Identifying Cross-Cultural Differences in Word Usage

Aparna Garimella and **Rada Mihalcea**
University of Michigan
{gaparna,mihalcea}@umich.edu

James Pennebaker
University of Texas at Austin
pennebaker@utexas.edu

Abstract

Personal writings have inspired researchers in the fields of linguistics and psychology to study the relationship between language and culture to better understand the psychology of people across different cultures. In this paper, we explore this relation by developing cross-cultural word models to identify words with cultural bias – i.e., words that are used in significantly different ways by speakers from different cultures. Focusing specifically on two cultures: United States and Australia, we identify a set of words with significant usage differences, and further investigate these words through feature analysis and topic modeling, shedding light on the attributes of language that contribute to these differences.

1 Introduction

According to Shweder et al. (1998), "to be a member of a group is to think and act in a certain way, in the light of particular goals, values, pictures of the world; and to think and act so is to belong to a group."

Culture can be defined as any characteristic of a group of people, which can affect and shape their beliefs and behaviors (e.g., nationality, region, state, gender, or religion). It reflects itself in people's everyday thoughts, beliefs, ideas, and actions, and understanding what people say or write in their daily lives can help us understand and differentiate cultures. In this work, we use very large corpora of personal writings in the form of blogs from multiple cultures[1] to understand cultural differences in word usage.

We find inspiration in a line of research in psychology that poses that people from different cultural backgrounds and/or speaking different languages perceive the world around them differently, which is reflected in their perception of time and space (Kern, 2003; Boroditsky, 2001), body shapes (Furnham and Alibhai, 1983), or surrounding objects (Boroditsky et al., 2003). As an example, consider the study described by Boroditsky et al. (2003), which showed how the perception of objects in different languages can be affected by their gender differences. For instance, one of the words used in their study is the word "bridge," which is masculine in Spanish and feminine in German: when asked about the descriptive properties of a bridge, Spanish speakers described bridges as being *big*, *dangerous*, *long*, *strong*, *sturdy*, and *towering*, while German speakers said they are *beautiful*, *elegant*, *fragile*, *peaceful*, *pretty*, and *slender*.

While this previous research has the benefit of careful in-lab studies that explore differences in world view for one dimension (e.g., time, space) or word (e.g., bridge, sun) at a time, it also has limitations in terms of the number of experiments that can be run when subjects are being brought to the lab for every new question being asked. We aim to address this shortcoming by using the power of large-scale computational linguistics, which allows us to identify cultural differences in word usage in a data-driven bottom-up fashion.

We hypothesize that we can use computational models to identify differences in word usage between cultures, regarded as an approximation of their differences in world view. Rather than starting with predetermined hypotheses (e.g., that Spanish and German speakers would have a different way of talking about bridges), we can use computational linguistics to run experiments on hundreds of words, and

[1] Throughout this paper, we use the term culture to represent the nationality (country) of a group of people.

consequently identify those words where usage differences exist between two cultures. We explore this hypothesis by seeking answers to two main research questions.

First, given a word W, are there significant differences in how this word is being used by two selected cultures? We build cross-cultural word models in which we use classifiers based on several classes of linguistic features and attempt to differentiate between usages of the given word W in different cultures. By applying them to a large number of words, these models are used to identify those words for which there exist significant usage differences between the two cultures of interest.

Second, if such significant differences in the usage of a word are identified, can we use feature analysis to understand the nature of these differences? We perform several analyses: (1) Feature ablation that highlights the linguistic features contributing to these differences; (2) Topic modeling applied to the words with significant differences, used to identify the dominant topic for each culture and to measure the correlation between the topic distributions in the two cultures; and (3) One-versus-all cross-cultural classification models, where we attempt to isolate the idiosyncrasies in word usage for one culture at a time.

2 Data

We base our work on personal writings collected from blogs, and specifically target word usage differences between Australia and United States. These two countries are selected for two main reasons: (1) they both use English as a native language, and therefore we can avoid the noise that would otherwise be introduced by machine translation; and (2) they have a significant number of blogs contributed in recent years, which we can use to collect a large number of occurrences for a large set of words.

We obtain a large corpus of blog posts by crawling the blogger profiles and posts from Google Blogger. For each profile, we consider up to a maximum of 20 blogs, and for each blog, we consider up to 500 posts. Table 1 gives statistics of the data collected in this process. We process the blog posts by removing the HTML tags and tagging them with part-of-speech labels (Toutanova et al., 2003).

Country	Profiles	Blogs	Posts
Australia	469	1129	320316
United States	374	1267	471257

Table 1: Blog statistics for the two target cultures.

Next, we create our pool of candidate target words by identifying the top $1,500$ content words based on their frequency in the blog posts, additionally placing a constraint that they cover all open-class parts-of-speech: of the $1,500$ words, 500 are nouns, 500 verbs, 250 adjectives, and 250 adverbs. These numbers are chosen based on the number of examples that exist for the target words; e.g., most (> 490) of the 500 selected nouns have more than 300 examples; etc. We consider all possible inflections for these words, for instance for the verb *write* we also consider the forms *writes*, *wrote*, *written*, *writing*. The possible inflections for the target words are added manually, to ensure correct handling of grammatical exceptions.

To obtain usage examples for the two cultures for these words, we extract paragraphs from the blog posts that include the selected words with the given part-of-speech. Of these paragraphs, we discard those that contain less than ten words. We also truncate the long paragraphs so they include a maximum of 100 words to the left and right of the target word, disregarding sentence boundaries. The contexts of the target words are then parsed to get the dependency tags related to the target word (Klein and Manning, 2003). We also explicitly balance the data across time. Noting there could be cases where the number of blog posts published in a specific year is higher compared to that in other years due to certain events (e.g., an Olympiad, or a major weather related event), we draw samples for our dataset from several different time periods. Specifically, for each culture, we consider an equal number of instances from four different years (2011-2014). Table 2 shows the per-word average number of data instances obtained in this way for each part-of-speech for each culture for the years 2011-2014.

Note that we do not attempt to balance the data across topics (domains), as we regard potentially different topic distributions as a reflection of the culture (e.g., Australia may be naturally more interested in water sports than United States is). We do instead explicitly balance our data over time, as described above, to avoid temporal topic peaks related to certain events.

Country	Noun	Verb	Adj	Adv
Australia	22461	18396	19206	19377
United States	15199	12347	12513	12952

Table 2: Average number of instances for the 1,500 target words for the years 2011-2014.

3 Finding Words with Cultural Bias

We start by addressing the first research question: given a word W, are there significant differences in how this word is being used by two different cultures? We formulate a classification task where the goal is to identify, for the given target word W, the culture of the writer of a certain occurrence of that word. If the accuracy of such a classifier exceeds that of a random baseline, this can be taken as an indication of word usage differences between the two cultures. We run classification experiments on each of the $1,500$ words described in the previous section, and consequently aim to identify those words with significant usage differences between Australia and United States.

3.1 Features

We implement and extract four types of features:

Local features. These features consist of the target word itself, its part-of-speech, three words and their parts-of-speech to the left and right of the target concept, nouns and verbs before and after the target concept. These features are used to capture the immediately surrounding language (e.g., descriptors, verbs) used by the writers while describing their views about the target word.

Contextual features. These features are determined from the global context, and represent the most frequently occurring open-class words in the contexts of the word W in each culture. We allow for at most ten such features for each culture, and impose a threshold of a minimum of five occurrences for a word to be selected as a contextual feature. Contextual features express the overall intention of the blogger while writing about the target word.

Socio-linguistic features. These features include (1) fractions of words that fall under each of the 70 Linguistic Inquiry and Word Count (LIWC) categories (Pennebaker et al., 2001); the 2001 version of LIWC includes about 2,200 words and word stems grouped into broad categories relevant to psychological processes (e.g., emotion, cognition); (2) fractions of words belonging to each of the five fine-grained polarity classes in OpinionFinder (Wilson et al., 2005), namely strongly negative, weakly negative, neutral, weakly positive, and strongly positive; (3) fractions of words belonging to each of five Morality classes (Ignatow and Mihalcea, 2012), i.e., authority, care, fairness, ingroup, sanctity; and (4) fractions of words belonging to each of the six Wordnet Affect classes (Strapparava et al., 2004), namely anger, disgust, fear, joy, sadness, and surprise. These features provide social and psychological insights into the perceptions bloggers have about the words they use.

Syntactic features. These features consist of parser dependencies (De Marneffe et al., 2006) obtained from the Stanford dependency parser (Klein and Manning, 2003) for the context of the target word. Among these, we select different dependencies for each part-of-speech: (1) nouns: root word of context (`root`), governor[2] if noun is nominal subject (`nsubj`), governor verb if noun is direct object (`dobj`), adjectival modifier (`amod`); (2) verbs: root, nominal subject (`nubj`), direct object (`dobj`), adjectival complement (`acomp`), adverb modifier (`advmod`); (3) adjectives: root, noun being modified (`amod`), verb being complemented (`acomp`), adverb modifier (`advmod`); (4) adverbs: root, adverb modifier (`advmod`). These features capture syntactic dependencies of the target word that are not always obtained using just its context.

[2] We follow the convention provided in `http://nlp.stanford.edu/software/dependencies_manual.pdf`

3.2 Cross-cultural Word Models

The features described above are integrated into an AdaBoost classifier.[3] This classifier was selected based on its performance on a development dataset, when compared to other learning algorithms. We compare the performance of the classifier with a random choice baseline, which is always 50%, given the equal distribution of data between the two cultures. This allows us to identify the words for which we can automatically identify the culture of the writers of those words, which is taken as an indication of word usage differences between the two cultures.

Throughout the paper, all the results reported are obtained using ten-fold cross-validation on the word data. When creating the folds, we explicitly ensure that posts authored by the same blogger are not shared between the folds, which in turn ensures no overlap between bloggers in training and test sets. This is important as repeating bloggers in both the train and the test splits could potentially overfit the model to the writing styles of individual bloggers rather than learning the underlying culture-based differences between the bloggers.

To summarize the cross-validation process: First, for each of the 1,500 target words, we collect an equal number of instances containing the given target word or its inflections from Australia and United States, from each of the selected years (2011-2014). We then divide the posts belonging to Australia and United States each into ten approximately equal groups, such that no two groups have bloggers in common. We finally combine the corresponding groups to form a total of ten bi-cultural groups that are approximately of equal size, which form our cross-validation splits. [4]

To compute the statistical significance of the results obtained, we perform a two-sample t-test over the correctness of predictions of the two systems namely, Adaboost and random chance classifiers. Disambiguation results that are significantly better ($p < 0.05$) than the random chance baseline of 50% are marked with $*$.

On average, the classifier leads to an accuracy of 58.36%*, which represents an absolute significant improvement of 8.36% over the baseline (a random chance of 50%). Table 3 shows the average classification results for each part-of-speech, as well as the number of words for which the AdaBoost classifier leads to an accuracy significantly larger than the baseline. These results suggest that there are indeed differences in the ways in which writers from Australia and United States use the target words with respect to all the parts-of-speech.

Part-of-speech	Average accuracy	Words with significant difference
Nouns	57.51*	393
Verbs	58.01*	395
Adjectives	59.25*	207
Adverbs	61.77*	215
Overall	58.36*	1210

Table 3: Average ten-fold cross-validation accuracies and number of words with an accuracy significantly higher than the baseline, for each part-of-speech, for United States vs. Australia.

4 Where is the Difference?

We now turn to our second research question: Once significant differences in the usage of a word are identified, can we use feature analysis to understand the nature of these differences?

4.1 Feature Ablation

We first study the role of the different linguistic features when separating between word usages in Australia and United States through ablation studies. For each of the feature sets specified in Section 3,

[3] We use the open source machine learning framework Weka (Hall et al., 2009) for all our experiments. We use the default base classifier for AdaBoost, i.e., a DecisionStump.

[4] This condition of equal data in each group is only approximate, as there will generally not be an exact division of bloggers with equal data. The average size of a cross-validation train split for a target word is 6246.57, while that for a test split is 819.88.

we retrain our concept models using just that feature set type, which helps us locate the features that contribute the most to the observed cultural differences.

The left side of Table 4 shows the ablation results averaged over the 1,500 target words for the four sets of features. We observe that the contextual and socio-linguistic features perform consistently well across all the parts-of-speech, and they alone can obtain an accuracy close to the all-feature performance.

Part-of-speech	**Loc**	**Con**	**Soc**	**Syn**	**All**	**LIWC**	**OF**	**ML**	**WNA**
Nouns	49.06	57.22*	56.52*	46.54	57.28*	56.62*	56.21*	54.17*	52.94
Verbs	47.97	57.65*	57.04*	47.71	57.90*	56.90*	56.28*	53.73*	53.25*
Adjectives	48.67	58.63*	57.90*	47.52	59.01*	58.03*	57.31*	55.42*	54.30*
Adverbs	50.72	61.09*	59.80*	46.85	60.81*	60.27*	59.80*	57.00*	56.57*
All	48.91	58.25*	57.47*	47.14	58.36*	57.55*	57.01*	54.70*	53.87*

Table 4: Feature ablation averaged over 1,500 target words. Loc: Local features, Con: Contextual features, Soc: Socio-linguistic features, Syn: Syntactic features, All: All feature types, LIWC: Linguistic Inquiry and Word Count, OF: OpinionFinder, ML: Morality Lexicon, WNA: WordNet Affect.

We also perform a feature ablation experiment to explore the role played by the various socio-linguistic features. The right side of Table 4 shows the classification accuracy obtained by using one socio-linguistic lexicon at a time: LIWC, OpinionFinder, Morality, and WordNet Affect. Among all these resources, LIWC and OpinionFinder appear to contribute the most to the classifier; while the morality lexicon and WordNet Affect also lead to an accuracy higher than the baseline, their performance is clearly smaller.

4.2 Topic Modeling

We next focus our analysis on the top 100 words (25 words for each part-of-speech) that have the most significant improvements over the random chance baseline, considered to be words with cultural bias in their use. The average accuracy of the classifier obtained on this set of words is 65.45%; the accuracy for each part-of-speech is shown in the second column of Table 7.

We model the different usages of the words in our set of 100 words by using topic modeling. Specifically, we use Latent Dirichlet Allocation (LDA)[5] (Blei et al., 2003) to find a set of topics for each word, and consequently identify the topics specific to either Australia or United States.

As typically done in topic modeling, we preprocess the data by removing a standard list of stop words, words with very high frequency ($> 0.25\% \times$datasize), and words that occur only once. To determine the number of topics that best describe the corpus for each of the 100 words, we use the average corpus likelihood over ten runs (Heinrich, 2005). Specifically, we choose that number of topics ($>= 2, <= 10$) for which the corpus likelihood is maximum.

For each data instance, we say that a topic dominates the other topics if its probability is higher than that of the remaining topics. For a given word, we then identify the *dominating topic* for each culture as the topic that dominates the other topics in a majority of data instances. We use this definition of dominating topic in all the analyses done in this section.

Quantitative Evaluation. To get an overall measure of how different cultures use the words that were found to have significant differences, we compute the Spearman's rank correlation between the topic distributions for the two cultures. For each topic, we get the number of data instances in which it dominates the other topics, in both cultures (Australia and United States). Subsequently, we measure the overall Spearman correlation coefficient between the dominating topic distributions for all 100 words. In other words, the distribution of topics is compared across cultures for each word. The Spearman coefficient is calculated as 0.63, which reflects a medium correlation between the usages of the words by the two cultures.

Qualitative Evaluation. For a qualitative evaluation, Table 5 shows five sample words for each part-of-speech, along with the identified number of topics and the dominating topic for Australia and United

[5] LDA has been shown to be effective in text-related tasks, such as document classification (Wei and Croft, 2006).

States. We associate labels to the hidden topics manually after looking at the corresponding top words falling under each of them.

As seen in this table, the number of topics that best describe each word can vary widely between two topics for words such as *start* and *economic*, up to ten topics (which is the maximum allowed number of topics) for words such as *color* or *support*. The dominating topics illustrate the biases that exist in each culture for these words; for instance, the word *teach* is dominantly used to describe academic teaching in Australia, whereas in United States it is majorly used to talk about general life teaching. Several additional examples of differences are shown in Table 5.

4.3 One-versus-all Classification

For additional insight into word usage differences between United States and Australia, we expand our study to develop word models to separate word usages in Australia (or United States) from a mix of ten different cultures. In other words, we conduct a one-versus-all classification using the same process as described in Section 3, but using Australia (United States) against a mix of other cultures to examine any features specific to Australia (United States).

In order to do this, we collect data from nine additional English speaking countries, as shown in the left side of Table 6. As before, the data for each country is balanced over time, and it includes an equal number of instances for four different years (2011-2014). The right side of Table 6 shows the average number of instances per word collected for each part-of-speech.

In this classification, for a given target word, one half of the data is collected from Australia (United States), and the other half is collected from the remaining countries, drawing 10% from each country. We run these classifiers for all the 100 words previously identified as having cultural bias in their use.

The average classifier accuracy for the Australia-versus-all classification, using ten-fold cross validation, is 64.23%, as shown in the third column of Table 7. We repeat the same one-versus-all classification for United States, with an average accuracy of 54.89%; the results of this experiment are listed in the last column of Table 7.

Overall, the performance improvement over the baseline is higher for Australia versus other countries (14.23% absolute improvement) than it is for United States versus others (4.89% absolute improvement). From this, we can infer that that the performance improvement over the baseline for the Australia versus United States task can be majorly attributed to the different word usages in Australia from the remaining countries. In other words, United States is more aligned with the "typical" (as measured over ten different countries) usage of these words than Australia is.

5 Related Work

Most of the previous cross-cultural research work has been undertaken in fields such as sociology, psychology, or antropology (De Secondat and others, 1748; Shweder, 1991; Cohen et al., 1996; Street, 1993). For instance, Shweder (1991) examined the cross-cultural similarities and differences in the perceptions, emotions, and ideologies of people belonging to different cultures, while Pennebaker et al. (1996) measured the emotional expressiveness among the northerners and southerners in their own countries, to test Montesquieu's geography hypothesis (De Secondat and others, 1748). More recently, the findings of Boroditsky et al. (2003) indicate that people's perception of certain inanimate objects (such as bridge, key, violin, etc.) is influenced by the grammatical genders assigned to these objects in their native languages.

To our knowledge, there is only limited work in computational linguistics that explored cross-cultural differences through language analysis. Our work is most closely related to that by Paul and Girju (2009), in which they identify cultural differences in people's experiences in various countries from the perspective of tourists and locals. Specifically, they analyzed forums and blogs written by tourists and locals about their experiences in three countries, namely Singapore, India, and United Kingdom, using an extension of LDA. One of their findings is that while topic modeling on tourist forums offered an unsupervised aggregation of factual data specific to each country that would be important to travelers (such as destination's climate, law, and language), topic modeling on blogs authored by locals showed cultural differences between the three countries with respect to several topics (e.g., fashion, pets, religion, health).

	Accuracy					NT	Dominating topic	
Word	All	Loc	Con	Soc	Syn		Australia	United States
Nouns								
store	67.14*	53.11*	67.15*	64.98*	50.79	10	ONLINE (blog, card, gifts, online, sale)	GROCERY (grocery, shopping, things)
support	66.05*	44.56	52.45	62.54*	51.29	10	EDUCATION (school, students, education)	LAW (public, police, law, social)
color	65.56*	50.74	67.35*	65.56*	50.27	10	BACKGROUND (pink, design, background)	PICTURE (paint, kit, picture, photo)
version	65.42*	47.75	65.47*	63.57*	47.28	4	RENDERING (story, thought, school)	ALBUM (song, music, album, classic)
phone	64.32*	53.93*	62.51*	58.19*	52.49	3	COMMUNICATION (calls, email, message)	FRIEND (night, friend, talking)
Verbs								
go	63.93*	47.29	64.15*	65.18*	42.77	2	TIME (time, day, night, love, today)	LIFE (life, world, work, god, children)
start	63.66*	46.80	63.46*	62.09*	54.08*	2	DAYBREAK (starting, day, love, morning)	SCHOOL (starting, school, softball)
know	63.51*	50.18	64.20*	62.40*	47.12	2	GOOD TIMES (love, good, things, life)	CHILDREN (children, school, year, book)
sing	63.54*	51.07	63.96*	58.19*	49.80	4	CHRISTMAS (christmas, happy, kids)	ROCK SINGER (band, guitar, rock, singer)
teach	62.66*	52.72	61.69*	59.88*	51.18	10	ACADEMICS (teachers, education, curriculum)	LIFE (time, young, life, thought, work, friends)
Adjectives								
various	65.64*	54.0*	64.70*	63.18*	48.40	10	POLITICS (political, party, war, revolution)	DIVERSITY (states, people, companies)
own	64.76*	56.89*	65.19*	62.71*	51.79	2	POWER (life, war, political, power)	LIFE (love, music, work, school)
economic	61.88*	44.82	42.11	59.16*	33.07	2	FINANCE (economy, financial, market, tax)	POLITICS (political, social, war, power)
old	66.45*	42.95	67.09*	64.74*	39.89	2	AGE (older, children, family, age)	PAST (back, school, days, love)
human	64.37*	52.78	60.92*	59.09*	48.14	9	RIGHTS (rights, law, freedom, civil)	LIFE (life, time, love, real)
Adverbs								
quite	73.56*	55.22*	70.98*	73.49*	51.50	2	EXTENT (time, back, good, love, thought)	EXTENT (time, back, good, love, thought)
else	67.35*	52.45	68.21*	64.66*	45.19	2	(make, find, world, invented, book, christ)	(time, life, night, love, work, home, god)
actually	65.62*	50.67	66.60*	65.51*	45.75	9	POLITICS (law war, people, government)	FAMILY (kids, fun, home, couple)
usually	68.37*	57.71*	70.16*	67.03*	44.35	2	SPORTS (softball, cricket, play)	ROUTINE (things, work, love, home)
certainly	64.64*	53.76*	63.25*	62.87*	43.03	2	SPORTS (game, series, softball, wrestling)	TIME (time, work, things, make)

Table 5: Five sample words per part-of-speech with significant usage difference in the Australian and American cultures. All: All the features, Loc: Local features, Con: Contextual features, Soc: Sociolinguistic features, Syn: Syntactic features, NT: Number of topics.

				Word instances			
Country	**Profiles**	**Blogs**	**Posts**	**Noun**	**Verb**	**Adj**	**Adv**
Barbados	440	830	32785	581	466	490	476
Canada	461	1097	397479	15015	12020	12965	12512
Ireland	473	978	231240	8936	7161	7919	7809
Jamaica	451	770	41495	1632	1318	1353	1297
New Zealand	450	1112	226900	8713	7313	7883	8284
Nigeria	464	908	223772	13719	9710	9796	7631
Pakistan	458	1404	135473	2861	2130	2243	1847
Singapore	406	803	208972	5623	5430	5447	6639
United Kingdom	473	934	282740	10887	9432	10021	11066

Table 6: Statistics for blog data collected for additional English speaking countries.

Part-of-speech	**United States vs. Australia**	**Australia vs. all**	**United States vs. all**
Nouns	65.54*	63.45*	57.07*
Verbs	64.20*	63.97*	53.87*
Adjectives	65.13*	64.36*	54.48*
Adverbs	66.92*	65.13*	54.13*
Overall	65.45*	64.23*	54.89*

Table 7: Ten-fold cross-validation accuracies averaged over the top 100 target words for United States vs. Australia; Australia vs. a mix of ten other countries; United States vs. a mix of ten other countries.

Yin et al. (2011) used topic models along with geographical configurations in Flickr to analyze cultural differences in the tags used for specific target image categories, such as cars, activities, festivals, or national parks. They performed a comparison over the topics across different geographical locations for each of the categories using three strategies of modeling geographical topics (location-driven, text-driven, and latent geographical topic analysis (LGTA) that combines location and text information), and found that the LGTA model worked well not only for finding regions of interest, but also for making effective comparisons of different topics across locations.

Ramirez et al. (2008) performed two studies to examine the expression of depression among English and Spanish speakers on the Internet. The first study used LIWC categories to process depression and breast cancer posts to identify linguistic style of depressed language. Significantly more first person singular pronouns were used in both English and Spanish posts, supporting the hypothesis that depressed people tend to focus on themselves and detach from others. The second study focused on discovering the actual topics of conversation in the posts using Meaning Extraction Method (Chung and Pennebaker, 2008). It was found that relational concerns (e.g., family, friends) were more likely expressed by depressed people writing in Spanish, while English people mostly mentioned medical concerns.

6 Conclusions

In this paper, we explored the problem of identifying word usage differences between people belonging to different cultures. Specifically, we studied differences between Australia and United States based on the words they used frequently in their online writings. Using a large number of examples for a set of $1,500$ words, covering different parts-of-speech, we showed that we can build classifiers based on linguistic features that can separate between the word usages from the two cultures with an accuracy higher than chance. We take this as an indication that there are significant differences in how these words are used in the two cultures, reflecting cultural bias in word use.

To better understand these differences, we performed several analyses. First, using feature ablation, we identified the contextual and socio-linguistic features as the ones playing the most important role in these word use differences. Second, focusing on the words with the most significant differences, we used topic modeling to find the main topics for each of these words, which allowed us to identify the dominant topic for a word in each culture, pointing to several interesting word use differences as outlined in Table

5. We also measured the correlation between the topic distributions for the top 100 words between the two cultures, and found a medium correlation of 0.63. Finally, we also performed a one-versus-all classification for these 100 words, where word use instances drawn from one of Australia or United States were compared against a mix of instances drawn from ten other cultures, which suggested that United States is a more "typical" culture when it comes to word use (with significantly smaller differences in these one-versus-all classifications than Australia).

In future work, we plan to extend this work to understand differences in word usages between a larger number of cultures, as well as for a larger variety of words (e.g., function words).

The cross-cultural word datasets used in the experiments reported in this paper are available at http://lit.eecs.umich.edu.

Acknowledgments

This material is based in part upon work supported by the National Science Foundation (#1344257), the John Templeton Foundation (#48503), and the Michigan Institute for Data Science. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation, the John Templeton Foundation, or the Michigan Institute for Data Science. Finally, we would like to thank Konstantinos Pappas and Steven Wilson for assisting us with the collection of the blog data.

References

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022.

Lera Boroditsky, Lauren A Schmidt, and Webb Phillips. 2003. Sex, syntax, and semantics. *Language in mind: Advances in the study of language and thought*, pages 61–79.

Lera Boroditsky. 2001. Does language shape thought?: Mandarin and english speakers' conceptions of time. *Cognitive psychology*, 43(1):1–22.

Cindy K Chung and James W Pennebaker. 2008. Revealing dimensions of thinking in open-ended self-descriptions: An automated meaning extraction method for natural language. *Journal of research in personality*, 42(1):96–132.

Dov Cohen, Richard E Nisbett, Brian F Bowdle, and Norbert Schwarz. 1996. Insult, aggression, and the southern culture of honor: An" experimental ethnography.". *Journal of personality and social psychology*, 70(5):945.

Marie-Catherine De Marneffe, Bill MacCartney, Christopher D Manning, et al. 2006. Generating typed dependency parses from phrase structure parses. In *Proceedings of LREC*, volume 6, pages 449–454.

Charles-Louis De Secondat et al. 1748. *The Spirit of Laws*. Hayes Barton Press.

Adrian Furnham and Naznin Alibhai. 1983. Cross-cultural differences in the perception of female body shapes. *Psychological medicine*, 13(04):829–837.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. 2009. The weka data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1):10–18.

Gregor Heinrich. 2005. Parameter estimation for text analysis. Technical report, Technical report.

Gabe Ignatow and Rada Mihalcea. 2012. Injustice frames in social media. Denver, CO.

Stephen Kern. 2003. *The culture of time and space, 1880-1918: with a new preface*. Harvard University Press.

Dan Klein and Christopher D Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pages 423–430. Association for Computational Linguistics.

Michael Paul and Roxana Girju. 2009. Cross-cultural analysis of blogs and forums with mixed-collection topic models. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3-Volume 3*, pages 1408–1417. Association for Computational Linguistics.

James W Pennebaker, Bernard Rimé, and Virginia E Blankenship. 1996. Stereotypes of emotional expressiveness of northerners and southerners: a cross-cultural test of montesquieu's hypotheses. *Journal of personality and social psychology*, 70(2):372.

James W Pennebaker, Martha E Francis, and Roger J Booth. 2001. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 71:2001.

Nairan Ramirez-Esparza, Cindy K Chung, Ewa Kacewicz, and James W Pennebaker. 2008. The psychology of word use in depression forums in English and in Spanish: Texting two text analytic approaches. In *ICWSM*.

Richard A Shweder, Jacqueline J Goodnow, Giyoo Hatano, Robert A LeVine, Hazel R Markus, and Peggy J Miller. 1998. The cultural psychology of development: One mind, many mentalities. *Handbook of child psychology*.

Richard A Shweder. 1991. *Thinking through cultures: Expeditions in cultural psychology*. Harvard University Press.

Carlo Strapparava, Alessandro Valitutti, et al. 2004. Wordnet affect: an affective extension of wordnet. In *LREC*, volume 4, pages 1083–1086.

Brian Street. 1993. Culture is a verb: Anthropological aspects of language and cultural process. *Language and culture*, pages 23–43.

Kristina Toutanova, Dan Klein, Christopher D Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 173–180. Association for Computational Linguistics.

Xing Wei and W Bruce Croft. 2006. Lda-based document models for ad-hoc retrieval. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 178–185. ACM.

Theresa Wilson, Paul Hoffmann, Swapna Somasundaran, Jason Kessler, Janyce Wiebe, Yejin Choi, Claire Cardie, Ellen Riloff, and Siddharth Patwardhan. 2005. Opinionfinder: A system for subjectivity analysis. In *Proceedings of hlt/emnlp on interactive demonstrations*, pages 34–35. Association for Computational Linguistics.

Zhijun Yin, Liangliang Cao, Jiawei Han, Chengxiang Zhai, and Thomas Huang. 2011. Geographical topic discovery and comparison. In *Proceedings of the 20th international conference on World wide web*, pages 247–256. ACM.