Identifying User Demographic Traits through Virtual-World Language Use

Aaron Lawson
Speech Technology and Research (STAR) Lab
SRI International
Menlo Park, California
aaron.lawson@sri.com

John Murray
Computer Science Laboratory
SRI International
Menlo Park, California
john.murray@sri.com

Abstract—The paper presents approaches to identifying real-world demographic attributes based on language use in the virtual world. We apply features developed from the classic literature on sociolinguistics and sound symbolism to data collected from virtual-world chat and avatar naming to determine participants' age and gender. We also examine participants' use of avatar names across virtual worlds and how these names are employed to project a consistent identity across environments, which we call "traveling characteristics."

Keywords—virtual worlds, linguistic features, machine learning

1 Introduction

This paper presents results from VERUS, a large-scale study of the relationship between virtual-world/online behavior and real-world characteristics. The program goal was automatically predicting major real-world (RW) demographic attributes using only virtual-world (VW) behavior. These RW attributes include age group, gender, ethnicity, income level, education level, leadership role, and urban/rural background among others. More than one thousand participants volunteered RW demographic information about themselves and allowed the recording of their online behaviors, including text chat data produced during online activities and names chosen for online personae ("avatars"). SRI gathered hypotheses from the theoretical sociolinguistics literature, phonology and sound symbolism, semantics, and discourse analysis and made empirical observations to generate features for determining RW attributes that could be combined in a global model using statistical classifiers.

Some of the major observable features had application to several sub-areas, such as factors that related to neophyte online behavior (using real-world names and written language syntax). In Table 1, we give examples of those features that correlated highly with the prediction of two conditions: age and gender.

Various windows into RW identity were provided by chat behavior, including word-choice factors such as swearing, insults, apologies, and expressions of certainty; and grammatical factors including verb forms and use of questions. Length of utterance, typographical features, spelling, part of speech, and semantic affect were also found to contribute to identity determination and to apply outside chat data [1]. Avatar-naming conventions in terms of typography (capitalization, special characters, spaces, etc.); phone class of characters used; syllable structure; name endings; use of numbers; and other factors enabled good prediction of both age group and gender [2]. Further, even when online avatar gender differed from RW gender, naming conventions still enabled the correct prediction of RW gender. Participants who used multiple online names exhibited behavior that enabled linking of different names across sessions [3].

TABLE 1. PROMINENT CHAT FEATURES.

Behavior	Class	Example
Hedging/Uncertainty	Female	Questions, "I don't know"
Command forms	Male	"Heal me!"
Use of slurs	Male	"You homo!"
Direct apologies	Female	"I'm sorry"
Indirect apologies	Male	"Ooops," "my bad"
Empathy	Female	"I like X," "u OK?"
Use of modal verbs	Female	can, could, would, should, etc.
Use of all caps	Youth	"STOP BEING DUMB"
Frequent use of ellipsis	Adult	"if you bring up your questlog..."
Using commas, apostrophes	Adult	"we're done. let's turn in"
Lowercase "i" for "I" and "u" for "you"	Youth	"u losted to 4 pokemno"
Single word utterances	Youth	"come," "yo"

The following sections delve more deeply into using language-based features to enable the detection of demographic categories. We focus on age and gender, but also discuss how avatar names relate to demographics and how these names are used across virtual-world environments.

2 Data

We obtained data for this study from the VERUS internal corpus of virtual-world chat and avatar demographics information from the Sherwood and Guardian Academy worlds datasets only. World of Warcraft (WoW) and Second Life data was used for hypothesis generation. The avatar names used in this study were chosen by participants when setting up their character information only for games on the VERUS server environment. Note that no actual avatar names are presented in this paper due to privacy concerns on the part of participants. Table 2 summarizes the total amount of data collected for the project.

TABLE 2. DATA DISTRIBUTION OF VIRTUAL-WORLD CHAT IN THE VERUS PROJECT.

Game	Turns	Talkers	Tokens
Guardian Academy	914	57	2,688
Sherwood	13,149	271	57,843
SecondLife	79	4	392
WoW	2,337	117	56,036
Total	11,214	445	89,521

3. GENDER

This section explores gender differences in virtual-world language use. We test the relevance of traditional sociolinguistic observations of males and females in face-to-face conversation to the contemporary space of VW chat interactions in online gaming and collaborative environments. In addition, we study the relationship between a player's real world (RW) gender and naming decisions for online personas, or avatars, in the light of linguistic observations based on sound symbolism and naming conventions. The approach taken in this study focused on applying sociolinguistic claims and observations to develop discourse features for characterizing gender in VW chat and looking at other linguistic factors, such as choice of avatar name, to detect gender trends. To expedite the development of features, we examine the rich empirical claims of the sociolinguistic literature to identify known factors that have tended to correlate with male or female speech, especially [4], [5], [6], [7], [8]. A primary goal of this study is determining whether these findings apply in the physically distant universe of VW interactions.

3.1 Sociolinguistic Features

Based on the literature, ten features were evaluated: 1) expressions of uncertainty; 2) strong swears; 3) light swears; 4) insults; 5) slurs; 6) questions; 7) modals verbs; 8) expressions of empathy; 9) strong apologies; and 10) indirect apologies. Rules for extracting each of these features were developed and the results were calculated to verify whether gender-based biases existed in the distribution.

Analysis of the results (Figure 1) is given as *lift*, the degree to which the feature rises above random. Results shows that many RW sociolinguistic claims about gender and discourse also hold true in the VW: women were much more likely to use modal verbs, ask questions, use expressions of uncertainty, and use strong apologies than were males. Males were much more likely to use strong swears, slurs, and indirect apologies. The results also demonstrate that some of the categories suggested in the sociolinguistic studies may be too coarse. For example, women are claimed to apologize more frequently than men. However, analyzing the kinds of apologies that occur makes clear that direct apologies (e.g. "I'm sorry") are more typical of female players, while indirect apologies ("oooops!" or "my bad!") are more associated with males. Similarly, with swears, a breakdown between the kinds of words used existed: light swears were associated more with females and strong swears, more with males. This observation is actually in keeping with the observations that men are more comfortable with profanity than females, as many of the "light swears" represent approaches to avoid offensive cursing. Slurs were the category most strongly associated with males, and most slurs were homophobic in nature.

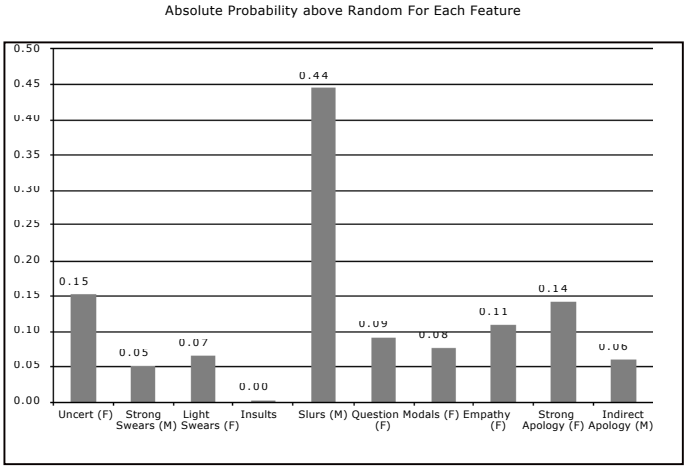

Figure 1: Prominence of each feature for gender.

3.2 Gender Differences for Emoticon and Ellipsis Use in Chat

Other groups have reported significant gender effects for the use of both emoticons and the ellipsis in online chat. The major hypotheses claim: 1) women tend to use ellipsis more often than men;, 2) Women tend to use emoticons more often than men;, but 3) men tend to use lewd (:P, :D, xD) emoticons more often. To determine whether these claims hold with the data collected by the VERUS team, 9,373 lines of in-game chat were process for these phenomena. Of this data, 5,149 turns were produced by males and 4,226 produced by women. This data came from the Guardian Academy and Sherwood.

The full ellipsis (...) frequency was calculated for both genders, with 152 occurrences for males and 116 for females. After adjusting for differences in the priors for both classes, the probability of the ellipsis being associated with females is 48.2%. The partial ellipsis (..) was also evaluated, with 208 occurrences for males and 163 for females. The probability of partial ellipsis being used by women was 48.8%. The distribution of all emoticons across genders was calculated, with 124 instances for males and 164 for females. After adjusting for differences in the priors for both classes, the probability of emoticons being associated with females is 61.2%. For non-lewd emoticons, males had 36 instances and females, 64 instances. The probability of females using non-lewd emoticons was 68.4%. The frequency of lewd emoticons was calculated for both genders: males had 57 instances and females, 50 instances. After adjusting for the differences in the priors, the probability of a male using lewd emoticons was 53.3%.

The conclusions were mixed. No gender effect was seen for either type of ellipsis in this VW data, with lift being less than 3%. (We consider significant any lift of greater than 5%.) Emoticons in general and non-lewd emoticons in particular were significant for females, with lifts of 34.0% and 36.8%, respectively. Lewd emoticons were not significant for males, with lift of 3.3%. A clear trend exists that female

The authors acknowledge the Air Force Research Laboratory at Wright Patterson Air Force Base for sponsorship of this research under contract FA8650-10-C-7009.

participants tended to use emoticons more often than did males. This use may be a function of hedging or qualifying a turn with an emoticon as a way of softening the potential impact of a statement or ensuring that a turn made in jest was not interpreted as a insult or slight.

3.3 Avatar Names and Gender

For investigating the relationship between gender and avatar naming, thirteen rules, largely based on observations from the sound symbolism research (especially [9], [10], and 11), were developed: four for females and nine for males. These included phonetic rules such as female names ending in low vowels; male names ending in back vowels; male names containing "z" or "x"; and female names containing "sh." The association between female names and final low vowels comes from the frequency of female grammatical endings in both Semitic and Indo-European languages. In addition, we included more basic rules, such as the use of female names for female players and male names for male players, based on 2010 U.S. census data. These rules are listed in Table 3.

TABLE 3. AVATAR RULES.

Rule #	Gender Affected	Formulation
1	FEMALE	ends in "a"
2	MALE	ends in back vowel
3	FEMALE	ends in "y"
4	MALE	ends in "er"
5	MALE	ends in back or alveolar stop
6	MALE	ends in any consonant
7	MALE	ends with fricative consonant
8	MALE	begins with capital letter
9	MALE	contains "x" or "z"
10	FEMALE	contains palatal fricative ("sh")
11	MALE	contains a title of nobility
12	MALE	is a male census name
13	FEMALE	is a female census name

The highest precision sound-based rules deal with word endings, with those words ending in a fricative consonant being strongly male, and words ending in the central vowel schwa (represented orthographically with "a") being strongly female. Applying the same rules to avatar names from individuals whose RW and VW gender were different enabled the detection of RW gender at a similar high rate of accuracy (>.7), despite the mismatched gender. This result was surprising and may show that avatar gender was not playing a significant role in the players' online personas.

4 AGE

In this section, we investigate the relationship between the choice of avatar names, chat, and participant age. A total of 305 participants (84 group 1, 178 group 2, 43 group 3) from collects spanning Guardian Academy and Sherwood were used in this section's analysis. Data was collected by matching avatar choice and chat within gaming sessions to demographic information collected in the project. The initial rules were developed based on an examination of the patterns in the avatar names. The notion was that younger avatars would be more innovative and more likely to break the conventions of standard naming, while older users would create names that conformed more to traditional naming conventions.

TABLE 4. PROPOSED AGE RULES.

Rule #	Age Characterized	Formulation
1	1	contains a number
2	2, 3	is a census name
3	2	contains space or separator in name

Three rules were proposed, which are defined in Table 4. The first rule looks at whether a number is used anywhere in the name. Rule two checks whether an avatar name is also a name listed in the US Census report of the 1,000 most common names for 2010. Rule three checks whether a name is divided into a "first name" and "last name" form, which one would associate with canonical names.

4.1 Age Results

These rules were tested against the 305 avatar names with age information from Guardian Academy and Sherwood. The results in Table 5 show that two rules achieved program goals (2 and 3), reaching precision levels of greater than 0.85. However, if one normalizes for distribution prior probabilities, rule 1 is actually superior to both rules 2 and 3, as far fewer examples of class 1 exist than do of class 2.

TABLE 5. AGE RESULTS.

Rule #	AvePrec.	Prec.	Recall	F-Score
1	0.82	0.76	0.26	0.39
2	0.76	0.89	0.32	0.47
3	0.62	0.87	0.07	0.13

4.2 Rule Combination of Names and Chat for Age

The rule combination results are shown in Table 6. These results take the best performing features from both chat and naming and combine the results by using machine-learning techniques. Very young participants tend to use a "telegraphic" chat style with low use of articles and other grammatical particles; low use of overt pronouns (when the meaning is understandable); and frequent use of "shouting" (all CAPS) in both names and chat. Adults have a chat style that conforms much more closely to standard written language with use of overt pronouns, qualifiers,

and articles; and use of the flourishes of written language, such as the ellipsis, etc.

TABLE 6. RULE COMBINATION.

Rule	Class	P	r
Avatar name is all caps + low use of articles	Youth	92%	12%
Frequent use of articles but low use of pronouns	Young Adult	80%	20%
Frequent use of both articles and qualifiers/tentative language	Adults	83%	14%

5 TRAVELING CHARACTERISTICS

Avatars are the primary means by which players navigate with and interact in most Massively Multiplayer Online Games (MMOGS). In recent years, both quantitative and qualitative studies have drawn attention to the opportunities for, and practices around, MMOG avatar creation and customization. Quantitative studies have explored player practices around avatar "gender-swapping," class choice, and comparisons of avatar customization options across different games. Qualitative (typically ethnographic) accounts of avatar customization have addressed the relationships and affiliations generated between players and their avatars (usually focusing on one avatar and in one MMOG). Celia Pearce's ethnographic look [12] at how a disenfranchised community of There.com users recreated the world within Second Life represents one of the few studies that examines players' transition from one virtual world to another. In doing so, Pearce opens up for consideration the ways that players create a sense of stability across experiences that are otherwise contingent upon specific games. Our hypothesis is that naming practices represent one primary (and underexplored) means through which players maintain continuity and a stable identity across their MMOG play. To explore this, we draw on more than 1700 avatar names, from more than 400 players, obtained through a mixed-methods, multi-site study of players across multiple MMOGs. Here, we report on the avatar-naming practices that we identified through both quantitative and qualitative analyses of this dataset.

The quantitative analysis consisted of unigram and bigram similarity measures, and Levenshtein or "Minimum Edit" distance metrics. The goal was to determine the extent to which players share either parts of their avatars across virtual worlds, measured by bigram similarity and Levenshtein distance, or characters from their avatar names, measured by unigram similarity. The analysis compared all of a participant's own avatar names with all other players' avatar names to determine the cohesion of a players' avatar names across worlds. The results showed that considerable "traveling" of names occurs across virtual worlds, with same-player bigrams per name averaging 3.4 for males and 2.9 for females, compared with 0.53 and 0.66 bigrams per name for different player names (see Figure 2). The Levenshtein distance was 17% greater for different player names than for same player names, and unigram similarity was 25% greater for same-player than different. These results demonstrate that players maintain a great deal of measurable continuity across virtual worlds using avatar naming.

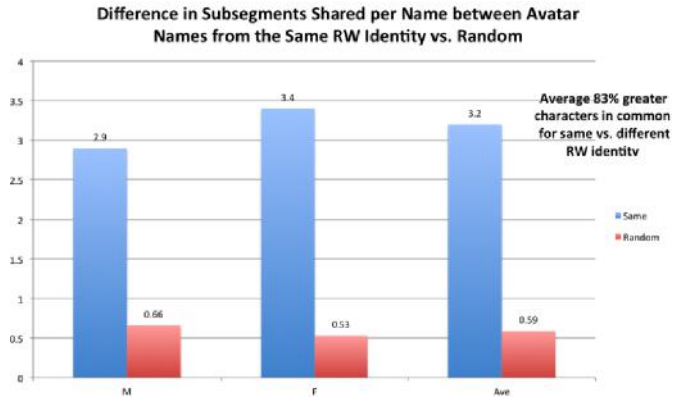

Figure 2: Reuse of Name Components across VWs.

Supplementing and extending this quantitative analysis, we also conducted a qualitative exploration and "open coding" of avatar names, identifying significant and widespread instances of naming practices that fall outside those quantitative measures. Specifically, we saw numerous examples of players' intertextual and thematic practices in "branding" multiple avatars. This discussion contributes both conceptual and methodological innovations to the field of MMOG studies. Conceptually, we can identify one significant "traveling characteristic," the means that players use to create a stable identity as they move across avatars and virtual worlds. Methodologically, we can show how quantitative and qualitative analyses of the same dataset can, when developed in tandem, provide a robust picture of an underexplored facet of player practice in MMOGs.

6 FUTURE RESEARCH

In this paper, we show that linguistic features offer high predictive value for identifying user demographics in the virtual world. We summarize our findings in Figure 3, with examples of representative chat and naming phenomena and of how they relate to demographics.

Despite the enormous amount of communication occurring online every day, research on virtual-world and online language use is still in its infancy. Three areas offer significant potential to expand our capabilities to identify real-world characteristics from VW language usage:

- Semantic-level analysis (word meanings in context) and syntax
- Association of participants through real-world cultural references
- Discourse and participant interaction

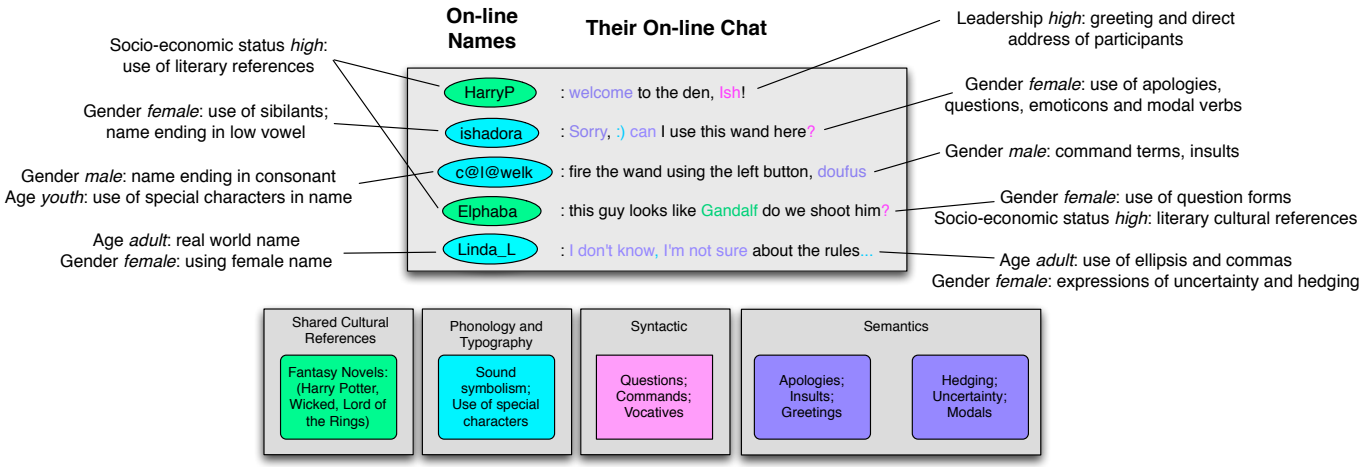

Figure 3: Summary of Chat and Naming Findings

Semantic analysis is clearly a gap in current approaches, and a pilot study into the kinds of shared real-world cultural references in both naming and chat reveals that this sort of information often provides a more focused and fine-grained picture of group association than lexical or typographical features do. In Figure 3 we see how linguistic features such as typography, syntax, and phonology enable the identifying real-world traits such as gender and age. Semantic and cultural references are crucial to understanding more specific multivalent characteristics such as socio-economic status and leadership roles. Semantic analysis also facilitates the effective clustering of individuals based on shared areas of interest and topic (for example, fantasy novels) to enable the linking of individuals who fit a certain profile (individuals who are interested in identified area X are also interested in Y).

Discourse factors—including interpersonal interaction, emotional response to others, non-native language use, intentionality assessment and other areas—enable the prediction of real-world demographics by combining low-level features with the larger semantic context and permit identity attribution via shared stylistic factors. For example, the discourse acts of greeting (especially being the first greeter) and addressing someone directly by name are associated with high real-world leadership scores. Though these "higher-level" features may prove difficult to extract automatically, they offer the potential for new insights into users' online and virtual-world behavior, with many possible applications to related settings, such as artificial-reality environments and games.

REFERENCES

[1] Wang, W. et al. "Automatic Detection of Speaker Attributes Based on Utterance Text," *Interspeech 2011*, Florence, Italy, October 2011.

[2] Lawson, A. et al. "Sociolinguistic Factors and Gender Mapping across Real and Virtual World Cultures," *2nd International Conference on Cross-Cultural Decision Making*, San Francisco, CA, July 2012.

[3] Lawson, A. and Taylor, N. "The Names People Play: Exploring MMOG Players' Avatar Naming Conventions," *Canadian Games Studies Association Symposium*, May 2012.

[4] Lakoff, Robin T. 1975. *Language and Woman's Place*. New York: Harper & Row.

[5] Tannen, D. 1994. *Gender and Discourse.* NY & Oxford: Oxford University Press.

[6] Tannen, D. 1984. *Conversational Style: Analyzing Talk among Friends.* Norwood, NJ: Ablex.

[7] Herring, S. C., and Paolillo, J. C. 2006. "Gender and Genre Variation in Weblogs," *Journal of Sociolinguistics*, 10(4), 439–459.

[8] Herring, S. 1994. "Gender Differences in Computer-Mediated Communication: Bringing Familiar Baggage to the New Frontier," *American Library Association Annual Convention,* Miami, FL.

[9] Gordon, M. and Heath, J. 1998. "Sex, Sound Symbolism, and Sociolinguistics," *Current Anthropology,* vol. 39, no. 4, August/October.

[10] Jespersen, O. 1922. *Language: Its Nature, Development and Origin*. London: Allen and Unwin.

[11] Ohala, J., Hinton, L. and Nichols, J. 1994. *Sound Symbolism*. New York: Cambridge University Press.

[12] Pearce, C. 2009. *Communities of Play: Emergent Cultures in Multiplayer Games and Virtual Worlds*. Cambridge, MA: MIT Press.