - processing only English leaves many users behind and reduces the validity of any conclusions that might be extracted from analyzing tweets.

Future studies in this area, when possible, should also analyze a larger section of the TwitterAAE corpus. It is important to have a large corpus size to prevent a single user's repetitive lexicon from distorting the results. Alternatively, this type of users could be detected as part of preprocessing and their tweets excluded.

Even though the corpus has been very useful and is relatively modern, considering how fast language can change online, it would be necessary to replicate the methodology of Blodgett and O'Connor (2017) to annotate more recent tweets. The methodology could also be applied to other languages for which NLP tools and demographic data are available (e.g. to analyze dialects of Spanish or German). The resulting datasets would be very valuable for sociolinguistic studies, but also to assess NLP tools' inclusivity – are NLP tools leaving some groups of people behind? Nonetheless, it must be noted that, as any model, the one employed to annotate the dataset used here showed some inaccuracies. Though the sociolinguistic validation performed by Blodgett and O'Connor (2017) proved it quite accurate for AAE, classification in other categories seemed more problematic (e.g. the large number of Spanish tweets in the Asian category). This may

be due partly to the even larger diversity of Asian and Hispanic groups, which makes classifiying people into four categories overly simplistic at times (e.g. where do Brazilians go, being racially very diverse and culturally close to the Hispanic demographic but outside it?). Problems may also have arisen due to the source of the data used to build the model: the US Census is known to undercount minorities.[6] Even though race and ethnicity are self-reported, the way the data are aggregated is problematic for some groups, such as Middle-Eastern populations, Afro-Latinos or Portuguese speakers.[7] Moreover, the way the Census data were linked to the tweets may have also introduced some inaccuracies: geolocation may not have been perfectly exact,[8] and it may sometimes have been false, given the large number of internet users who connect through VPN.[9]

Finally, we would like to emphasize the same message that Blodgett and O'Connor (2017) leave at the end of their paper: African Americans are underrepresented in the Computer Science community, which makes it much harder for their voices to be taken into account. This conclusion is also valid for the Hispanic demographic, though for the Asian demographic there seems to be adequate representation.[10]

5 Conclusion and future work

In this paper, we have carried out an analysis of a corpus with geolocated tweets and we have compared the lemmas used to WordNet. As the corpus contained text from social media, we have dealt with non-standard language and we have seen that it still presents a challenge for mainstream NLP resources, which may put them at risk of leaving behind some speakers and varieties. As a result of this study, we encourage linguistic work on different registers and non-standard varieties.

In the future, we plan to expand the analysis to a bigger sample of the corpus and apply this methodology to study other languages e.g. Spanish with *Corpus de Referencia del Español Actual* (CREA) corpus. Moreover, we are preparing a list of candidate synsets to propose to the English WordNet (McCrae et al., 2020) following the open source and collaborative initiative. Moreover, we would like to study the possibility of adding register/ geographical information to synsets as e.g. Huber and Hinrichs (2019) are proposing for the Swiss variety of German. Analysing other Twitter tokens such as hashtags or mentions that were left out in the preprocessing could lead also to other studies.

[6] https://journalistsresource.org/studies/government/2020-census-research-undercount/

[7] https://www.census.gov/topics/population/race/about/faq.html

[8] https://www.singlemindconsulting.com/blog/what-is-geolocation/

[9] https://blog.globalwebindex.com/chart-of-the-day/vpn-usage-2018/

[10] https://www.wired.com/story/computer-science-graduates-diversity/

Acknowledgments

This work has been partially funded by the project DeepReading (RTI2018-096846-B-C21) supported by the Ministry of Science, Innovation and Universities of the Spanish Government, Ixa Group-consolidated group type A by the Basque Government (IT1343-19) and BigKnowledge – Ayudas Fundación BBVA a Equipos de Equipos de Investigación Científica 2018.

References

Izaskun Aldezabal, Xabier Artola, Arantza Díaz de Ilarraza, Itziar Gonzalez-Dios, Gorka Labaka, German Rigau, and Ruben Urizar. 2018. Basque e-lexicographic resources: linguistic basis, development, and future perspectives. In *Workshop on eLexicography: Between Digital Humanities and Artificial Intelligence*.

S. L. Blodgett and B. O'Connor. 2017. Racial disparity in natural language processing: A case study of social media african-american english.

Su Lin Blodgett, Lisa Green, and Brendan O'Connor. 2016. Demographic dialectal variation in social media: A case study of african-american english. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1119–1130.

Su Lin Blodgett, Johnny Wei, and Brendan O'Connor. 2018. Twitter universal dependency parsing for african-american and mainstream american english. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1415–1425.

Shehzaad Dhuliawala, Diptesh Kanojia, and Pushpak Bhattacharyya. 2016. SlangNet: A WordNet like Resource for English Slang. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4329–4332.

H. Liu F. Morstatter, J. Pfeffer and K. M. Carley. 2013. Is the sample good enough? comparing data from twitter's streaming api with twitter's firehose. In *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media*, pages 400–408.

A. Farzindar and D. Inkpen. 2016. Natural language processing for social media. *Computational Linguistics*, 42(4):833–836.

Eva Huber and Erhard Hinrichs. 2019. Including Swiss Standard German in GermaNet. In *Proceedings of the Tenth Global Wordnet Conference*, pages 24–32.

Candida T Kohler, Ruth Huntley Bahr, Elaine R Silliman, Judith Becker Bryant, Kenn Apel, and Louise C Wilkinson. 2007. African american english dialect and performance on nonword spelling and phonemic awareness tasks. *American Journal of Speech-Language Pathology*.

Edward Loper and Steven Bird. 2002. NLTK: The Natural Language Toolkit. In *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics*, pages 63–70.

John P McCrae, Ian Wood, and Amanda Hicks. 2017. The colloquial wordnet: Extending princeton wordnet with neologisms. In *International Conference on Language, Data and Knowledge*, pages 194–202. Springer.

John Philip McCrae, Alexandre Rademaker, Ewa Rudnicka, and Francis Bond. 2020. English WordNet 2020: Improving and Extending a WordNet for English using an Open-Source Methodology. In *Proceedings of the LREC 2020 Workshop on Multimodal Wordnets (MMW2020)*, pages 14–19.

George A Miller. 1995. WordNet: a lexical database for English. *Communications of the ACM*, 38(11):39–41.

Nicole Patton-Terry and Carol Connor. 2010. African american english and spelling: How do second graders spell dialect-sensitive features of words? *Learning Disability Quarterly*, 33(3):199–210.