An Analysis of WordNet's Coverage of Gender Identity Using Twitter and The National Transgender Discrimination Survey

Amanda Hicks
University of Florida
Gainesville, FL, USA
aehicks@ufl.edu

Michael Rutherford
University of Arkansas for Medical Sciences
Little Rock, AR, USA
mwrutherford@uams.edu

Christiane Fellbaum
Princeton, University
Princeton, NJ, USA
fellbaum@prinecton.edu

Jiang Bian
University of Florida
Gainesville, FL, USA
bianjiang@ufl.edu

Abstract

While gender identities in the Western world are typically regarded as binary, our previous work (Hicks et al., 2015) shows that there is more lexical variety of gender identity and the way people identify their gender. There is also a growing need to lexically represent this variety of gender identities. In our previous work, we developed a set of tools and approaches for analyzing Twitter data as a basis for generating hypotheses on language used to identify gender and discuss gender-related issues across geographic regions and population groups in the U.S.A. In this paper we analyze the coverage and relative frequency of the word forms in our Twitter analysis with respect to the National Transgender Discrimination Survey data set, one of the most comprehensive data sets on transgender, gender non-conforming, and gender variant people in the U.S.A. We then analyze the coverage of WordNet, a widely used lexical database, with respect to these identities and discuss some key considerations and next steps for adding gender identity words and their meanings to WordNet.

1 Introduction

Gender identity is richly lexicalized in American English. Nevertheless, a cursory investigation of gender identity in WordNet (Miller, 1995) suggests that coverage of non-binary gender identity is low. The goal of our research is to measure the coverage of WordNet's gender identity and to suggest steps to improve it.

There is increasing incentive to include gender identity terms and other words that are relevant to transgender, gender variant, non-binary, and gender non-conforming people in WordNet. For example, the Institute of Medicine (IOM) recently recommended (1) gathering data on sexual orientation and gender identity in Electronic Health Records (EHR) as part of the meaningful use objectives in EHRs, (2) developing standardization of sexual orientation and gender identity measures to facilitate synthesizing scientific knowledge about the health of sexual and gender minorities, and (3) supporting research to develop innovative methods of conducting research with small populations to determine the best ways to collect information on LGBT minorities. Furthermore, it is important for the medical community to use words that are common among patients and research participants since the use of language that is familiar to the participant has been shown to improve response rates in data collection (Catania et al., 1996; Institute of Medicine, 2011; Alper et al., 2013).

However, there are challenges to determining which words to include in WordNet and how to define them. Based on the limited research available, some evidence (Dargie et al., 2015; Kuper et al., 2012; Scheim and Bauer, 2015) suggests that vocabulary for self-identifying gender and sexual orientation varies by community. There is clear evidence of lexical variation associated with geography in linguistics studies (Carver, 1987; Chambers, 2001; Nerbonne, 2013). Also, through discussions with members of the trans community and health care providers at LGBT clinics across the country, we have learned that new words are frequently coined to describe gender identity and that the connotations of existing words may vary across communities. We use 'trans' broadly to refer to transgender, transsexual, gender non-conforming, gender variant, and non-binary individuals.

User generated content on social media, such as Twitter, is a valuable resource because it can pro-

vide a source for gleaning information about people's daily lives to answer scientific questions. In our previous work, we produced a data set to investigate words used to discuss gender in the general population and among self-identifying trans persons using Twitter (Hicks et al., 2015). With 'self-identifying' we refer to people who have stated that they have a trans identity either through their tweets or in the National Transgender Discrimination Survey (NTDS) (Grant et al., 2011). We believe that we can augment our Twitter data set with the NTDS data to produce a data set that is in sync with current speakers' language, that can serve as a starting point for enriching WordNet's coverage of gender identity, and that can contribute to the medical and clinical goals outlined at the beginning of this section.

The National Transgender Discrimination Survey (NTDS) is the largest survey of the trans population in the United States to date (Harrison et al., 2012). The survey was designed to collect information about "the broadest possible swath of experiences of transgender and gender nonconforming people" in the U.S.A., including questions about how participants identify their own gender and an option to write in one's own identity (Harrison et al., 2012). We have compiled a list of the gender-identity word forms (henceforth simply 'words') from this survey and performed a normalized frequency analysis that can be compared to our Twitter data set.

In our previous work we built a data set and visualization tools that show relative frequency and co-occurrence networks for American English trans words on Twitter (Hicks et al., 2015). Our goal in this paper is to perform a two-fold coverage analysis of WordNet with respect to American English gender identity.

Our hypothesis is that a comprehensive list of words used to self-identify gender will require examining the words trans people use in different contexts. In order to evaluate this hypothesis, we perform a frequency analysis of words from both sets.

Our approach is as follows. First, we compare the trans identity words that we identified in our previous work with the words from the NTDS to assess the coverage of the Twitter set. Next, we produce an updated set of words using the NTDS and compare WordNet's coverage of gender identity against this list.

2 Methods

Here we describe our language analysis of the Twitter data and the NTDS data.

2.1 Language Analysis of Twitter Data

The general idea underlying our approach is to identify tweets that are relevant to the discussion of trans related issues and then examine the variations in language used for gender identification by different communities, that is, by population (trans people vs. the general public) and by geographical location (U.S. states). The analysis workflow consists of five main steps, as depicted in Figure 1: 1) collect tweets that are potentially related to discussions about gender identification; 2) preprocess and geotag tweets with their corresponding U.S. state; 3) build supervised classification models based on textual features in the tweets to a) filter out irrelevant tweets and b) find people who are self-identified as trans; 4) collect relevant (both self-identifying trans users and users in the general public who discussed trans related issues) users' Twitter timelines which consists of all of their tweets in chronological order; and 5) compare the usage of gender identification words by geographical locations (i.e., by U.S. states) and by population groups (self-identifying trans people vs. the general public).

Some of the search terms are ambiguous and their meanings are context dependent. For example, the tweet 'That Hot Pocket is full of trans fats' is not related to discussions of gender identification even though it contains the keyword 'trans'. To account for this observation, we engineered a binary classifier to determine the likelihood that a tweet is relevant to the discussion of gender identification and to remove those that are unlikely to be relevant from the corpus in step 3. We also leverage a number of visualization techniques to provide straightforward and easy-to-understand visual representations, namely, word clouds, co-occurrence matrices, and network graphs to substantiate our findings. A full description of this work and analysis of terms can be found in (Hicks et al., 2015).

2.2 Language Analysis of NTDS Data

Unlike the Twitter study data processing techniques, the NTDS dataset did not require the preprocessing for language filtering, geotagging or the mining techniques for the identification of rel-

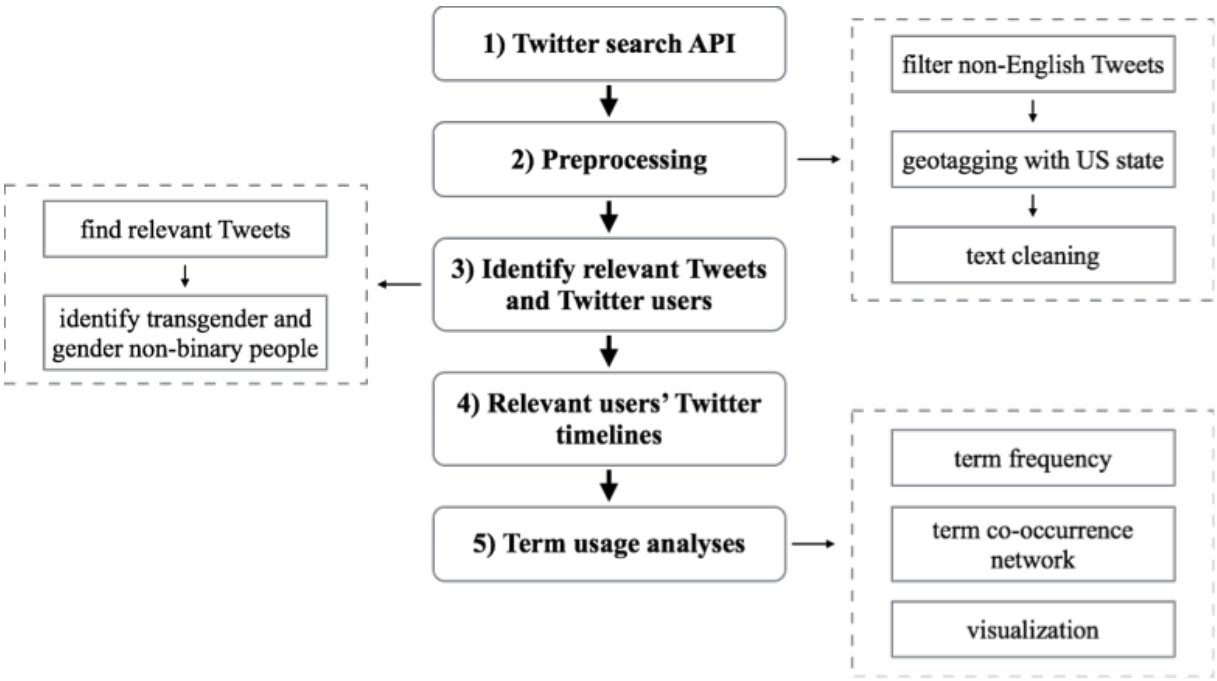

Figure 1: The analysis workflow for identifying tweets related to trans issues

evant trans individuals. Knowing that the records were all of unique self-identified trans individuals, we were able to skip ahead to Step 5, the term usage analysis.

The Twitter data analysis methods were duplicated and restricted to the term extraction and usage analysis, including term frequencies and word cloud generation.

We utilized questions three and four from the NTDS. These questions asked what gender identity the respondent identified with at the time of the survey and how strongly they identified with certain identities. Figure 2 shows these questions.

Term frequency analyses were generated based on all words utilized, no matter the degree with which the respondent specified (strongly, somewhat, or not at all). The frequencies were then measured both at a state and national level for coverage comparisons with the Twitter set.

2.3 Coverage Analysis of Twitter Words

We performed a coverage analysis of the words in the Twitter data set with those from the NTDS data set. We collated all of the words in the NTDS questions three and four as well as the identity words used in the write-in responses. We removed terms that were preceded by a hash tag in the Twitter set and words that were only used once in the NTDS set, and then we measured the number of common words from both the Twitter list and the NTDS list. Due to the character limit on Twitter, abbreviations are common in Tweets as are alternate spellings of words (e.g., 'gender queer' and 'gender-queer'). We also gathered words into groups consisting of alternative spellings and abbreviations. 'Genderqueer' and 'gender-queer' are in the same group. Henceforth we call these groups of word forms simply 'groups'. We measured the degree of overlap of groups in Twitter and in NTDS which is reported in the results section of this paper.

2.4 Coverage Analysis of WordNet

Our next step was to generate a list of words to use in the coverage analysis of WordNet. We removed the Twitter terms that contained a hash tag from the Twitter data set and removed word forms that only had one occurrence in the NTDS set. We then took the union of these sets to produce a set of words for evaluating the coverage of WordNet. Similarly, we produced a list of groups with alternate spellings and abbreviations by taking the union set of corresponding groups for the Twitter list and NTDS list. For example, the NTDS word groups contained the group (gender non-conforming, gender non conforming) and the Twitter word groups contained (gender non-conforming, gnc). The compiled set of groups contains (gender non-conforming, gender non conforming, gnc).

3. What is your primary gender identity today?

- ❍ Male/Man
- ❍ Female/Woman
- ❍ Part time as one gender, part time as another
- ❍ A gender not listed here, please specify _____

4. For each term listed, please select to what degree it applies to you.

	Not at all	Somewhat	Strongly
Transgender	❍	❍	❍
Transsexual	❍	❍	❍
FTM (female to male)	❍	❍	❍
MTF (male to female)	❍	❍	❍
Intersex	❍	❍	❍
Gender non-conforming or gender variant	❍	❍	❍
Genderqueer	❍	❍	❍
Androgynous	❍	❍	❍
Feminine male	❍	❍	❍
Masculine female or butch	❍	❍	❍
A.G. or Aggressive	❍	❍	❍
Third gender	❍	❍	❍
Cross dresser	❍	❍	❍
Drag performer (King/Queen)	❍	❍	❍
Two-spirit	❍	❍	❍
Other, please specify _____	❍	❍	❍

Figure 2: Questions 3 and 4 from the National Transgender Discrimination Survey that asks respondents to report their gender identity

We automatically searched for words and groups of synonymous words ('synsets') that corresponded to words and groups using the Natural Language Tool Kit's (NLTK) interface for WordNet 3.0 (Bird et al., 2009). We then manually evaluated which synsets were relevant to gender identity. We did not evaluate whether the WordNet definition accurately characterized the intended meaning of the word, in part because we do not have a reliable method for ascertaining the intended meaning of the word and also because that is outside of the scope of our coverage analysis.

Many of the groups that did not have a corresponding synset in WordNet 3.0 were compounds such as 'trans person of color'. Our next step was to produce a list of words in compounds and search for corresponding synsets in WordNet. We manually identified compounds and then generated a set of words in the compounds. We removed stop words from the set with NLTK. Once again we programmatically searched for synsets using NLTK and then manually evaluated whether the retrieved synset was relevant to gender identity. We classified the compounds into three groups: (1) those that were partially covered by WordNet, meaning they contained at least one word that corresponded to a relevant synset and at least one that did not, (2) those that were completely covered by WordNet, meaning every word in the compound (excluding stop words) was represented in WordNet, and (3) those that had no coverage in WordNet.

3 Results

First we discuss the results of analysis of our Twitter data. Then we discuss our analysis of WordNet's coverage of trans related terms.

3.1 Language Analysis of Twitter Data

We collected over 53.8 million tweets matching the search queries during a 116-day period from January 17, 2015 to May 12, 2015 inclusive. Out of the collected tweets, about 29 million tweets (54.2%) were in English. We were able to extract location information for 368,518 tweets (1.26% of English tweets from 119,778 unique users), which we retained for further processing. We eliminated the tweets that were deemed irrelevant (15,478 tweets from 3,785 users) based on a classification model we developed (Hicks et al., 2015). From the remaining records, 115,993 Twitter users were classified as relevant, of which 1,921 users were classified as self-identifying trans. In addition to the data we collected using the search API, we

	Unique	Shared
NTDS Words	79.66% (141 / 177)	20.34% (36 / 177)
NTDS Groups	81.82% (117 / 143)	18.18% (26 / 143)
Twitter Words	80.65% (150 / 186)	19.35% (36 / 186)
Twitter Groups	67.50% (54 / 80)	32.50% (26 / 80)

Table 1: The percentage of overlap among NTDS and Twitter words and groups

crawled more than 337.9 million tweets from the 115,993 relevant Twitter users' timelines. Out of the 337.9 million tweets, 872,340 Twitter messages contain one or more of the keyword forms of our interest. These 872k tweets comprise the corpus we used for language usage analysis.

3.2 Coverage of Twitter Word Groups

Table 1 contains a summary of the degree of overlap between the set of Twitter trans words and their groups and the NTDS trans words and their groups. Only about 18% of the NTDS groups were represented in the Twitter data set. Section 4.2 contains a discussion of some of the main reasons for the most frequent word forms not being in the Twitter data set.

The word clouds in Figure 3 illustrate two interesting facts about word usage to self-describe trans identity.

First, different words appear in different contexts. For example, 'cis' and 'shemale' are prevalent on Twitter but not in the NTDS. Second, even words that are common across contexts are used with different frequency. For example, 'genderqueer' is prominent in the NTDS word cloud but relatively small in the Twitter word cloud (top left-hand quadrant). Conversely, 'Transgender' is more prominent in the Twitter word cloud than the NTDS.

3.3 WordNet's Coverage of Gender Identities

We found that 39% of the words in our compiled list of trans groups have a corresponding synset in WordNet 3.0. Another 28% of the words were compounds that contain at least one component word with a corresponding synset in WordNet and one without. 33% of the words did not have any corresponding entries in WordNet. These results are summarized in Figure 4. Table 2 shows a numerical analysis of WordNet's 3.0 coverage of our trans related words.

4 Discussion

4.1 Limitations

We note that our previous study is limited by the user demographics available on social media platforms. The users of social media tend to be younger; 37% of Twitter users are under 30, while only 10% are 65 or older, as of 2014 (Duggan, Ellison, Lampe, Lenhart, & Madden, 2014). There are also power users who exhibit a substantially greater level of activity than the average user (Pew Research Center, 2015). These characteristics are likely to create sample bias and impose limitations on mining meaningful information from Twitter that represents a broader population. For instance, Twitter data may not be reliable for mining information about older people who may not use Twitter.

The NTDS was published in 2011, but more current data are being collected at the time of writing this paper. The Transgender Survey 2015 was launched in August 2015 (U.S, 2015) and the PRIDE study in June 2015 (PRI, 2015). We expect these newer data sources to be completed within the next year or two. Both studies collect demographic data on trans individuals, including identity words. This will provide insight into which words are relatively stable over time and may also reveal words that are emerging as more prevalent.

4.2 Words Excluded From Twitter Search Terms

While compiling a list of words for Twitter, we observed the distinctions among trans identities, intersex conditions, and sexual orientation. As a result we excluded words that were specifically intersex related or that describe sexual orientation from the Twitter set. However, intersex and sexual orientation words were among participant responses in the NTDS so were included in our NTDS data set. The heterogeneous nature of the Twitter term lists and NTDS term lists may skew the coverage analysis of our Twitter list. However, this heterogeneity is valuable for our analysis of WordNet's coverage since it provides a more comprehensive list of words that trans people use to describe their own identities.

Figure 3: Word clouds representing the relative frequency of trans words used by self-identifying trans people on Twitter in the U.S.A. (left) and self-identifying trans people in questions three and four of the NTDS (right)

	Word Groups in WordNet
Full WordNet Coverage	**71**
Non-Compounds In WordNet	39
Compound - Full Coverage	32
Compound with Partial WordNet Coverage	**50**
Non-compounds not in WordNet	61
Compound - No WordNet Coverage	13
No WordNet coverage	**74**
Total Trans* Word groups	**195**

Table 2: Analysis of trans word groups in WordNet 3.0 reported by number

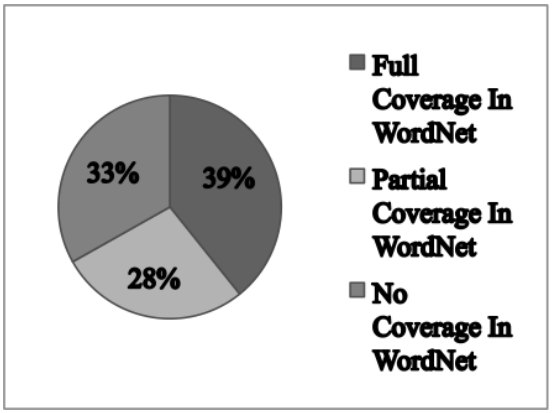

Figure 4: Summary of WordNet 3.0's coverage of trans word groups

An examination of tables 3 and 4 reveals three main reasons words from the NTDS term lists were not included in the Twitter term lists: (1) Polysemy - 'Aggressive' is polysemous and would result in too many false hits in the Twitter search. Similarly 'androgynous' produced too many false hits since many people who used this word were tweeting about fashion. (2) Gender words that are not trans specific -'male', 'female', 'woman', and 'man', are used with such prevalence that we excluded them in the Twitter set since they are unhelpful in identifying tweets about trans issues. (3) Identity words that are not trans specific - 'butch' and 'intersex' were deliberately excluded from the Twitter set since we were following the conceptual distinctions among sexual orientation, gender identity, and intersex. However, the NTDS data set shows that when individuals describe their gender identities, they do not limit their descriptions to these high level distinctions.

4.3 Suggestions for Integrating Gender Identity Into the WordNet Database

Approximately one third of the compounds with partial or no coverage have 'gender' as a compo-

NTDS Term	NTDS %	Included in Twitter Set
Aggressive	10.4%	No
Genderqueer	5.5%	Yes
Transgender	5.3%	Yes
Butch	5.2%	No
Female-to-male	5.2%	Yes
Androgynous	5.2%	No
Male-to-female	5.2%	Yes
Transsexual	5.2%	Yes
Two-Spirit	5.2%	Yes
Intersex	5.2%	No

Table 3: Ten most frequent words in NTDS

NTDS Term	NTDS %	Included in Twitter Set
Genderqueer	16%	Yes
Male	8.7%	No
Female	8.2%	No
Woman	4.9%	No
Queer	4.7%	No
Transgender	3.5%	Yes
Trans	2.7%	Yes
Man	2.7%	No
Butch	1.8%	No
Female-to-male	1.7%	Yes

Table 4: The ten most frequent words in the NTDS write-in fields in questions three and four

nent term. The synsets for 'gender' in WordNet are tied to biological properties and reproductive roles, and there is no synset for gender as a social role independently of reproductive features. Other words that would have a significant effect on WordNet's coverage of compounds are 'trans', 'genderqueer', and 'femme'. Some words that are relevant to the trans issues such as 'agender', 'cisgender' (describing somebody who is not trans), and 'binarism' are missing.

In addition to adding more words to integrate gender identity in WordNet, efforts should be made to craft informed definitions and example sentences of new words and to evaluate the accuracy of existing entries. Likewise, more work needs to be done to identify synsets. The word groups that we used for this study grouped morphologically similar words such as 'gender queer' and 'gender-queer'. However, we did not group words like 'agender' and 'genderless' into synsets. Methods for reliably detecting synonyms of gender identity words should be developed and tested.

Finally, methods also need to be developed for establishing hierarchy relations among gender identity words. Such methods may include testing established lexical patterns with English speakers who are competent with trans vocabulary (Hearst, 1992). Another approach may include leveraging the responses in question 4 of the NTDS to detect hierarchy relations. For example, if most participants who identify strongly as transgender also identify strongly as genderqueer but not vice versa, this could indicate that 'genderqueer' is a hypernym of 'transgender'.

4.4 Future Work

Wordnets have been built in some seventy different languages, and each reflects the culture of the speakers. Mapping gender identity words across languages should reveal interesting similarities and differences. For example, India allows its citizens to officially identify as 'third gender', or *hijra*, a term that encompasses biological males dressing in women's clothes as well as intersex individuals. Future research within the global wordnet community could ask whether such officially sanctioned words cover distinct words used in specific communities and if so, how do they correspond to the English words identified in our work? Twitter corpora can show which terms are used in similar or identical contexts (n-grams), suggesting synonymy and shared synset membership. Additionally, questionnaires could be developed and submitted to the trans population for input on how to accurately represent the terms. Reflecting geographic and group differences poses additional challenges, akin to dialectal variation that is currently marked in WordNet with usage flags.

5 Conclusion

Our hypothesis was that a comprehensive list of words used to describe gender identity will require sets of words taken from different contexts. To test this hypothesis we performed a coverage analysis of trans words taken from two different contexts, Twitter and the National Transgender Discrimination Survey. We found that while there was some overlap, there was significant variation of words used between these contexts. As a result, we generated a more comprehensive list of trans words from both sources. A second aim of this paper was to assess WordNet's coverage of trans identity. We found that, while there is some coverage of trans words in WordNet, there is more work to be done to ensure more comprehensive coverage.

Acknowledgements

We are grateful to the National Center for Transgender Equality (NCTE) for providing the dataset from the National Transgender Discrimination Survey. We are also grateful to Naomi Ardjoman-Kermani for their helpful comments on previous drafts. This work was supported in part by the NIH/NCATS Clinical and Translational Science Awards to the University of Florida UL1 TR000064 and the University of Arkansas for Medical Sciences UL1 TR000039. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health or the NCTE.

References

Joe Alper, Monica N Feit, Jon Q Sanders, et al. 2013. *Collecting Sexual Orientation and Gender Identity Data in Electronic Health Records: Workshop Summary*. National Academies Press.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python*. O'Reilly Media, Inc.

Craig M Carver. 1987. *American Regional Dialects: A Word Geography*. University of Michigan Press.

Joseph A Catania, Diane Binson, Jesse Canchola, Lance M Pollack, Walter Hauck, and Thomas J Coates. 1996. Effects of interviewer gender, interviewer choice, and item wording on responses to questions concerning sexual behavior. *Public Opinion Quarterly*, 60(3):345–375.

Jack K Chambers. 2001. Region and language variation. *English world-wide*, 21(2):169–199.

Emma Dargie, Karen L Blair, Caroline F Pukall, and Shannon M Coyle. 2015. Somewhere under the rainbow: Exploring the identities and experiences of trans persons. *The Canadian Journal of Human Sexuality*.

Jaime M Grant, Lisa Mottet, Justin Edward Tanis, Jack Harrison, Jody Herman, and Mara Keisling. 2011. *Injustice at Every Turn: A Report of the National Transgender Discrimination Survey*. National Center for Transgender Equality.

Jack Harrison, Jaime Grant, and Jody L Herman. 2012. A gender not listed here: Genderqueers, gender rebels, and otherwise in the National Transgender Discrimination Survey. *LGBTQ Public Policy Journal at the Harvard Kennedy School*, 2(1).

Marti A Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th conference on Computational linguistics-Volume 2*, pages 539–545. Association for Computational Linguistics.

Amanda Hicks, R. Hogan, William, Michael Rutherford, Bradley Malin, Mengjun Xie, Christiane Fellbaum, Zhijun Yin, Daniel Fabbri, Josh Hanna, and Jiang Bian. 2015. Mining Twitter as a first step toward assessing the adequacy of gender identification terms on intake forms. In *Proceedings of the AMIA 2015 Annual Symposium*. American Medical Informatics Association.

Institute of Medicine. 2011. The health of lesbian, gay, bisexual, and transgender people: Building a foundation for better understanding.

Laura E Kuper, Robin Nussbaum, and Brian Mustanski. 2012. Exploring the diversity of gender and sexual orientation identities in an online sample of transgender individuals. *Journal of Sex Research*, 49(2-3):244–254.

John Nerbonne. 2013. How much does geography influence language variation? *Space in Language and Linguistics: Geographical, Interactional, and Cognitive Perspectives*, pages 220–36.

2015. The Population Research in Identity and Disparities for Equality (PRIDE) Study. `http://www.pridestudy.org`. Accessed: 2015-08-24.

Ayden I Scheim and Greta R Bauer. 2015. Sex and gender diversity among transgender persons in Ontario, Canada: Results from a respondent-driven sampling survey. *The Journal of Sex Research*, 52(1):1–14.

2015. U.S. Trans Survey 2015. `http://www.transsurvey.org`.