Generating Ethnographic Models from Communities' Online Data

Tomek Strzalkowski[1], Anna Newheiser[2], Nathan Kemper[2], Ning Sa[2], Bharvee Acharya[2] and Gregorios Katsios[1]
[1]Rensselaer Polytechnic Institute, Troy, NY, USA
[2]University at Albany, SUNY, Albany, NY, USA
{tomek, katsig}@rpi.edu, {anewheiser, nkemper, nsa, bacharya}@albany.edu

Abstract

In this paper, we describe computational ethnography studies to demonstrate how machine learning techniques can be utilized to exploit bias resident in language data produced by communities with online presence. Specifically, we leverage the use of figurative language (i.e., the choice of metaphors) in online text (e.g., news media, blogs) produced by distinct communities to obtain models of community worldviews that can be shown to be distinctly biased and thus different from other communities' models. We automatically construct metaphor-based community models for two distinct scenarios: debates on gun rights and marriage equality. We then conduct a series of experiments to validate the hypothesis that the metaphors found in each community's online language convey the bias in the community's worldview.

1 Introduction

Recent advances in machine learning, particularly deep learning, have led to successful exploitation of vast amounts of human-generated internet data and have produced remarkably accurate computational models of complex semantic and social phenomena in language, speech, vision, and other media, thus bringing us closer to the practical reality of artificial intelligence. These models are often considered objective and universal because the volume of data on which they are based is so vast that it is believed to be free of sampling limitations plaguing earlier research. And yet, the models that can be derived are only as good as the data from which they are built; the data, however vast, may still be biased. For one, people who post on the internet are not necessarily representative of the general population. Furthermore, society is composed of various communities and groups whose opinions and worldviews differ dramatically on a range of important issues. When data are oversampled from some sources over others (which can easily happen due to different rates of production), the resulting model is bound to be biased accordingly. This bias can lead to unwanted consequences, for example in government planning or resource allocation.

The flip side of the data bias, however, is that each community produces a unique information footprint that may be used to understand how its members perceive the world. The objective of our project is to investigate whether online information generated by various communities can serve as raw data for developing reliable and accurate ethnographic models of these communities, thus augmenting costly and limited-scale field studies. Clearly, the methods of computational ethnography will be different from its traditional counterpart and will rely extensively on substantial volumes of largely un-directed data, from which the critical information (e.g., relationships, opinions, conceptualizations) can be learned. One rich source of data is language, which serves as a communication vehicle, but also, as it evolves, encodes social, cultural, and often physical experiences of its user communities. These experiences are often vividly captured in the use of figurative language constructs, such as metaphors, that directly link abstract notions to collective physical experiences (Lakoff & Johnson, 1980; Thibodeau & Boroditsky, 2011). This observation has been confirmed in earlier work that constructed metaphor repositories across language and cultural dimensions (Western/American, Latin American/Mexican, Eastern European/Russian, Middle Eastern/Persian) related to such notions as government, economic inequality, and democracy (e.g., Shutova, 2010; Strzalkowski et al., 2013;

Proceedings of the Second Workshop on Figurative Language Processing, pages 165–175

https://doi.org/10.18653/v1/P17

Wilks et al., 2013; Mohan et al., 2013). Furthermore, within each linguistic-cultural society, various communities project their views on weighty social issues onto metaphorical language (Charteris-Black, 2002).

The objectives of the current Computational Ethnography (COMETH) project are thus twofold: (1) confirm experimentally that computational models of figurative language use capture communities' uniquely biased worldviews; and (2) demonstrate experimentally that such models, when used generatively, can mimic communities' reactions to novel information. Accordingly, the COMETH project developed an automated system that (a) rapidly ingests quantities of unstructured language data produced by communities of interest; and (b) uses natural language processing and machine learning techniques to construct ethnographic models for these communities. In this paper, we report preliminary results from applying this approach to two distinct scenarios: debates on gun rights and marriage equality.

2 Initial Case Study and Approach

Our initial approach was to develop and evaluate ethnographic models of two U.S. communities: (1) a community whose members prefer individual oversight of guns and prioritize gun rights, which we refer to as INDO; and (2) another community whose members prefer government oversight of guns and prioritize gun control, which we refer to as GOVTO. Our research demonstrates that these two communities' cultural models differ fundamentally from one another in their representation and valuation of concepts related to the gun debate. These concepts include broad notions such as gun rights and gun control, as well as narrower issues such as the Second Amendment, school shootings, and assault weapons. Each community is defined by the set of valuations they assign to these concepts. In order to extract these valuations, we identify culturally biased correlations, expressed in the use of metaphorical language, between key gun-related concepts and more basic, concrete, and imageable source domains, such as war, barrier, disease, animals, natural force, water, and foodstuffs. Additionally, we capture the prevalent sentiment that members of each community apply when referring to these concepts, in both literal and metaphorical contexts. We leverage language data available through public online sources produced by the target communities. These sources include mainstream media as well as public community blogs, newsletters, and websites. Data are collected and processed automatically using simple internet crawlers and natural language processing software capable of analyzing sentences for grammatical components, sentiment, and presence of metaphors. The processed data are then deposited in searchable structured repositories that are unique to each community. We estimate that the amount of data required to support both confirmatory and exploratory studies is approximately 5 million words per community. In the initial feasibility demonstration stage of this project, we collected 6 million words for both the INDO and GOVTO community.

In the remainder of this section, we provide detailed descriptions of data collection and processing procedures, as well as the repository construction process. These steps prepare us for the experiments described in Sections 5 and 6.

2.1 Data Harvesting and Processing

Our first objective was to identify the metaphors that are used to characterize the gun rights debate in the U.S. The topics of guns, gun rights, and gun control are well represented in U.S. media and finding related metaphors is not difficult. We use the automated system developed during the IARPA Metaphor Program (Strzalkowski et al., 2013) in order to extract examples of metaphors across a variety of information outlets.

For extracted metaphors to be useful for model development purposes, they must be assigned to a particular community or protagonist, in this case INDO or GOVTO, the two sides of the gun debate. To do so, we identified appropriate media outlets that cater to the INDO and GOVTO communities, as follows:

1. *Identification of spokespersons and spokesperson sites representing each community*. This step typically requires input from a cultural/political expert; however, it may be approximated using distance calculation based on metaphor distribution. For the gun debate scenario, we leveraged known correlations of opinions with the liberal-conservative political spectrum.
2. *Array sites along an opinion spectrum*. In most cases, there will be a spectrum of opinions within each community. We were initially particularly interested in the most orthodox

and extreme positions, because these provide the strongest contrast with other communities. This step is also assisted by input from cultural experts; however, it may be approximated by the Topical Positioning method (Lin et al., 2013) that tracks sentiment polarity in addition to metaphor choice. We leveraged Pew Research Center (2014, 2015, 2017) studies for establishing ground truth for both scenarios.

3. *Start collection of data from extreme positions*. This helps to establish a reasonably balanced collection of evidence from each community that can be used for confirmatory studies. Prior research (e.g., Shaikh et al., 2015) shows that the overall output of generally comparable communities may be quite unbalanced, depending on political context and related factors (Taylor et al., 2014).
4. *Data collection from sites of a more general nature.* These sites will be general news and opinion sites that may be considered relatively "opinion balanced" or "objective". These data provide a cultural backdrop against which the selected communities may be compared. In the current project, we collected such data as part of the exploratory marriage equality scenario.

This data collection and segmentation method is founded on the fact that language (including metaphors) is used as a group marker or a signal of group membership (Lakoff, 2001). Observation suggests that subgroups taking an extreme position define important markers for the more general group. Those in the middle of an opinion spectrum may employ language that reflects some of the extreme positions, some of the middling positions, and possibly some of the opposite positions – reflecting not only their middle-of-the-road approach to the issue, but also their willingness to identify with a range of views.

The positions of various participants in the U.S. gun debate range on a scale from radically in favor of government oversight of gun ownership, through a more moderate position in favor of this oversight, to a moderate position against such oversight, ending in a radical position against government oversight. In the U.S., this range corresponds roughly to a spectrum of U.S. political thought, arrayed typically on a scale from the radical left through the center to the radical right.

2.2 Identifying Metaphorical Targets

The process of identifying key concepts relevant to the case scenario has been fully automated. It proceeds in the following three steps:

1. *Locate frequently occurring topics in text*. The initial candidates are noun phrases, proper names (of locations, organizations, positions, events, and other phenomena, but less so of specific individuals). These are augmented with co-referential lexical items: pronouns, variants, and synonyms. The process of selection is quite robust but requires some rudimentary processing capability in the target language: part-of-speech tagging, basic anaphor resolution, and a lexicon/thesaurus.
2. *Down-select frequent topics to a set of 20-30 concepts*. The two key criteria are length and polarization. Topic length is measured by the number of references to the topic (either direct or indirect) that form "chains" across the "utterances" that are part of the scenario-related debate. Topic polarization is measured by the proportion of polarized references to the topic, either positive or negative. For example, the terms *gun rights* and *gun safety* are both frequently used and polarized in the gun debate.
3. *Select metaphorical targets*. Although all topics selected in Step 2 are important to the scenario, only some of them are likely to be targets of metaphors. We determine this simply by probing metaphor extraction for each of the selected topics and then eliminating these that do not bring back a sufficient number of metaphors or where the metaphor-to-literal ratio is too low. For example, "gun" is mostly used literally and is a poor metaphorical target. We used a 2% cut-off threshold for productive targets (a typical metaphor to literal ratio is 8-10%).

2.3 Data Collection Procedure

The data collection procedure consists of several steps as explained below. All steps are automated.

1. *Selection of target terms.* Target terms denote the key concepts of interest that the analyst wishes to investigate. For the gun debate, target concepts include *gun control*, *gun rights*, and *Second Amendment*, among others. This initial set of seed target terms need not be more than a few terms (e.g., less than 10).

2. *Search.* Selected data source websites were visited using an automated script. For sites that supported a search function, queries were posted directly to it. All text files matching any of the search terms were downloaded.
3. *Data cleaning.* All downloaded material was automatically segmented into passages so that at most five consecutive sentences were extracted: the sentence containing at least one search term, and up to two sentences on either side (before and after). Each full document yields one or more such passages, some of which may be overlapping.
4. *Data pre-processing*. All extracted passages were automatically pre-processed by a tokenizer that removes spurious characters and non-textual content and properly separates words.
5. *Target term set expansion*. Extracted passages were analyzed for presence of other terms besides the seed targets. All bigrams including only content words (not prepositions, determiners, etc.) were extracted and normalized for lexical variations. The most frequent bigrams were selected as additional target terms. This expansion was applied only once.

2.4 Scenarios Investigated

We investigated two distinct scenarios during the course of this project. The initial scenario involved two distinct views of gun rights versus gun control in the U.S. and developing ethnographic models of the communities representing these views. This initial scenario was partly based on the preliminary work conducted in the IARPA Metaphor program. The second scenario developed models for communities within the U.S. that hold different views on the topic of marriage equality, including same-sex marriage. Unlike the gun rights scenario that is essentially binary, the marriage equality topic produced multiple views, thus making the modeling task significantly harder. Nonetheless, we demonstrated that our approach can successfully support derivation of multi-faceted models. We summarize the first scenario only briefly; see Shaikh et al., 2015 for a complete description. The second scenario is described in more detail.

Gun Rights Scenario. Within the U.S., a public debate is ongoing concerning the Constitutionally and socially appropriate management of gun ownership, between those favoring Federal Government oversight (GOVTO) and those favoring individual oversight (INDO). At their extremes, the two sides are far apart. They view the issue in different conceptual terms, the GOVTO side relying heavily on DISEASE related metaphors and the INDO side relying on WAR related metaphors. These views appear reasonably constant over the years, even as the volume of output from each side changes.

Marriage Equality Scenario. Similar to gun rights versus gun control, people also disagree about the issue of marriage equality (i.e., same-sex marriage or gay marriage). Clashes in opinion on this topic became apparent during Obergefell v. Hodges (2015), the landmark case in which the U.S. Supreme Court ruled in favor of recognizing same-sex marriage. The lead-up to and aftermath of this case rippled through the media, with people voicing various stances on the issue. We identified seven basic stances one might take on the concept of marriage equality.

The first stance, labeled *expansion*, holds that we must nationally and internationally continue to expand rights for the LGBT community in regard to marriage, adoption, etc. The second stance, labeled *maintenance*, focuses on preserving the hard-won rights of gay couples and protecting them from infringement. The third stance, labeled *celebration*, is oriented toward commemorating the history of activism and legal battles that led to the Supreme Court decision legalizing same-sex marriage in the U.S. These three stances can be grouped together in the more general category of the *progressive community*, or those who believe that the institution of marriage should be open to all, regardless of sexual orientation or gender identity.

The fourth stance, labeled *reconciliation*, holds that traditional institutions such as the church should begin adapting to the changing moral and legal landscape surrounding marriage and family. The fifth stance, labeled *navigation*, is oriented toward working within changing laws surrounding marriage and family without compromising one's own values. These two stances can be grouped into the more general category of the *moderate community*, or those who default to legal precedent and consensus.

The sixth stance, labeled *incorrect interpretation*, holds that any extension of the institutions of marriage and family beyond

heterosexual couples is an incorrect interpretation of the concept of marriage. Finally, the seventh stance, labeled *infringement,* focuses on preventing emerging legal definitions of marriage and family from infringing on personal and religious liberties. These last two stances can be grouped into the more general category of the *traditional community*, or those who believe that marriage and family should be reserved for heterosexual couples and that it is not the place of the government to define these terms.

2.5 Data Sources

For the gun debate case scenario, we identified 62 internet sources that include both extreme and moderate positions on both sides of the issue. These sources included both mainstream news reporting (e.g., New York Times, The New Yorker, Fox News) as well as blogs and websites of relevant organizations (e.g., nra.com). In selecting data sources for the gun debate scenario, we relied on the fact that in the U.S. these issues align quite closely with the political spectrum. We could thus utilize publications such as Pew Research Center reports to identify initial media on the political left and right. Our final collection consisted of 33,000 documents from which 55,000 passages were extracted, for a of approx. 6.1 million words.

Data collection for the marriage equality scenario involved 75 online sources that yielded nearly 1 million text passages. After removing duplicates and ill-formed content, we obtained 620,000 passages (each containing up to 5 sentences) with the cumulative content of approx. 30 million words. As with the gun debate scenario, we deployed search terms that represent the most frequently used concepts in the domain. The larger size of the marriage equality dataset reflects its greater complexity of stances.

3 Metaphor Extraction Approach

We distinguish two levels of metaphor identification: (1) text (or linguistic) metaphors that consist of a metaphorical target, typically an abstract concept, and a relation adopted from a concrete source domain; and (2) conceptual metaphors that generalize across multiple occurrences of text metaphors involving the same target. While text metaphors have the semantic form of *shared-property (Target, Source)*, the conceptual metaphor usually conveys a more definite mapping *Target=Source* or *Target* $\in$ *Source*. For example, the textual metaphor "erosion of gun rights" alludes a shared property between gun rights and a geological landmark, thus invoking "Gun Rights is a Geological Landmark" conceptual metaphor. We note that conceptual metaphors are often implied rather than directly stated. Accordingly, the metaphor extraction process follows these two steps: we extract text metaphors first and then fuse them into conceptual ones. For ethnographic modeling purposes we use conceptual metaphors espoused in the language generated by a particular group of people.

We have developed a data-driven computational approach to extracting text metaphors that combines topical structure and imageability analysis in order to locate the candidate metaphorical expressions within text (Strzalkowski et al., 2013). To analyze topical structure, we identify nouns and verbs in a text passage and link their repeated occurrences, including co-references, synonyms, and hyponyms, and combine them into topic chains. Content words (e.g., verbs, nouns, adjectives) found outside these topical chains are candidate source relations if they also carry high imageability scores. Imageability ratings of most lexical items are looked up in an expanded MRC psycholinguistic database, which were built for several languages (Liu et al., 2014). The candidate relations are then used to compute and rank possible source domains in an emerging conceptual metaphor. Full details of the metaphor extraction process can be found in the cited papers.

Our approach to metaphor extraction is contrasted with more traditional computational approaches based on selectional restriction violations (Wilks, 1975; Fass, 1991; Martin, 1994; Carbonell, 1980; Feldman & Narayan, 2004; Shutova & Teufel, 2010; inter alia, also Shutova, 2010 for an overview) which do not scale well due to their heavy reliance on domain knowledge. More recent variants of this general approach (e.g., Rosen, 2018) utilize more robust deep learning methods but their utility remains limited to only some forms of text metaphors.

4 Metaphor-based Ethnographic Models

In this section, we outline the ethnographic models derived for each of the two scenarios. We provide only top-level characterization of each domain in terms of selection and distribution of metaphors

that define each community's viewpoint: two communities for the gun debate scenario and three communities for the marriage equality scenario.

4.1 Characterization of the INDO and GOVTO Metaphor Repositories

We applied the metaphor extraction system (Broadwell et al., 2013) to the 2018 GOVTO and INDO datasets. All passages were processed by the software to determine whether a target term was used metaphorically or literally. In both cases, the semantic relation involving the target term was identified so that sentiment toward the target could be computed. For metaphorical cases, the relations were further classified into one of several dozen metaphorical source domains (see Table 1), such as War, Disease, or Barrier. The processed passages form the metaphor repository database, from which community models are derived.

SOURCE DOMAIN	DEFINITION	ANCHOR TERMS
BARRIER	anything blocking someone from going somewhere or from doing something	barrier, obstacle, wall, obstruction
WATER	the part of the earth's surface covered with water (such as a river or lake or ocean)	watercourse, ocean, lake, river, pond, sea
DISEASE	a disordered or incorrectly functioning organ, part, structure, or system of the body	illness, sickness, ailment, disease, cancer
MAZE	a confusing network of intercommunicating paths or passages; labyrinth.	labyrinth, web, tangle, snarl, warren, maze
MEDICINE	any substance or substances used in treating disease or illness; medicament; remedy	medication, drug, remedy, medicine
FORCEFUL EXTRACT	to get, pull, or draw out, usually with special effort, skill, or force	pull, draw, extract, force out
WAR	a conflict carried on by force of arms, as between nations or between parties within a nation; warfare	warfare, combat, hostilities, war, battle, conflict

Table 1. *A subset of metaphorical source domains used in this study*

We also analyzed the distribution of metaphors with respect to the source domains. A source domain is a concrete semantic class to which the target is likened. The assignment of a metaphor to a source domain is determined by metaphorical relations that are applied to the target in a particular instance. For example, in "the plague of gun violence" the metaphorical relation is "plague," which is a sub-concept of DISEASE (e.g., in Wordnet; Miller, 1995). Therefore, this metaphor is classified as DISEASE metaphor – that is, "gun violence" is likened to disease.

Table 1 shows a partial list of source domains we used, along with definitions and anchor terms that are representative members of each domain. The complete list of 67 source domains was compiled by the IARPA Metaphor Program.

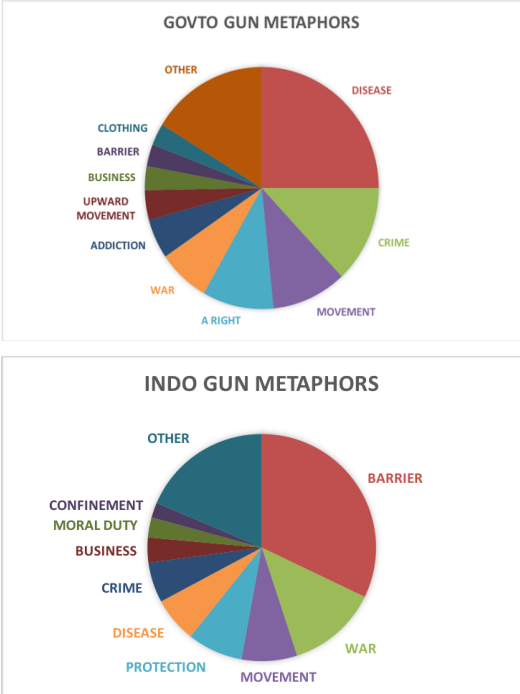

Figure 1: *Metaphor source domains for the gun debate scenario: DISEASE and CRIME dominate the GOVTO community, whereas BARRIER and WAR are most common within the INDO community.*

In Figure 1, we present the top choices of source domains for metaphors associated with the target concepts, including *gun control*, *gun rights*, and *gun violence*. We note that *DISEASE and CRIME* dominate on the GOVTO side, while *BARRIER and WAR* explain nearly half of INDO metaphors. This analysis illustrates one type of strong bias that is found in the data and confirms some earlier findings (Shaikh et al., 2015) over new data.

4.2 Characterization of Metaphors in the Marriage Equality Domain

We applied our metaphor extraction system to the marriage equality data, initially concentrating on the three major stances noted above (i.e., the progressive, moderate, and traditional communities). We used the same list of source domains as with the gun debate scenario. Overall,

we extracted 8305 metaphors including targets such as *marriage equality, same-sex marriage,* and *gay rights*. As expected, the selection of source domains was different than in the gun debate scenario, but again it showed marked contrast across the stances. Moreover, unlike in the gun debate scenario, we did not have an *a priori* classification of media sources as representing a particular stance. Instead, the set of all metaphors was split 3-ways using K-means clustering applied on the metaphor distribution statistics, taking into account the metaphor target, the metaphoric relation, the target role in the relation, and the source domain. Figure 2 shows near-perfect 3-way separation between sources representing progressive, moderate, and traditional views on marriage equality. A further attempt to separate the finer-grained seven stances described above was somewhat less successful, producing an Adjusted Rand Index (ARI) score of only 0.27, partly due to soft boundaries between some of the stances and insufficient data.

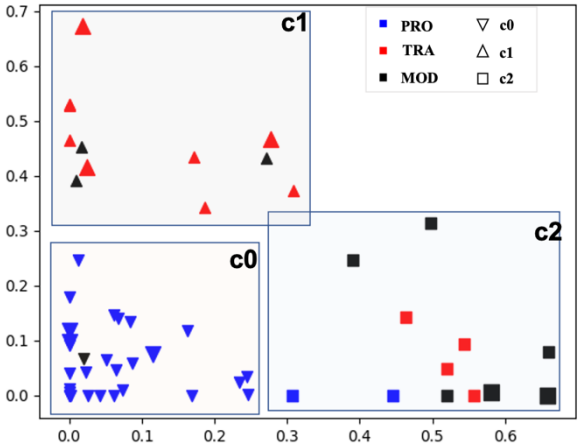

Figure 2. *Automatically derived clusters of information sources based on metaphor distribution show a good split between progressive (c0), traditional (c1), and moderate sources (c2), with ARI of 0.69.*

We note that alternative approaches to obtaining automatic separation of stances based on topic and sentiment distribution did not come close to the result seen in Figure 2. A doc2vec based method (Le & Mikolov, 2014) only achieved an ARI score of 0.17 on the 3-way split; an LDA-based approach (Blei et al, 2003) did only slightly better at 0.37.

Figures 3 to 5 show the metaphor distribution across the three main stances in the marriage equality domain. The first analysis (Figure 3, Table 2) shows metaphor distribution in language collected from progressive sources. The dominating metaphor is Forceful Extraction, which involves relations such as "ban" and "prohibit." Other common metaphors, along with their frequent relations, are shown in Table 2.

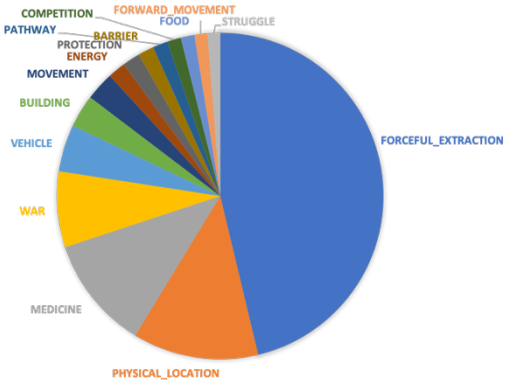

Figure 3. *Distribution of marriage equality metaphors in the progressive stance sources*

Marriage Equality is...	Sample Metaphorical Relations
forceful extraction (38.4%)	ban, prohibit
physical location (10.3%)	disagree, divide, poll, reiterate, strike, subject
medicine (9.3%)	legalize, ban, approve
war (6.2%)	battle, fight, defend, victory
vehicle (3.9%)	overturn, drive, ban
building (2.7%)	enter, condemn, preside, celebrate, way for
movement (3.4%)	embrace, move, movement, proceed

Table 2. *Selected metaphoric relations for the most frequent source domains in the progressive stance.*

Figure 4 and Table 3 show the analysis of the moderate stance. The most frequent metaphor, representing about 23% of all collected examples, is Physical Location. This is followed by Medicine, which explains another 14% of the examples. This community had a relatively low output volume, producing a mere 5% of metaphors in our data set, with another 20% attributed to "neutral" sources.

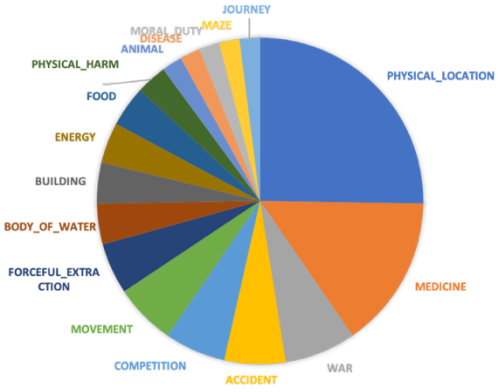

Figure 4. *Distribution of marriage equality metaphors in the moderate stance sources.*

Marriage Equality is...	Sample metaphorical relations
physical location (22.7%)	argument, recognition
medicine (13.6%)	legalize, treat
war (6.4%)	undermine, authorize
accident (5.5%)	rule
competition (5.5%)	promote
movement (5.5%)	embrace, perform
forceful extraction (4.6%)	ban
body of water (3.6%)	spawn, surround
building (3.6%)	enter

Table 3. *Selected metaphoric relations for the most frequent source domains in the moderate stance.*

Figure 5 and Table 4 show the analysis of the traditional stance on marriage equality. Here the dominating metaphor is Medicine (20%). When coupled with a related and quite frequent metaphor of Addiction (5%), these two together form a hybrid "bad medicine" metaphor. Other frequent metaphors (Physical Location and Forceful Extraction) are also quite visible, which can be explained partly by the mixed content of the traditional stance cluster as well as frequent critical references to the progressive sources. At this time, we lack reliable means, beyond distribution frequency, of separating expressions that characterize one's own stance as compared to other people's stances. We note that sources representing traditional views account for only about 10% of extracted metaphors.

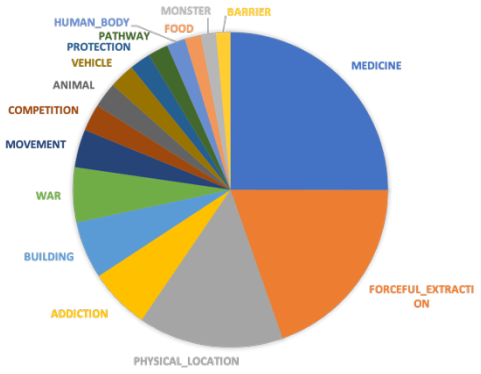

Figure 5. *Distribution of marriage equality metaphors in the traditional stance sources.*

Marriage Equality is...	Sample metaphorical relations
medicine (20.2%)	legalize
forceful extraction (15.8%)	ban, prohibit, force
physical location (12%)	disagree, subject, argue
addiction (5%)	legalization
building (4.7%)	enter, way for, preside
war (4.5%)	fight, battle, defend, undermine
movement (3.1%)	perform, embrace

Table 4. *Selected metaphoric relations for the most frequent source domains in the traditional stance.*

5 Confirmatory Studies on the Gun Debate

We conducted a study to experimentally confirm the bias in community data in relation to the gun debate scenario. A subset of passages containing metaphors was selected from both the INDO and GOVTO communities' metaphor repository and were displayed to human participants, whose task was to categorize each passage as advocating for either individual or government oversight of guns. The objective of this study was to confirm that the bias was captured in the metaphors used by each community and that this bias can be detected by human raters. We thus predicted that participants would be able to categorize the passages as representing the intended community viewpoints at rates above chance. This result would confirm that our metaphor repositories accurately reflect the language use of the two communities relevant to this target scenario (i.e., INDO and GOVTO).

A sample of 338 respondents completed the study via Amazon Mechanical Turk. Raters viewed 20 passages from INDO sources and 20 passages from GOVTO sources. Overall accuracy scores were calculated by dividing the total number of correct categorizations by the total number of passages (i.e., 40). As predicted, participants categorized passages with above-chance accuracy (mean accuracy=66%, SD=14%), $t(337)=21.94$, $p<.001$, $d=1.19$. INDO and GOVTO categorization accuracy scores were calculated by dividing the number of correct categorizations for each passage type by 20. Participants categorized passages from GOVTO sources (mean accuracy=70%, SD=15%) more accurately than passages from INDO sources (mean accuracy=63%, SD=16%), $t(337)=7.62$, $p<.001$, $d_z=0.41$. Thus, human raters were able to

determine whether passages came from INDO or GOVTO media sources at reliably greater-than-chance (above 50% accuracy) rates.

We replicated this study with another sample of 906 participants who rated 40 randomly selected (vs. researcher-selected, as in the previous study) passages from the same metaphor repositories on a continuous scale (0=*definitely in favor of government oversight* to 100=*definitely in favor of individual oversight*). As predicted, participants rated INDO passages as being reliably more in favor of individual oversight than the scale midpoint (M=56.48, SD=11.26), $t(905)$=17.47, p<.001, d=0.58. Participants also rated GOVTO passages as being reliably more in favor of government oversight than the scale midpoint (M=38.31, SD=12.71), $t(905)$=-27.62, p<.001, d=0.92. These results again suggest that participants were able to detect the bias present in the passages and rated them accordingly.

6 Confirmatory Study on the Marriage Equality Debate

We conducted a study to confirm the bias in community data in relation to the marriage equality scenario. Following the same procedure as in the first study, 285 participants categorized a total of 45 passages automatically selected from progressive, moderate, and traditional sources as representing a progressive, moderate, or traditional stance on marriage. As predicted, participants categorized passages with above chance accuracy, with 33% accuracy representing chance (mean accuracy=38%, SD=8%), $t(284)$=11.49, p<.001, d=0.68. Moreover, participants categorized passages from moderate sources (M=41%, SD=16%) more accurately than passages from progressive sources (mean accuracy=36%, SD=18%), $t(568)$=-3.70, $p_{bonferroni}$<.001, d_z=-0.20. Accuracy scores for passages from traditional sources (mean accuracy=38%, SD=15%) did not differ from any other accuracy score. Thus, human raters were able to determine whether passages regarding marriage equality originated from progressive, moderate, or traditional media sources at reliably greater-than-chance rates.

7 Conclusion

In this paper, we presented results from a project in which we built ethnographic models of communities based on the choice of metaphors in online language use. We investigated two distinct scenarios, a binary gun rights vs. gun control debate and a non-binary marriage equality debate. We demonstrated in both cases that metaphor choice provides strong clues to a community's identity and bias, and that automatically derived models can adequately delineate target communities. Future work will focus on improving the accuracy of metaphor classification and exploiting other forms of figurative language that capture deeply held collective meanings and stances.

This research explored a new avenue of computational ethnography, conducted entirely online. The long-standing benchmark is traditional field ethnography that involves typically small-scale field work, which takes months or years to complete. Such studies, while producing "thick" models, are not always feasible, especially in the areas of conflict or disturbance nor when a rapid response is required. Furthermore, field ethnography by its nature is heavily reliant on regional, local language speaking subject matter experts who may be hard to find due to lack of expertise or risks involved.

Current approaches to online ethnography include application of traditional methods of observation of online behavior aimed at modeling online communities, as opposed to the real-world communities (e.g., Miller & Salter, 2000; Safar & Mahdi, 2012). Efforts aimed at deriving offline models from online data are either small-scale (Martey et al., 2011) or limited to superficial analysis of social media (e.g., sentiment extraction) that cannot easily separate transient views from entrenched opinions (e.g., Turney & Littman, 2003). These approaches produce low-quality results also due to over-sensitivity to data noise and are thus unreliable. We believe that our research shows the value of advanced sociolinguistic analysis and natural language processing in studying the online human terrain.

Acknowledgments

This work was supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR0011-18-9-0016. All statements of fact, opinion or conclusions contained herein are those of the authors and should not be construed as representing the official views or policies of USN, DARPA, or the U.S. Government.

References

Blei, David M., Andrew Y. Ng, and Michael I. Jordan (2003) Latent dirichlet allocation. Journal of machine Learning research. (vol. 3, pp 993-1022).

Broadwell, George Aaron; Umit Boz, Ignacio Cases, Tomek Strzalkowski, Laurie Feldman, Sarah Taylor, Samira Shaikh, Ting Liu, Kit Cho (2013) Using Imageability and Topic Chaining to Locate Metaphors in Linguistic Corpora. SBP-2013 Conference, Washington, DC.

Carbonell, Jaime (1980) Metaphor: a key to extensible semantic analysis. *Proceedings of the 18th Annual Meeting on Association for Computational Linguistics*.

Charteris-Black, Jonathan (2002) Second language figurative proficiency: A comparative study of Malay and English. Appl Linguistics, 23(1):104–33.

Fass, Dan (1991) met*: A Method for Discriminating Metonymy and Metaphor by Computer. *Computational Linguistics, Vol 17:49-90*

Feldman, J. and S. Narayanan (2004) Embodied meaning in a neural theory of language. *Brain and Language*, 89(2):385–392.

Lakoff, George and Johnson, Mark (1980) *Metaphors We Live By*. University of Chicago Press.

Lakoff, George (2001) *Moral politics: what conservatives know that liberals don't.* University of Chicago Press, Chicago, Illinois.

Le, Quoc and Mikolov, Tomas (2014) Distributed representations of sentences and documents. In Proceedings of the 31st International Conference on Machine Learning, ICML'14.

Lin, Ching-Sheng; Samira Shaikh, Jennifer Stromer-Galley, Jennifer Crowley, Tomek Strzalkowski, Veena Ravishankar (2013) Topical Positioning: A New Method for Predicting Opinion Changes in Conversation. Proceedings of the Language Analysis in Social Media workshop, NAACL 2013 Conference, Atlanta, GA.

Liu, Ting; Kit Cho, George Aaron Broadwell, Samira Shaikh, Tomek Strzalkowski, John Lien, Sarah Taylor, Laurie Feldman, Boris Yamrom, Nick Webb, Umit Boz and Ignacio Cases (2014). Automatic Expansion of the MRC Psycholinguistic Database Imageability Ratings. LREC Conference, Reykjavik.

Martey, Rosa Mikeal, Jennifer Stromer-Galley, Mia Consalvo, Kelly Reene, Tomek Strzalkowski, Michelle Weihmann-Purcell, Kevin Shiflett, Jingsi Wu, Jaime Banks, Sharon Small and Michael Ferguson. (2011) "Gamer culture versus the culture of the game: An Analysis of Player Behavior and Gamer Identity in Second Life," in Proceedings of the 12th annual conference of the Association of Internet Researchers (AoIR), Seattle.

Miller, G. A. (1995). WordNet: A Lexical database for English. Comm. of the ACM, 38(11): 39-41.

Miller, D., and Slater, D. (2000) The Internet: An Ethnographic Approach, Oxford; New York: Berg.

Mohler, Michael; David Bracewell, David Hinote, and Marc Tomlinson (2013) Semantic signatures for example-based linguistic metaphor detection. In *Proceedings of the First Workshop on Metaphor in NLP, (NAACL)*, pages 46–54.

Pew Research Center (2013) A survey of LGBT Americans: Attitudes, experiences and values in changing times. Retrieved from https://www.pewsocialtrends.org/2013/06/13/a-survey-oflgbt-americans/

Pew Research Center (2014) Political polarization and media habits: From Fox News to Facebook, how liberals and conservatives keep up with politics. Retrieved from http://www.journalism.org/2014/10/21/political-polarization-media-habits/58

Pew Research Center (2017) America's complex relationship with guns: An in-depth look at the attitudes and experiences of U.S. adults. Retrieved fromhttp://www.pewsocialtrends.org/2017/06/22/americas-complex-relationship-with-guns/

Rosen, Zachary (2018) Computationally Constructed Concepts: A Machine Learning Approach to Metaphor Interpretation Using Usage-Based Construction Grammatical Cues. Proceedings of the ACL Workshop on Figurative Language Processing, New Orleans, Louisiana.

Safar, M., & Mahdi, K. (2012) Social Networking and Community Behavior Modeling: Qualitative and Quantitative Measures, (pp. 1-400). Hershey, PA: IGI Global. doi:10.4018/978-1-61350-444-4

Shaikh, Samira; Tomek Strzalkowski, Ting Liu, George Aaron Broadwell, Boris Yamrom, Sarah Taylor, Laurie Feldman, Kit Cho, Umit Boz, Ignacio Cases, Yuliya Peshkova and Ching-Sheng Lin (2014) A Multi-Cultural Repository of Automatically Discovered Linguistic and Conceptual Metaphors. LREC Conf., Reykjavik.

Shaikh, Samira; Tomek Strzalkowski, Sarah Taylor, Ting Liu, John Lien, George Aaron Broadwell, Laurie Feldman, Boris Yamrom, Kit Cho and Yuliya Peshkova (2015). Understanding Cultural Conflicts using Metaphors and Sociolinguistic Measures of Influence. Proc. of 3rd workshop on Metaphor in NLP, NAACL-2015, Boulder, CO.

Shutova, Ekaterina (2010) Models of metaphor in nlp. *Proc. of 48th Meeting of the Assoc. for Computational Linguistics*, pages 688–697.

Shutova, E. and S. Teufel (2010) Metaphor corpus annotated for source - target domain mappings. *In Proceedings of LREC 2010, Malta*

Strzalkowski, Tomek; George Aaron Broadwell, Sarah Taylor, Laurie Feldman, Boris Yamrom, Samira Shaikh, Ting Liu, Kit Cho, Umit Boz, Ignacio Cases and Kyle Elliott. (2013) Robust Extraction of Metaphors from Novel Data. Proceedings of NAACL Workshop on Metaphors in NLP. Atlanta, GA.

Strzalkowski, Tomek; George Aaron Broadwell, Sarah Taylor, Laurie Feldman, Boris Yamrom, Samira Shaikh, Ting Liu, Kit Cho, Umit Boz, Ignacio Cases and Kyle Elliott (2014) Computing Affect in Metaphors. Proceedings of the 2nd workshop on Metaphor in NLP, ACL-2014 Conference, Baltimore, MD.

Taylor, Sarah; Laurie Beth Feldman, Kit Cho, Samira Shaikh, Ignacio Cases,Yuliya Peshkova, George Aaron Broadwell Ting Liu, Umit Boz, Kyle Elliott. Boris Yamrom, and Tomek Strzalkowski (2014) Extracting Understanding from automated metaphor identification: Contrasting Concepts of Poverty across Cultures and Languages. AHFE Conference, Cracow, Poland.

Thibodeau, Paul, H. and Lera Boroditsky (2011) Metaphors We Think With: The Role of Metaphor in Reasoning. PLoS ONE 6(2): e16782.

Turney, P. D., & Littman, M. L. (2003) "Measuring Praise and Criticism," ACM Trans. Inf. Syst., 21(4), 315–346.

Wilks, Yorick (1975) Preference semantics. *Formal Semantics of Natural Language*, E. L. Keenan, Ed. Cambridge University Press, Cambridge, U.K., 329--348.

Wilks, Yorick; Lucian Galescu, James Allen, Adam Dalton (2013) Automatic Metaphor Detection using Large-Scale Lexical Resources and Conventional Metaphor Extraction. In the *Proceedings of the First Workshop on Metaphor in NLP, (NAACL)*. Atlanta.