An Exploration of Observable Features Related to Blogger Age

John D. Burger and **John C. Henderson**
The MITRE Corporation
202 Burlington Road
Bedford, Massachusetts 01730
{john,jhndrsn}@mitre.org

Abstract

Accurate prediction of blogger age from evidence in the text and metadata of blog entries would be valuable for marketing, privacy, and law enforcement concerns. This paper offers an initial exploratory data analysis of candidate features for blogger age prediction.

Introduction

Personal blogs are an emerging publishing medium containing information of a different type from traditional broadcast news, newswire, and newsgroup data. Text found in blogs is often intimate and detail-oriented, and rarely crafted. The novel types of information prevalent in blogs are perhaps most useful in aggregate form. Phenomena with broad impacts such as disease spread, brand awareness, and reactions to rising fuel prices are interesting to a variety of communities. These trends can be measured in the aggregate in blogs.

Author ages can be found by inspecting profiles of bloggers. This is a new feature, prevalent only in association with this particular text medium. Accurate prediction of blogger age from evidence in the text and metadata of blog entries would be valuable for marketing, privacy, and law enforcement concerns. Models that can predict blogger age from text alone might also be used to predict the age of authors in other media. While highly accurate prediction of blogger age is not yet attainable, we have investigated several informative features of blog posts in the course of evaluating candidate information sources for blogger age prediction. In this paper we look at a large sample of personal blogs and explore how blogger age relates to several other variables.

Stylistic differences of blogs along gender and age lines have been discussed in the literature (Herring *et al.* 2004; Huffaker 2004), but there is little previous work on automatically identifying age from blog content.

Data

From July, 2004, to July, 2005 we accumulated a blog dataset of approximately 85 million posts. As expected, occasional network and hardware failures have interfered with accumulation to varying degrees. This has left us with some gaps in the dataset, but those gaps do not significantly reduce utility. For this study, a subsample of 100,000 posts was randomly drawn from the continuous period of harvesting from December 15, 2004 through January 15, 2005. The number of posts per day in this subsample is shown in Figure 1. Some bloggers are more prolific than others, but 87,883 unique posters were found in this sample. Thus, for most of the bloggers represented in the sample, only one post was selected in the random draw. Validation of marginalizations of the age profile distributions by day of week, day of year or month of year with a comparison to a standard reference dataset from census collectors has proven more difficult than expected, and the authors would welcome pointers to such reference distributions for future work.

To characterize these blogs in a qualitative sense, they are entries in personal journals. They are not community-oriented or agglomerations of expertise on technical sub-

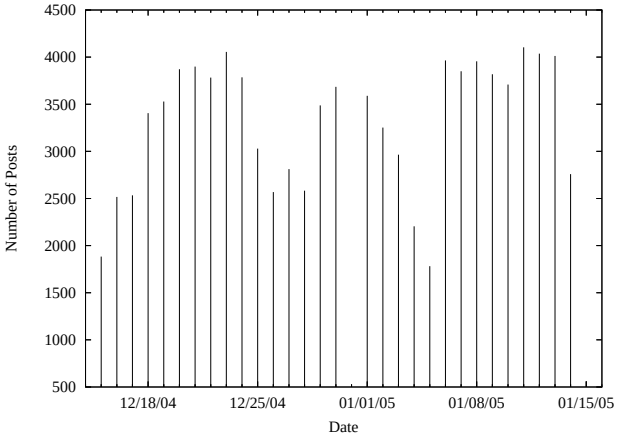

Figure 1: Posts per day in the sample.

DOB in Profile	Entries	Bloggers
Full sample	100,000	87,883
Contains at least YYYY part	53,072	48,157
Full and valid DOB	52,449	47,605

Figure 2: Birthdate reporting in profiles associated with the sample.

Today has been different... good up until just now too. I woke up and went over to my neighbor's house and helped them hook up a new DVD player. That took me all of five minutes.. after that I ran up to the mall to run some errands for my mom. That proved to be semi-fruitless as well. I had to go to the Hallmark store to pick up something that my mom had them put on hold for her. The mall was busy too... I've never recalled it being this busy *sighs*... I thought I'd be coming home to a ghost town, but this place has grown. Probably due to the hurricanes and such. Anyway I went to the store and it was packed. There must have been thirty people in this little mall store... I walk up to the desk and ask the lady if she has what my mother put on hold, and I just watch all the intellegence drain from her face. Needless to say, she knew nothing of holding anything for my mom so... I left empty handed.

Figure 3: Text from a blog used in this study.

jects. They are subjectively-written rather than well-crafted texts. The topics such as announcements of the arrival of the coming weekend, discussions of best friends' attendance at parties, cries of depression at losing a job are intimate and generally personal subjects one would expect to find in a diary. Figure 3 shows an excerpted first paragraph from a blog entry in the sample.

When registering their blogs, bloggers are given the opportunity to declare some of their personal traits such as name, age and gender. The source from which we drew this sample allows registrants to omit the age if they prefer, or to fill in only partial dates. While bloggers can lie about these traits if they choose to, they can just as easily leave the fields blank to provide partial anonymity. In our sample, we found that 47,605 of the 87,883 bloggers (55%) filled out the birthdate field in its full MM-DD-YYYY form. These users were responsible for 52,449 posts in our sample. Some may have lied about their age to appear older or younger, but we felt that in this large sample such effects would be drowned out by those answering honestly. As mentioned above, comparisons of appropriate marginalizations of this distribution to a well-established reference distribution would help validate this hypothesis.

Exploring Blogger Age

Year of birth has been chosen to represent age in this study because it will more easily allow for longitudinal comparisons. The age of a person changes over time, but year of birth remains unchanged. Some care must be taken in calculating a mean year of birth for a sample. A mean of truncated birth years from a sample can differ from the truncated mean of birthdates in a more fine-grained representation. In calculating means mentioned in this work, we first change all DOBs to a fixed offset in days from a reference day. Then, after the mean offset in the sample is calculated, we determine which year contains that offset (truncation).

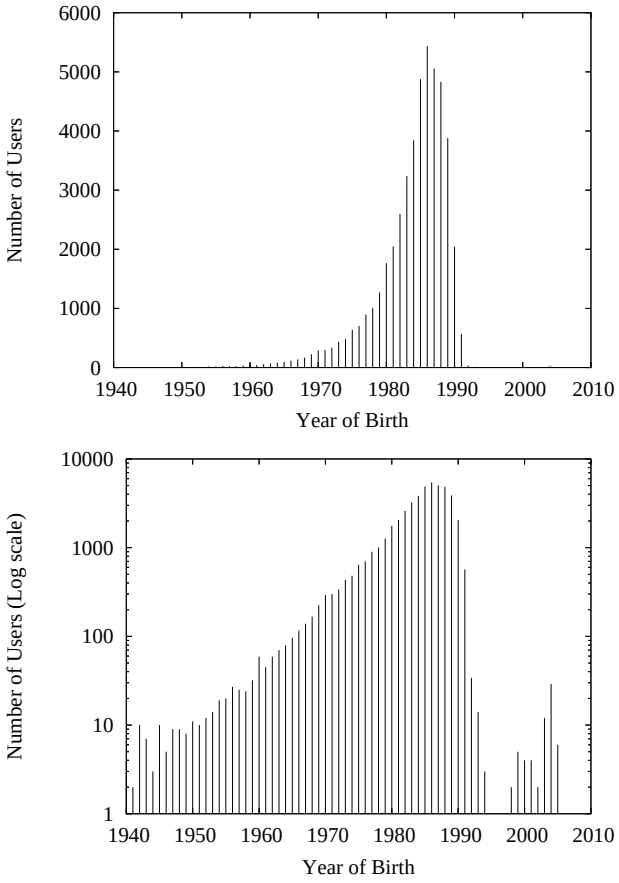

Figure 4: Number of bloggers indicating their year of birth.

	Year of Birth		
Country	**Mean**	**Std.D.**	**# Users**
Philippines	1985	3.5	300
Finland	1985	4.4	292
Spain	1984	4.7	138
Scotland	1984	5.6	113
Netherlands	1984	5.8	240
United States	1984	6.5	50450
Singapore	1984	7.1	351
Australia	1983	6.6	1278
United Kingdom	1983	6.7	2683
Canada	1983	7.1	3597
New Zealand	1983	7.3	183
None	1983	9.1	17361
Japan	1982	4.9	173
Germany	1982	5.7	348
France	1981	6.4	118
Estonia	1981	6.6	110
Russian Federation	1980	8.6	3979
Belarus	1980	9.1	114
Ukraine	1979	9.1	383
Israel	1977	7.6	242

Figure 5: Mean blogger birth year by country ($n \geq 100$).

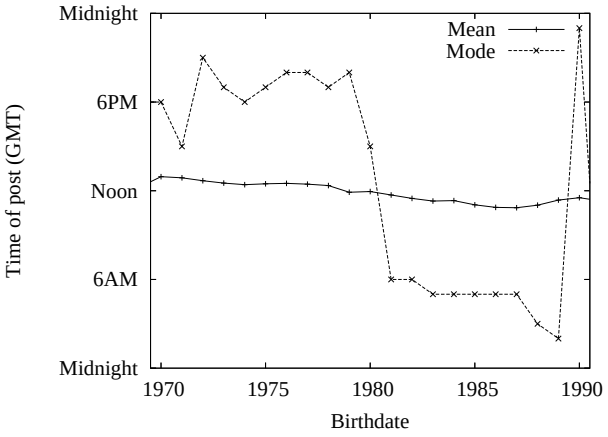

Figure 6: Mean (to the minute) and mode (binned by hour) of posting time (Greenwich Mean Time).

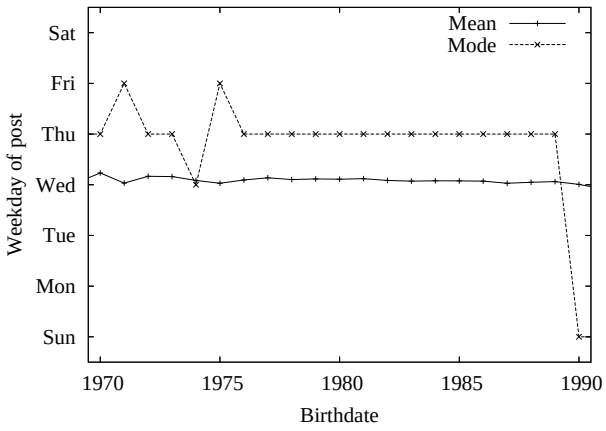

Figure 7: Mean and mode of weekday of posting.

Figure 4 shows the general distribution of the self-reported year of birth in our sample. The top plot shows the smoothness of the curve and the tightness of the predominance of posters in the age range of 14–24. The bottom plot shows the log-linear regularity of the curve as it moves to older bloggers (with earlier years of birth). According to this data, bloggers start posting around age 14 and taper off gradually after age 18. The cohorts from the years 1970–1990 each have at least 100 users, large enough samples to be examined in more detail in later graphs. The small set of posters around the age of 5 is inconsistent with the rest of the plot, and likely consists of blogs either written by or strongly encouraged by their parents.

Location

Figure 5 shows the mean year of birth, standard deviation, and number of bloggers in countries that were represented by at least 100 bloggers. Unsurprisingly, the majority of bloggers in our sample state that they reside in the U.S. with a mean age of 21. The Philippines has the youngest bloggers and Israel's bloggers report the highest ages. The eight-year difference in mean age between bloggers in those countries is surprisingly wide, suggesting one or more cultural biases influencing the age of computer users or an underlying demographic difference.

Time

Figure 6 shows the mean and mode of the time of day that bloggers post as given by the submission time on their entry. The units shown are relative to the GMT timezone. The mode is the more indicative statistic for this data because of the periodic nature of the dependent variable. The trend of the mean with respect to the year of birth is indicative, but the offset is an artifact of the coordinate system with zero set to midnight.

Keeping in mind that the majority of the bloggers are from the U.S., we see that bloggers under the age of 24 are submitting their post late in the evening. 5 AM GMT is midnight EST, suggesting that younger bloggers in the three timezones of the continental U.S. are posting between 9 PM and midnight. Also notice that bloggers' posting times creep later in the evening until they reach age 23 or 24 (perhaps college graduation) at which point afternoon posting is more common.

Figure 7 shows the mean and mode of the day of the week that blog posts are made. These curves suggest that there is no dominant change in the day of the week that bloggers post. The right-most point on the mode curve, at 1990, could be an artifact of flatness of the distributions, or suggestive of multimodality. These two phenomena are indistinguishable from this view, but the lack of a predominant differentiation by year of birth makes the outcome of further study of this variable pessimistic.

This dataset spans only 30 days of posting. While there were five opportunities to observe postings on most of the days of the week, there were only four opportunities to observe Monday postings and Tuesday postings. The holidays observed during that period could also have affected the counts from the day of the week. While these factors were independent of bloggers' ages, the effect from time-of-day of posting observed above could have interacted with observation of this variable. A follow-up study spanning more weeks could resolve these issues with more evidence.

Language/script

Reliable language identification of noisy text such as that found in blogs would be useful. As a stand-in for language ID, script identification is an easily observed shallow textual feature. Figure 8 shows the mean year of birth of users of the most-represented scripts in our dataset. Note that these counts are not disjoint, as many bloggers use a mix of scripts in their blogs. This data is consistent with our observations about blogger location above. Users writing Cyrillic characters, are older, on average than those users writing only with Latin. A larger sample would allow us to make more claims about users of these scripts.

Script	Year of Birth Mean	Std.D.	# Users
HANGUL	1986	2.4	6
GREEK	1985	3.3	9
KATAKANA	1985	2.6	10
LATIN	1984	7.1	50108
FULLWIDTH	1984	1.7	3
KATAKANA-HIRAGANA	1984	2.1	6
CJK	1983	6.2	45
HIRAGANA	1983	5.4	16
CYRILLIC	1979	9.2	4351
HEBREW	1977	3.5	4

Figure 8: Mean blogger birth year for various script usages ($n \geq 2$).

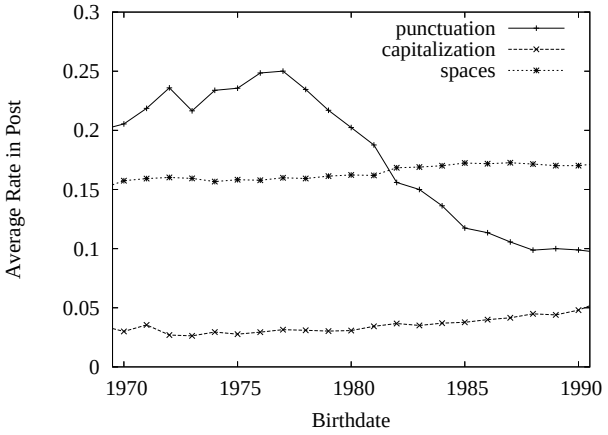

Figure 9: Rates of punctuation, uppercase and space character usage.

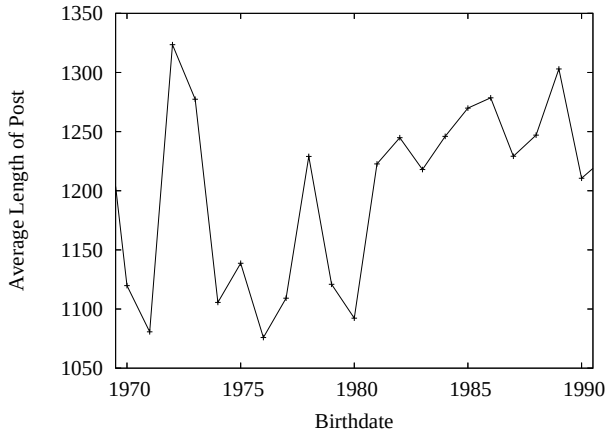

Figure 10: Mean post length.

Figure 11: Mean number of non-image links per post

Text features

We hoped to find some indicators of blogger age in the textual content of the post. The most obvious textual feature is simply the length of a posting. Figure 10 shows the length of the text part of a blog entry as measured in characters versus the age of the blogger. Posting length increases with year of birth. That is, older bloggers are posting shorter entries. This is a noisy trend, but it appears strongly when looking to either side of the 25-year-old dividing threshold (1980).

Other shallow textual features include the fraction of characters in a post that are punctuation, or that are capitalized letters. The portion of characters that are spaces can stand in as inversely related to average word length. As Figure 9 indicates, these are opposites in that younger bloggers use more capitalization, but less punctuation and slightly shorter words. Furthermore, punctuation is the most indicative of this set of features.

Deeper textual features, such as the strings of characters in a post, senses of words, or rates of dangling pronominal references might also be useful indicators of blogger age. The quality of language processing for these features varies even on clean text, though. This dataset provides an interesting challenge for researchers interested in developing language processing techniques for simultaneously dealing with multiple languages, noisy data, and sparse features that so far have only proven to be indicative of age in the aggregate.

Links and images

Bloggers often post links to other web pages or images that they find interesting. Figure 11 shows the average number of links, excluding embedded images, for the various age cohorts. Excepting the sharp spike in the age 24 cohort, there appears to be a gentle downward trend—apparently younger bloggers post fewer links. The spike appears to be caused by a single post with over 500 links in it. We have examined similar averages for embedded images, and there appears to be no trend with respect to age for image references in this dataset.

Friends

Figure 12 shows the number of friends indicated by the profiles of bloggers. These are counts of explicitly declared

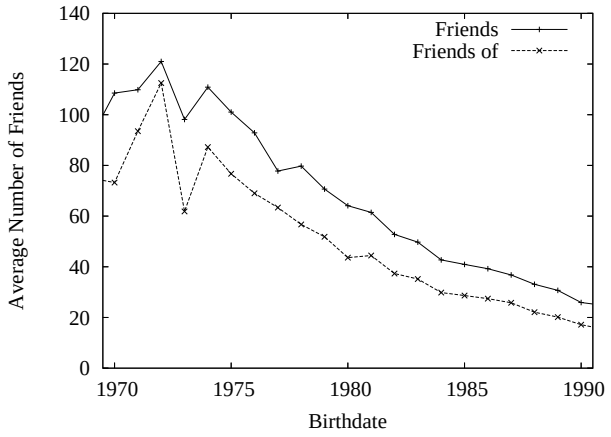

Figure 12: Mean number of declared "friends" and "friends-of".

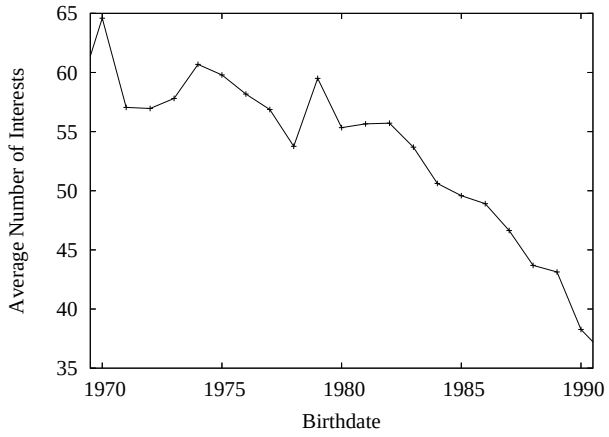

Figure 13: Mean number of interests.

Feature	Posts	Birthyear
Mood - No	22436	1982
Mood - Yes	30013	1984
Music - No	30611	1982
Music - Yes	21838	1985

Figure 14: Mood and music are indicated by younger bloggers.

"friend links" to other bloggers at the same site, offered as an aide for quick navigation and declaration of a relationship. Older bloggers declare more friends. On average, they accumulate at a rate of roughly 4 per year between the ages of 15 and 30. The graph also displays the average counts for the reverse links, "friends of". The latter counts lag the former for all age cohorts, quite consistently. If the graph included all of the data for the blogger site in question, we would expect the areas under the two curves to be equal, but no such constraint applies to a sample of the data such as this. More friendship links point out of our sample than point into it, and this appears to be true for all age cohorts. Follow-up investigation should explore distributional effects of the prevalence of large friendship "hub" users at different ages.

Interests

While creating their profile, some bloggers indicate their interests by entering collections of short phrases. Figure 13 is a plot of the number of interests declared by bloggers of various ages. It generally indicates that older bloggers have more interests. They acquire an additional 15 interests between the ages of 15 and 25.

Figure 15 shows the mean age for twenty of the most age-skewed interests declared by more than 1000 users. The youngest users report interests in musical bands and new technology such as "cell phone", while the older users are interested in more mature subjects and arcane technology such as "science fiction" books.

Oddities

Figure 14 shows the mean age of bloggers that indicate their mood and their music and the mean age of those that do not. On mean, bloggers that indicate either of these features are two years younger than bloggers who do not indicate. The differences of these means are statistically significant with high confidence levels. The statistics are supported by large populations on either side. Note also that these two signals are observed independently in the two different populations.

Age Prediction

Many of the features described above seem to have some correlation with blogger age, and we have begun to explore using them in statistically trained classifiers. Ideally, a regression model could be employed to accurately predict actual blogger age as a continuous value, but our initial experiments involve the simpler task of binary classification with respect to an age threshold, in this case the significant 18-year-old boundary with an associated large set of legal and privacy ramifications. Even this simpler task has turned out to be surprisingly difficult, however. Our best classifier is able to reduce the error by a only a few percent, compared to a baseline that always guesses the majority class (18 and over).

We have also begun to experiment with language models built from the under- and over-18 postings, using the SRI Language Modeling Toolkit (Stolcke 2002). We hoped to leverage possible stylistic differences in minority and majority authors, but thus far this has proven fruitless as well—the language models do not appear to sufficiently discriminate between the two text classes. Even when various language model scores are combined with some of the other features described above in a maximum entropy classifier (Daumé III 2004), we have yet to find good discriminators.

Conclusion

We have presented explorations of several features of bloggers and blog entries change with blogger age, with an eye toward predicting age from textual content of a blog entry and other blogger metadata. Much of the metadata appears

Interest	Year of Birth Mean	Std.D.	# Users
fuse	1988	1.8	125
mall	1988	1.9	361
pac sun	1988	2.0	119
hawthorne heights	1988	2.0	591
the spill canvas	1988	2.0	170
from first to last	1988	2.0	276
silverstein	1988	2.0	318
daphne loves derby	1988	2.1	132
thirteen	1988	2.3	159
the mall	1988	2.3	166
smile empty soul	1988	2.4	313
cell phone	1988	2.4	124
hollister	1988	2.4	406
chad michael murray	1988	2.6	181
laguna beach	1988	2.7	237
phone	1988	7.3	351
letter kills	1987	1.9	142
early november	1987	1.9	108
hellogoodbye	1987	2.1	248
chillin	1987	2.1	160
dead poetic	1987	2.2	257
the bled	1987	2.2	260
rooney	1987	2.2	358
boys night out	1987	2.2	279
american hi-fi	1987	2.2	109
unix	1977	8.6	179
bdsm	1977	8.8	565
science fiction	1977	8.9	1366
herbs	1977	9.2	195
дети (“children”)	1977	9.2	194
neal stephenson	1977	9.4	122
costuming	1977	10.0	171
водка (“vodka”)	1977	10.2	111
женщины (“women”)	1977	10.3	269
garlic	1977	10.4	120
dead can dance	1977	10.4	218
latex	1977	10.4	153
goddess	1976	8.0	113
babylon 5	1976	8.5	317
bears	1976	8.9	224
reiki	1976	9.0	115
doctor who	1976	9.4	159
shamanism	1976	10.8	160
spanking	1976	12.3	115
highlander	1976	12.5	120
живопись (“painting”)	1976	13.5	124
sca	1975	7.4	197
gardening	1975	10.2	429
parenting	1974	7.4	133
polyamory	1974	7.4	291

Figure 15: Mean birth year for blogger interests (youngest and oldest 25 with $n \geq 100$).

to be indicative of the birth year of the blogger, and some of the shallow textual evidence is weakly indicative.

In future work we plan to incorporate more fine-grained text features into our classifiers, with the hope of predicting blogger age for previously unseen posts. Given how difficult it has proven to build an age predictor, we are also planning some small-scale human annotation experiments, to see how well people can judge postings with respect to the 18-year-old boundary. Following accurate prediction of blogger age, we would be interested in determining how well such models carry over to prediction of author age in other textual genres such as newswire, broadcast news transcriptions, chat logs, and forensic studies.

Finally, we welcome others in the community to use this dataset to try out other approaches to age prediction. This task may provide a good proving ground for new algorithmic insights, and perhaps for deeper language understanding as well.

Acknowledgements

This work was funded under MITRE’s internal research program.

References

Daumé III, H. 2004. Notes on CG and LM-BFGS optimization of logistic regression. Paper available at `http://www.isi.edu/~hdaume/docs/daume04cg-bfgs.ps`, implementation available at `http://www.isi.edu/~hdaume/megam/`.

Herring, S. C.; Kouper, I.; Scheidt, L. A.; and Wright, E. L. 2004. Women and children last: The discursive construction of weblogs. In Gurak, L.; Antonijevic, S.; Johnson, L.; Ratliff, C.; and Reyman, J., eds., *Into the Blogosphere: Rhetoric, Community, and Culture of Weblogs*. `http://blog.lib.umn.edu/blogosphere/`.

Huffaker, D. 2004. Gender similarities and differences in online identity and language use among teenage bloggers. Master’s thesis, Georgetown University. `http://cct.georgetown.edu/thesis/DavidHuffaker.pdf`.

Stolcke, A. 2002. SRILM—an extensible language modeling toolkit. In *Proc. Intl. Conf. on Spoken Language Processing*.