

¿Es un plátano? Exploring the Application of a Physically Grounded Language Acquisition System to Spanish

Caroline Kery

University of Maryland Baltimore County, Baltimore, Maryland
ckery1@umbc.edu

Cynthia Matuszek

University of Maryland Baltimore County, Baltimore, Maryland
cmat@umbc.edu

Francis Ferraro

University of Maryland Baltimore County, Baltimore, Maryland
ferraro@umbc.edu

Abstract

In this paper we describe a multilingual grounded language learning system adapted from an English-only system. This system learns the meaning of words used in crowd-sourced descriptions by grounding them in the physical representations of the objects they are describing. Our work presents a framework to compare the performance of the system when applied to a new language and to identify modifications necessary to attain equal performance, with the goal of enhancing the ability of robots to learn language from a more diverse range of people. We then demonstrate this system with Spanish, through first analyzing the performance of translated Spanish, and then extending this analysis to a new corpus of crowd-sourced Spanish language data. We find that with small modifications, the system is able to learn color, object, and shape words with comparable performance between languages.

1 Introduction

With widespread use of products like Roombas, Alexa, and drones, robots are becoming commonplace in the homes of people. We can see a future where robots are integrated into homes to provide assistance in many ways. This could be especially beneficial to elders and people with disabilities, where having someone to help with basic tasks could be what allows them to live independently (Broekens et al., 2009). Natural language is an intuitive way for human users to communicate with robotic assistants (Matuszek et al., 2012a). Grounded Language Acquisition is the concept of learning language by tying natural language inputs to concrete things one can perceive. This field of study looks to train language and perceptual skills simultaneously in order to gain a better understanding of both (Mooney, 2008). Work in this field is critical for building robots that can learn

about their environments from the people around them.

For such a system to truly be useful for the average user, it is not enough to merely train a robot how to recognize everyday objects and actions in a lab. Much like toddlers who grow up in a family and surrounding culture, a service robot should be ideally able to learn the acronyms, nicknames, and other informal language that happens naturally in human interaction. It logically follows that a truly well-designed system should not only be able to handle vocabulary differences between users but also users that speak different languages. There are thousands of official languages spoken around the world, and many more dialects. In the United States alone, around 21 percent of residents speak a non-English language as their primary language at home (United States Census Bureau, US Department of Commerce, 2017). Grounded Language Acquisition takes many of its roots from Natural Language Processing, which in the past has had an unfortunate habit of focusing on English-centric methods. This often leads to systems that perform very well in English and “well enough” in other languages.

In this paper, we take an existing grounded language acquisition system (Matuszek et al., 2012b; Pillai and Matuszek, 2018) designed for grounding English language data and examine what adaptations are necessary for it to perform equally well for Spanish. We explore the extent to which machine translated data can assist in identifying linguistic differences that can impact system performance. We then collect a new comparable corpora of crowd-sourced Spanish language data and evaluate it on the system with and without our proposed modifications.

2 Previous Work

In this section, we describe relevant previous works in grounded language acquisition and multilingual natural language processing. While there has been past work to apply grounded language learning systems to multiple languages (Chen et al., 2010; Alomari et al., 2017) to our knowledge there have been few efforts in the space of non-English grounded language learning where comprehensive analysis was done to diagnose differences in performance between languages and work to mitigate these differences.

2.1 Grounded Language Acquisition

Language grounding can be done in many ways. There is a significant community within computer vision that works on object recognition with the help of captions (Krishna et al., 2017; Gao et al., 2015). These efforts ground objects found in images with words and relations stated in the captions. A multilingual example of this by Gella et al. (2017), used images as pivots between English and German image descriptions. This paper has a similar task of mapping language to images, but does so on a token level, and does not attempt to combine information between the Spanish and English corpora. In addition, the image data we are using includes depth information, as we are simulating the visual percepts of a robot. It must be noted that this differs from other works that use additional products of robotic percepts like video data, eye tracking, and other forms of gesture recognition (Chen et al., 2010; Alomari et al., 2017; Kollar et al., 2014; Yu and Ballard, 2004). In the robotics space, many works tie language grounding to enable actions like pathfinding (Matuszek et al., 2012a), household tasks (Alomari et al., 2017), and building (Brawer et al., 2018). While performing practical tasks is the eventual goal of our grounded language system, the current system focuses on the first step: building representations of objects and how they are described (nouns and adjectives).

There are a few examples of language grounding in multiple languages (Chen et al., 2010; Alomari et al., 2017). Several works tested their system in a language besides English and presented the results for both. While this showed that their systems could handle multiple languages, none provided an in-depth analysis into the differences in performance for their systems, or extrapolated

past the two languages. Our work seeks to examine and identify causes of differences in performance. While our current work only displays this system with Spanish, we plan to extend our framework to additional languages in the near future.

2.2 Multilingual Natural Language Processing

There is a strong multilingual community in the broader field of NLP working in many different aspects, such as machine translation (Wu et al., 2016) or multilingual question answering (Gao et al., 2015). Some works dive deep into specific language pairs to evaluate how differences between the languages complicate translation (Alomari et al., 2016; Gupta and Shrivastava, 2016; Ghasemi and Hashemian, 2016). Several work with Spanish and English specifically (Le Roux et al., 2012; Pla and Hurtado, 2014). Analyses such as these helped to shape our analysis when comparing the English and Spanish data performance, and enabled us to predict points where linguistic differences could impact performance.

There are quite a few examples in literature of taking a system designed for English and adapting it for multilingual use (Daiber et al., 2013; Gamon et al., 1997; Macdonald et al., 2005; Poesio et al., 2010; Jauhar et al., 2018). Sometimes this involves manually recreating aspects of the system to match the rules of the other language (Gamon et al., 1997), or utilizing parallel corpora to transfer learning between languages (Jauhar et al., 2018). Other projects look to make an English system “language agnostic” (not biased towards any one language) by editing parts of the preprocessing algorithm (Daiber et al., 2013; Wehrmann et al., 2017). The first method introduces a lot of additional complications such as manual rule-building, so it may seem attractive to make a system that is completely language-blind. The problem with this is that even generalized preprocessing techniques are often still biased towards languages with English-like structures (Bender, 2009), and in avoiding specifying anything about the language one can miss out on common properties within language families that could increase performance. For this paper, we strive to find common ground between making our system as generalized as possible and taking specific linguistic structures into account if necessary.

One significant difference between our research

and many works in grounded language acquisition is that our system is entirely trained off of noisy short descriptions collected without filtering. This has very different characteristics from the more common corpora built off of newswire and other forms of well-written text (a very common one is multilingual Wikipedia), or data that has been placed into structures like trees (Le Roux et al., 2012). Our data is prone to errors in grammar and misspellings; in this regard, our data is most like that of works that use Twitter data (Pla and Hurtado, 2014). However, in contrast to (Pla and Hurtado, 2014), our system only uses token extraction to find the relevant images to extract features from, rather than extracting all features from just the language.

3 Approach

In this paper, instead of building a new grounded language system, we chose to start with an existing system presented by Pillai and Matuszek (2018), which we will refer to as the GLS (Grounded Language System). This system attempted to learn physical groundings of colors, shapes, and objects by tying color and depth data derived from images of various items with natural language descriptions of the images. As a broader research goal, we seek to discover how effective the GLS is at handling non-English data. We decided to start with Spanish, due to it being very similar to English. We wanted to see if and how the slight differences between the two languages would affect the relative performance of the system.

To begin our analysis we explored the performance of the system on translated Spanish data with minimal modifications. Our analysis of these results concentrated on identifying language differences between Spanish and English that introduced new complications in grounding language. We used our insights from this analysis to inform our experiments on real Spanish data collected using Amazon Mechanical Turk.

3.1 Data

Pillai and Matuszek (2018) used a Kinect depth camera to collect images of fruit, vegetables, and children’s blocks of various shapes (see figure 1 for examples). There were a total of 18 object types, with four instances of each object. Each instance had around five images taken using the depth camera. For each of these images, RGB and

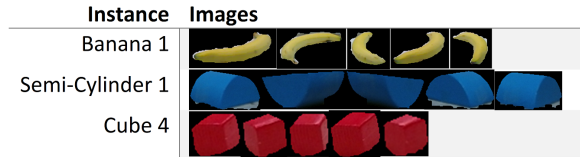


Figure 1: Examples of images of objects in the original dataset. Each object had several examples called “instances” and images of each instance were taken from several angles.

HMP-extracted kernel descriptors (Bo et al., 2011; Lai et al., 2013) were extracted from the RGB-D data. The authors then collected descriptions of these images in English using Amazon Mechanical Turk. About 85 descriptions were collected for each instance, for a total corpus of about six thousand descriptions. As we discuss in section 6, our own data collection process replicated this setup.

3.2 Grounding System

The GLS learned classifiers for meaningful tokens in the categories of color, shape, and object in an unsupervised setting. The system used the Mechanical Turk descriptions to identify which images were positive or negative examples of each token. Images that were described with a particular token often were assumed to be examples of that token. To find negative examples of tokens, the GLS used document similarity metrics to compare the descriptions for instances in vector space. The instances that were sufficiently far away in vector space from the identified positive instances for a token that had also never been described with that token were then chosen as negative examples of that token. For example, suppose the system were finding positive and negative instances for the token “carrot.” A positive instance identified might be “carrot 4.” In the document vector space, the instances with the descriptions most different from “carrot 4” would be “arch 1” and “cuboid 4,” while instances like “tomato 2” and “cucumber 3” are closer but still different enough to possibly qualify as negative examples of the token “carrot.”

Tokens that did not have any negative examples or had fewer than three positive examples were thrown away, with the assumption that there was not enough data to learn a classifier. The final classifiers were scored using the downstream task of image classification. Held-out positive and negative examples were presented, and the classifiers were judged by how well they could identify

which examples were positive or negative.

3.3 Our Modifications

Our research focused on taking the existing system and expanding it to work with Spanish. In the immediate sense, there were low level changes that had to be made throughout the code. English uses very few accents and many of the files had to have their encoding specified as unicode to handle non-ASCII characters. These changes, though minor, reflect a potential barrier to the application of research in new settings.

In addition to these minor fixes, more substantial changes had to be made to the system that preprocessed the image descriptions. The original GLS used an English WordNet lemmatizer. Lemmatizers are tools that take conjugated words like “walking” or “brushes” and attempt to turn them into un-conjugated versions like “walk” or “brush.” This step can be very helpful for making sure different versions of the same word are conflated. While this system worked well for English tokens, non-English lemmatizers proved difficult to find. Since we would ideally like our adaptations to the system to generalize well to other future languages, we decided to first remove the lemmatization step entirely, and later when this proved unsatisfactory for Spanish (see Sect. 5), we replaced the lemmatization step with a more available but rougher stemming step. Stemmers also attempt to remove conjugations from words, but they typically do so by chopping off common affixes without attempting to end up with a real word at the end. Words like “eating” will become “eat,” but words like “orange” may become “orang.”

Another step that we modified was the removal of “stop words.” In the original system non-meaningful words like “the,” “and,” “or,” and “if” were removed from the English data using a list of predefined words. This was an important step as it ensured that the system did not attempt to learn groundings for words like “and.” At the same time, we found that there were a number of words like “object,” “picture,” or “color” that were used so often in the descriptions that they held little physical meaning. These are designated as “domain-specific stop words,” which refer to words that in general cases hold meaning, but for the particular domain have been rendered meaningless by their frequent and varied use. We found that these words could be iden-

tified by their inverse-document-frequency (IDF), where each “document” is the concatenation of all descriptions for an instance.

4 Analysis with Translated Data

For our preliminary experiments, we only had access to the English corpora from Pillai and Matuszek (2018). We wanted to get baselines in how a Spanish corpora might perform. To do this, we translated the existing English phrases to Spanish through Google Translate’s API (Wu et al., 2016).

4.1 Translation Accuracy

As a sanity check on the quality of translation, the translated text was translated back into English (once again with Google translate’s API) and the English and back-translated English phrases were compared manually to see if their overall meanings were preserved. A total of 2,487 out of the 6,120 (around 40%) phrases remained exactly the same between translations. For the remaining 60%, five hundred back-translated phrases were randomly selected and manually compared to their original English version (see table 3 for examples). Approximately 87% of the phrases examined preserved their meaning between translations, so we estimated from this that about 90% of the phrases were translated accurately (shown in figure 2).

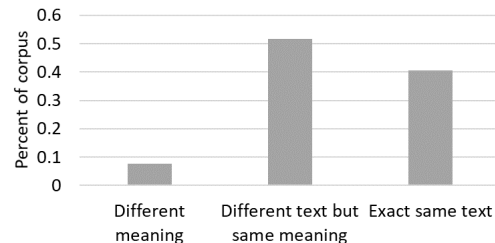


Figure 2: Breakdown of meaning preservation for English and English-Spanish-English translation.

For those phrases that did not translate accurately back to English, we observed a number of patterns. Some of them were simply due to ambiguities with the meaning of a word where the wrong one was selected during one of the translations (as an example, for the bottom row of table 3, “forma” can mean “shape” or “way”). A common example of this was the phrase “this is a red cabbage” becoming “this is a red wire,” which happened six times out of the five hundred selected phrases. Another error that occurred three times was “laying on its side” becoming “set

Image ID	Original English	Spanish Translation (Google API)	Back-translated English (Google API)	Same Meaning?
Orange 2	This fruit is called an orange	Esta fruta se llama naranja	This fruit is called orange	Yes
Cuboid 4	This is a picture of rectangular shaped blue coloured solid block	Esta es una foto de bloque sólido de color azul con forma rectangular.	This is a solid block photo of blue with rectangular shape.	No
Lime 2	It is a lime	Es una lima	It's a lime	Yes
Cuboid 2	This is a block The block is green The background is black The green block is laying on its side	Esto es un bloque El bloque es verde El fondo es negro El bloque verde está de lado	This is a block The block is green The background is black The green block is on its side	Yes
Cuboid 3	THIS IS A SHAPE	Esto es una forma	This is a way	No

Figure 3: Samples of English descriptions that were translated into Spanish and then back into English. The column on the right indicates if the meaning of the original English text matches the final back-translated English

aside,” since the Spanish phrase “puesta de lado” can mean “put sideways” but also “set aside.”

Other translation errors could be related to differences in Spanish and English structures. The pronoun “it” commonly became “he,” as Spanish nouns are gendered. Phrases with many adjectives saw them switching places with each other and the nouns they were attributed to. For example, “This is a picture of rectangular shaped blue coloured solid block” became “This is a solid block photo of blue with rectangular shape.” This confusion could be due to differences in the rules of adjective ordering between English and Spanish.

5 Scores for English and Google Translated Spanish

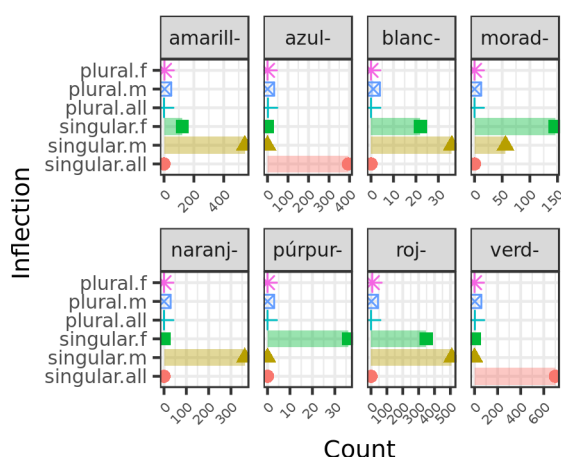


Figure 4: Proportion of color word forms in raw translated Spanish.

For the first experiment, we trained the model on the translated Spanish and English corpora with

minimal preprocessing (lowercasing and removing punctuation), and tested the color tokens only. Our goal was to get a baseline for how the system would perform using words that would be easy to compare between languages. It was expected that the Spanish corpus would perform worse, since it was not perfectly translated. When the tests were run, the translated Spanish did perform slightly worse (see figure 5), but an additional interesting issue emerged.

Spanish is a highly inflected language (Le Roux et al., 2012) and unlike English has adjective-noun agreement. This means that a simple color word like “red,” could translate to “rojo,” “roja,” “rojos,” or “rojas” depending on the gender and plurality of the noun it is describing. For the learning system this meant that the possible positive instances for color words could be split between the various forms, since different descriptions of the same object might use a different form depending on the structure of the sentence. We can see from figure 4, that in the overall translated corpus, the color words were split between different conjugations. This led to the hypothesis that some form of lemmatization or stemming would be necessary for Spanish, in a way that would have been less essential for English.

We processed both the translated Spanish and English descriptions with a Snowball stemmer (Porter, 2001). We chose this stemmer as it is readily available for a wide variety of languages through the nltk library. See results in Fig. 5.

We can see from figure 5 that applying stemming to the translated Spanish descriptions had a small positive effect on the F1-scores of the color

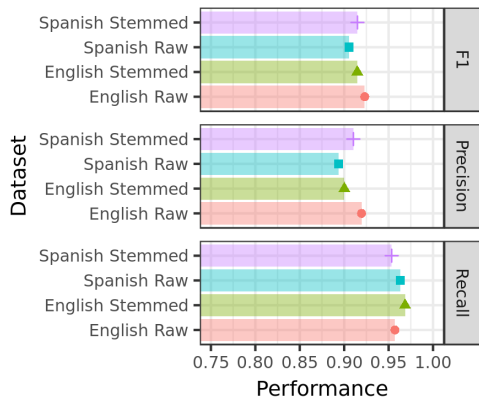


Figure 5: Average Scores for English and Google Translated Spanish color words.

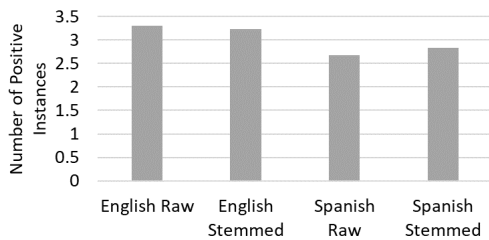


Figure 6: Average number of positive instances for English and Google Translated Spanish, stemmed and unstemmed.

classifiers. It also slightly raised the average number of positive instances per token, since stemming allowed instances that were split between small counts of several forms of a word to see them as the same word. We can see this in more detail in figures 7 and 8, which show the difference between the average of the scores for the various forms of color words in the unstemmed data (for example amarilla and amarillo would be averaged as amarill*), and the stemmed score of the stemmed form.

We can see in figure 7 that for the three colors shown, stemming always increased the average precision for that color, but could reduce recall. In addition from figure 8, we see that some of the colors had a large increase in average positive instances, while others did not. This was likely due to a case where many instances labeled with “rojo” also saw enough “roja” that it was a positive instance for both. When looking at the counts per instance, we found that for the 23 instances that had the token “roj” in their stemmed descriptions, 16 were positive examples of both “roja” and “rojo” in the un-stemmed version. For objects like cabbages (coles) and plums (ciruelas), “roja”

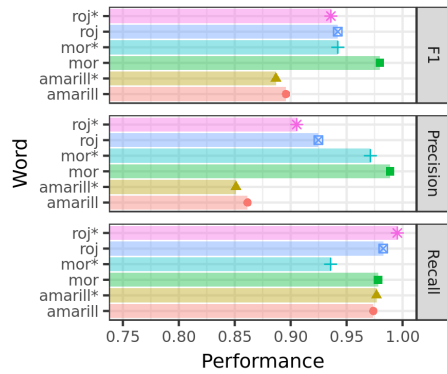


Figure 7: Comparison between the average of scores for various conjugations of color words (shown as *) and the scores of the stemmed versions.

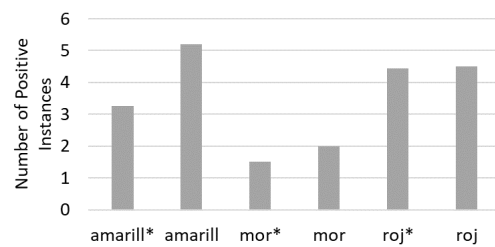


Figure 8: Comparison between the average number of positive instances across color word conjugations (see figure 6) and the number of positive instances of their stemmed forms.

was used dramatically more, while for tomatoes (tomates), cubes (cubos), and cylinders (cilindros) “rojo” appeared more.

As a final check, we examined the number of occurrences over all descriptions of each instance of the stemmed and un-stemmed versions of color words. For most of the colors, instances were often split between possible conjugations. For “amarill” (yellow), there were five instances where the individual counts of both un-stemmed forms of yellow: “amarillo” and “amarilla” were less than the threshold for a positive instance, while the stemmed version “amarill” was able to overcome that threshold. This is shown in the more dramatic increase in number of positive examples in figure 8. The effect on the scores is more complicated, since very yellow instances often had 50 or more occurrences of “amarill.” Because of the inherent messy nature of the data, instances with low but still significant counts of tokens (more than five occurrences) were much more likely to be falsely positive examples that could damage a classifier. We see this in figure 9 where the instance “eggplant 1” was called green seven times in the En-




Instance	Occurrences of "green"	Images
Cube 2	75	
Cucumber 2	31	
Eggplant 1	7	

Figure 9: Sample of instances that had more than five occurrences of “green” in the English corpora.

glish data. This is clearly because the stem of the fruit is green. However, a simple classifier may be confused by this instance, since it is mostly purple.

6 Collection of Real Spanish Data

Exploring comparisons between English and translated Spanish enabled us to get a basic idea of how Spanish descriptions might differ from English. However, in order to truly compare the languages, we needed to collect real Spanish data. We attempted to follow the methods described by [Pillai and Matuszek \(2018\)](#) as closely as possible to obtain comparable Spanish data to their English data. We utilized Amazon Mechanical Turk to collect Spanish descriptions of the images in the database.¹ In addition, workers were required to have at least fifty HITs accepted before being eligible to work on our HITs. To avoid biasing the workers towards a particular type of description, we provided no example descriptions.

We excluded data from a small number of workers who did not follow the directions (for example, responding in English or randomly selecting adjectives) and obtained additional high quality data to replace their submissions. All other submissions were accepted. This allowed for a wide variety of answers. One worker might simply name a carrot, while another would describe how it tastes, what foods it goes well in, or where it comes from. The English dataset was similarly noisy. This is desirable, as a robot that is trying to learn language from an average user must be able to handle the many ways in which a user might choose to introduce a new object.

One possible danger in collecting Spanish data that we considered was that someone might be responding in English and using a translation tool. We attempted to check for this by comparing our real Spanish data to the translated Spanish data. We found that short descriptions like “Esto es un limón” (this is a lemon) had a large amount of

¹This was accepted as an IRB exempt study.

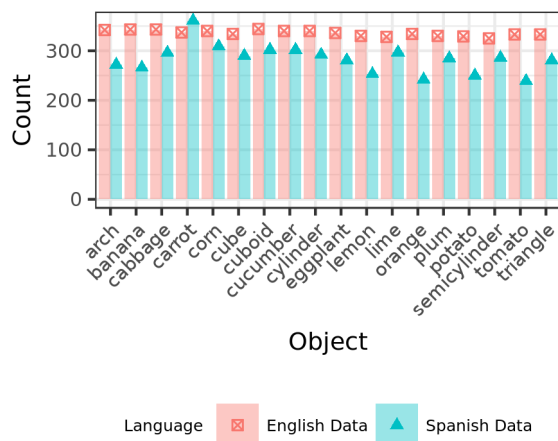


Figure 10: Total number of descriptions collected per object in Spanish and English.

overlap, but in general most of the Spanish descriptions were longer and did not mirror any of the translated results. In future work, we hope to find a better method to control for respondents who don’t actually speak the language, likely by requiring the completion of some short preliminary task like text summarization or more complex image captioning.

The total number of Spanish descriptions per object type was on average slightly lower than in the English corpus (see figure 10). We controlled for this in the results (section 7) by taking several random subsets of both corpora such that each instance had an equal number of Spanish and English descriptions and averaging the results.

7 Comparison of Spanish and English

7.1 Overall Scores

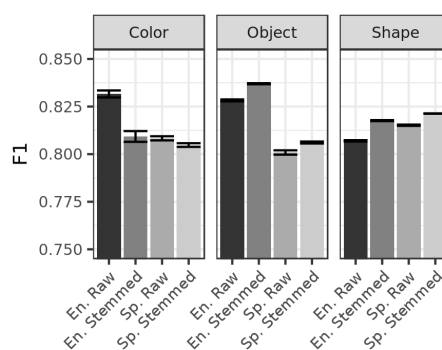


Figure 11: Average F1 scores for English and Spanish classifiers, stemmed and un-stemmed, for each classifier type. The error bars show the variance of these scores across all runs, which was fairly low.

In figure 11, we see the final averaged F1-Score for the color, shape, and object classifiers between the original English and the collected Spanish descriptions. Each score was found by averaging the results of twenty evaluation runs each of ten train-test splits. These scores were averaged across all tokens learned, without specifically sub-setting for the tokens that naturally represented colors, shapes, or objects. In general, the scores were fairly similar, varying between 0.8 and 0.84. From the small differences we see that stemming appeared to benefit the Spanish data for learning object and shape classifiers, but slightly hurt the performance for color classifiers. Un-stemmed English performed better than either Spanish version for color and object classifiers. Much like with Spanish, stemming appeared to help the shape and object classifiers, and hurt the color ones.

7.2 The Effect of Stemming

As one can see from figure 11, the effect of Stemming on the F1-Scores of the English and Spanish classifiers was not consistent. For both the object and shape classifiers, stemming appeared to either benefit or have little impact on the object recognition task. For the color tokens, stemming either barely impacted or lowered the scores.

Stemming can cause words to be conflated correctly or incorrectly. Incorrect stemming can certainly cause problems, where tokens are conflated that shouldn't be (Porter, 2001), or words that should be conflated are not. However, as discussed earlier, it is also possible for correct stemming to cause an instance to barely meet the threshold for being a positive example of a particular token 9, when perhaps that instance is not a good example of that token in reality. This was a particularly likely occurrence due to the inherent messiness of the data and the fact that the GLS based the classification label off of these messy descriptions. Due to this, and the high amount of conjugation in Spanish, it was decided that stemming likely would not negatively impact the learning system, and should most likely be employed.

7.3 Accents

One interesting difference that stood out when examining the real Spanish data was the use of accents. Unlike with the translated data, the real Spanish data was inconsistent with its usage of accents. While a majority of workers used accents where they were supposed to go, a not-

insignificant percentage of them left them out (see figure 12 for examples). This is likely because those workers did not have easy access to a keyboard with accented characters, and thus chose to leave them off. We can see in figure 12 that for common accented words, this had the effect of splitting the data. Luckily, the snowball stemmer (Porter, 2001) automatically removed these accents. We can see in figure 12 that after stemming, the counts for the accented and unaccented versions of the token were combined. The combined classifier did not always have a higher score on the testing data, for similar reasons to those discussed in section 7.2.

7.4 Stopwords

Without employing stop word removing during preprocessing, the system learned a total of ten words that could be classified as general stop words for English and eight for Spanish (see figure 13). This means that for these words, there was at least one instance where the word did not appear in any description. For Spanish, the tokens “de,” “es,” “una,” “y,” and “se,” and for English the tokens “this,” “is,” and “a” all had zero negative instances and were appropriately removed.

Figure 13 also shows tokens that appeared in the bottom 2% of tokens when sorted by IDF score. This was our way of estimating “domain-specific stop words.” Note that there were quite a few nltk stop words that also had very low IDF scores. The IDF method identified tokens like “object”, or “looks” which were used very often in the English descriptions and had little physical meaning. Figures 14 and 15 show how removing each type of stop word impacted the scores of the raw classifiers. For both languages, the greatest impact appeared to come from removing both general purpose stop words and low-IDF tokens, though the impact was small in all cases.

For the Spanish data, the tokens “amarillo” (yellow) and “roja” (red) were included in the bottom 2% of tokens by IDF score. These were common due to the prevalence of red and yellow objects in the dataset, suggesting a more nuanced approach such as lowering the threshold for the percent of low-IDF tokens to be thrown out.

8 Future Work

The work presented in this paper is ongoing. In the near future we intend to expand the analysis on

	Token	Count	F1-Score	Token	Count	F1-Score	Token	Count	F1-Score
English	corn	261	0.926508	banana	261	0.82946	lemon	252	0.907535
Spanish (accented)	maíz	117	0.802675	plátano	90	0.65640	limón	165	0.898421
Spanish (no accent)	maiz	65	0.793374	platan	47	0.66132	limon	118	0.866587
Spanish stemmed	maiz	182	0.835578	platan	140	0.65773	limon	283	0.840359

Figure 12: Object Scores for three Spanish that could be written with and without accents. Note that stemming removed accents, conflating stemmed and un-stemmed versions together.

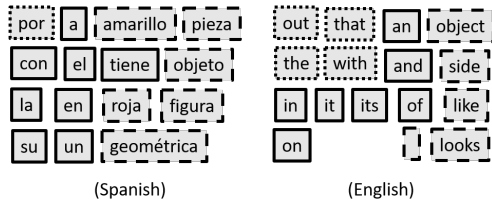


Figure 13: Stop words that appeared often enough to have classifiers trained on them. A dotted border indicates a stop word from the language's nltk stop word list. A dashed border indicates this token was in the top 2% tokens by ascending IDF score. A solid border means the token appeared in both lists.

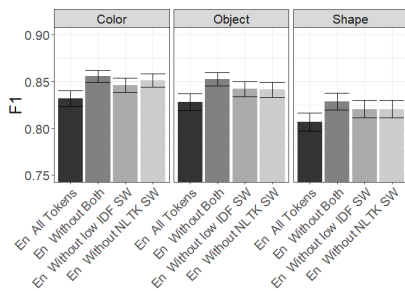


Figure 14: The impact on the average F1-score of removing nltk stop words versus removing the lowest 2% tokens by IDF score for English.

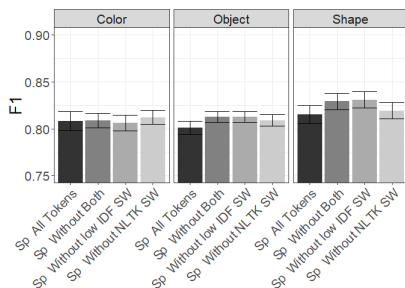


Figure 15: The impact on the average F1-score of removing nltk stop words versus removing the lowest 2% tokens by IDF score for Spanish.

the Spanish data. In addition many other possible techniques like spell-checking or synonym identification could be used to improve the ability of the system to handle the messy data.

A major next step for this research is to run our analysis on a language that is very different from English. For this, we intend to look next at Hindi. Hindi is the native language for hundreds of millions of people (India: Office of the Registrar General & Census Commissioner, 2015). It is from a different language family than English or Spanish, has a wide variety of dialects with small linguistic differences, and uses its own script. We anticipate that these properties will make Hindi a complicated and interesting language to analyze, and that doing so will introduce many new considerations for the grounded language system.

9 Conclusion

We have proposed adaptations to expand an existing unsupervised grounded language acquisition system (Pillai and Matuszek, 2018) to work with Spanish data. We discussed our initial observations with Google translated Spanish, and explored the extent to which these observations could be extended to real Spanish data collected through Amazon Mechanical Turk. Through our experiments, we were able to identify several differences between the two languages that had to be addressed in the system to attain comparable results. At the same time, we did not find that Spanish did significantly worse than English even before applying additional steps. In general, the existing system with slight modifications seems to work fairly well for both languages, which is promising when considering its applicability to real-life situations.

References

- Jawharah Alasmari, J Watson, and ES Atwell. 2016. A comparative analysis between arabic and english of the verbal system using google translate. In *Proceedings of IMAN'2016 4th International Conference on Islamic Applications in Computer Science and Technologies*, Khartoum, Sudan.
- Muhannad Alomari, Paul Duckworth, David C Hogg, and Anthony G Cohn. 2017. Natural language acquisition and grounding for embodied robotic systems. In *Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17)*.
- Emily M Bender. 2009. Linguistically naive!= language independent: why nlp needs linguistic typology. In *Proceedings of the EACL 2009 Workshop on the Interaction between Linguistics and Computational Linguistics: Virtuous, Vicious or Vacuous?*, pages 26–32.
- Liefeng Bo, Kevin Lai, Xiaofeng Ren, and Dieter Fox. 2011. Object recognition with hierarchical kernel descriptors. In *Computer Vision and Pattern Recognition*.
- Jake Brawer, Olivier Mangin, Alessandro Roncone, Sarah Widder, and Brian Scassellati. 2018. Situated human–robot collaboration: predicting intent from grounded natural language. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 827–833.
- Joost Broekens, Marcel Heerink, Henk Rosendal, et al. 2009. Assistive social robots in elderly care: a review. *Gerontechnology*, 8(2):94–103.
- David Chen, Joohyun Kim, and Raymond Mooney. 2010. Training a multilingual sportscaster: Using perceptual context to learn language. *J. Artif. Intell. Res. (JAIR)*, 37:397–435.
- Joachim Daiber, Max Jakob, Chris Hokamp, and Pablo N Mendes. 2013. Improving efficiency and accuracy in multilingual entity extraction. In *Proceedings of the 9th International Conference on Semantic Systems*, pages 121–124. ACM.
- Michael Gamon, Carmen Lozano, Jessie Pinkham, and Tom Reutter. 1997. Practical experience with grammar sharing in multilingual nlp. *From Research to Commercial Applications: Making NLP Work in Practice*.
- Haoyuan Gao, Junhua Mao, Jie Zhou, Zhiheng Huang, Lei Wang, and Wei Xu. 2015. Are you talking to a machine? dataset and methods for multilingual image question. In *Advances in neural information processing systems*, pages 2296–2304.
- Spandana Gella, Rico Sennrich, Frank Keller, and Mirella Lapata. 2017. Image pivoting for learning multilingual multimodal representations. *arXiv preprint arXiv:1707.07601*.
- Hadis Ghasemi and Mahmood Hashemian. 2016. A comparative study of” google translate” translations: An error analysis of english-to-persian and persian-to-english translations. *English Language Teaching*, 9:13–17.
- Ekta Gupta and Shailendra Shrivastava. 2016. Analysis on translation quality of english to hindi online translation systems- a review. In *International Journal of Computer Applications*.
- India: Office of the Registrar General & Census Commissioner. 2015. [Comparative speakers’ strength of scheduled languages -1971, 1981, 1991 and 2001](#). Archived 2007-11-30.
- Sujay Kumar Jauhar, Michael Gamon, and Patrick Pantel. 2018. Neural task representations as weak supervision for model agnostic cross-lingual transfer. *arXiv preprint arXiv:1811.01115*.
- Thomas Kollar, Stefanie Tellex, Deb Roy, and Nick Roy. 2014. Grounding verbs of motion in natural language commands to robots. In *Experimental Robotics. Springer Tracts in Advanced Robotics*, Springer, Berlin, Heidelberg.
- Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123:32–73.
- Kevin Lai, Liefeng Bo, Xiaofeng Ren, and Dieter Fox. 2013. Rgb-d object recognition: Features, algorithms, and a large scale benchmark. In *Consumer Depth Cameras for Computer Vision: Research Topics and Applications*, pages 167–192.
- Joseph Le Roux, Benoit Sagot, and Djamel Seddah. 2012. Statistical parsing of spanish and data driven lemmatization. In *ACL 2012 Joint Workshop on Statistical Parsing and Semantic Processing of Morphologically Rich Languages (SP-Sem-MRL 2012)*, pages 6–pages.
- Craig Macdonald, Vassilis Plachouras, Ben He, Christina Lioma, and Iadh Ounis. 2005. University of glasgow at webclef 2005: Experiments in per-field normalisation and language specific stemming. In *Workshop of the Cross-Language Evaluation Forum for European Languages*, pages 898–907.
- Cynthia Matuszek, Nicholas FitzGerald, Evan Herbst, Dieter Fox, and Luke Zettlemoyer. 2012a. Interactive learning and its role in pervasive robotics. In *ICRA Workshop on The Future of HRI*, St. Paul, MN.
- Cynthia Matuszek, Nicholas FitzGerald, Luke Zettlemoyer, Liefeng Bo, and Dieter Fox. 2012b. A joint model of language and perception for grounded attribute learning. In *Proceedings of the 2012 International Conference on Machine Learning*, Edinburgh, Scotland.

- Raymond Mooney. 2008. Learning to connect language and perception. In *Proceedings of the 23rd AAAI Conference on Artificial Intelligence (AAAI)*, pages 1598–1601, Chicago, IL.
- Nisha Pillai and Cynthia Matuszek. 2018. Unsupervised selection of negative examples for grounded language learning. In *Proceedings of the 32nd National Conference on Artificial Intelligence (AAAI)*, New Orleans, USA.
- Ferran Pla and Lluís-F Hurtado. 2014. Political tendency identification in twitter using sentiment analysis techniques. In *Proceedings of COLING 2014, the 25th international conference on computational linguistics: Technical Papers*, pages 183–192.
- Massimo Poesio, Olga Uryupina, and Yannick Versley. 2010. Creating a coreference resolution system for italian. In *International Conference on Language Resources and Evaluation*, Valletta, Malta.
- Martin F. Porter. 2001. Snowball: A language for stemming algorithms. *Retrieved March*, 1.
- United States Census Bureau, US Department of Commerce. 2017. *American community survey*. Data collected from 2012-2016.
- Joonatas Wehrmann, Willian Becker, Henry EL Cagnini, and Rodrigo C Barros. 2017. A character-based convolutional neural network for language-agnostic twitter sentiment analysis. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pages 2384–2391. IEEE.
- Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google’s neural machine translation system: Bridging the gap between human and machine translation. In *CoRR*.
- Chen Yu and Dana H. Ballard. 2004. A multimodal learning interface for grounding spoken language in sensory perceptions. In *ACM Transactions on Applied Perception*, pages 57–80.