Computational analysis to explore authors' depiction of characters

Joseph Bullard
Dept. of Computer Science
Rochester Institute of Technology
jtb4478@cs.rit.edu

Cecilia Ovesdotter Alm
Dept. of English
Rochester Institute of Technology
coagla@rit.edu

Abstract

This study involves automatically identifying the sociolinguistic characteristics of fictional characters in plays by analyzing their written "speech". We discuss three binary classification problems: predicting the characters' gender (male vs. female), age (young vs. old), and socioeconomic standing (upper-middle class vs. lower class). The text corpus used is an annotated collection of August Strindberg and Henrik Ibsen plays, translated into English, which are in the public domain. These playwrights were chosen for their known attention to relevant socioeconomic issues in their work. Linguistic and textual cues are extracted from the characters' lines (turns) for modeling purposes. We report on the dataset as well as the performance and important features when predicting each of the sociolinguistic characteristics, comparing intra- and inter-author testing.

1 Introduction

A speech community has sociolinguistic properties. Social variables influencing verbal interaction include, for example, geographical background, gender, age, ethnicity, and class. Writers and playwrights, in turn, use their knowledge of social verbal markers to generate credible and compelling characters. The focus of this study is the creation of an annotated dataset and computational model for predicting the social-biographical aspects of fictional characters based on features of their written "speech" in dramatic plays. The plays used here are authored by August Strindberg and Henrik Ibsen, two Scandinavian playwrights known for creating characters and stories that acted as social commentary and were controversial when they were first written. These authors are also recognized for their contributions in shaping modern drama. Their attention to social issues makes these plays and characters highly relevant in constructing such a model to shed light on how these authors' translated texts portray social variables. Interlocutors' social attributes (such as their gender, age, social class, and ethnicity) are known to correlate with language behavior, and they tap into dimensions of language behavior that are of central interest to the humanities. For instance, anecdotal evidence suggests that large-scale corpus analysis can show how society collectively ascribes certain roles to male versus female referents in text (cf. Lindquist, 2009).

Studying these authors and texts from the point of view of corpus-oriented computational sociolinguistics can also help us examine the authors' differences in production, descriptively. This is useful as a complementary approach to the more traditional close reading methodology common in literary research, through which their texts are usually approached. On a broader scale, the study can contribute valuable insights to a theory of linguistic text criticism. These authors are part of a global literary canon, and their plays are arguably more often performed in translation than in their Scandinavian originals. Accordingly, we focus on analyzing texts translated into English.

We focus on sociolinguistic characteristics that are assigned to each character and that can be described as translating into three binary classification problems: predicting the characters' gender (*male* vs. *female*), age (*young* vs. *old*), and socioeconomic standing or class (*upper-middle class* vs. *lower class*). The text corpus is annotated by assigning each of the characters that match specified criteria a value in each of the characteristics. We do this at the character level, joining all dialogic lines of a character into one instance. The work was accomplished through the use of computational tools

Proceedings of the 3rd Workshop on Computational Linguistics for Literature (CLfL) @ EACL 2014, pages 11–16,
Gothenburg, Sweden, April 27, 2014.

for natural language processing, including Python (http://www.python.org/), the Natural Language Toolkit (http://www.nltk.org/) for part of the preprocessing, and the `scikit-learn` machine learning library for the computational modeling. Translated texts that reside in the public domain were collected from the Gutenberg Archive (http://www.gutenburg.org/wiki/Main_Page/).

2 Previous Work

A pilot study by Hota et al. (2006) on automatic gender identification in Shakespeare's texts, as well as a few primarily gender-oriented studies surveyed in Garera and Yarowsky (2009), have set the stage for further inquiry. The latter study expanded on previous work by exploring three attributes: gender, age, and native/non-native speakers. There have been previous avenues of research into categorizing speakers based on different individual sociolinguistic factors. However, not many studies have attempted this categorization with fictional characters. Literary texts are complex, reflecting authors' decision-making and creative processes. From the perspective of digital humanities, such a focus complements computational sociolinguistic modeling of contemporary user-generated text types (such as emails, or blogs (Rosenthal and McKeown, 2011)). As Lindquist (2009) points out, social data for interlocutors is less often attached to openly available linguistic corpora, and interest is strong in developing corpus methods to help explore social language behavior (see Lindquist (2009) and Baker (2010)).

Previous investigation into social dimensions of language has established strong links between language and social attributes of speech communities (for an overview, see Mesthrie et al. (2009)). However, such inquiry has generally had a firm foundation in field-based research and has usually focused on one or just a few linguistic variables (such as how the pronunciation of certain sounds aligns with social stratification (Labov, 1972)). Moreover, previous scholarship has chiefly focused on the spoken rather than the written mode. Garera and Yarowsky (2009) and Boulis and Ostendorf (2005) take into account the interlocutors' speech for analysis. In contrast, we experiment with the challenge of using only sociolinguistically relevant knowledge coded in the text of characters' lines. Thus, our approach is more similar to Hota et al.'s (2006) work on Shakespeare. The characters' lines do not include the metadata needed for considering spoken features, since usually these are added at the discretion of the performer. This may make our problem more challenging, since some of these indicators may be reliable for identifying gender, such as backchannel responses and affirmations from females, and assertively "holding the floor" with filled pauses from males (Boulis and Ostendorf, 2005). Moreover, there are prosodic features that clearly differ between males and females due to physical characteristics (e.g. F_0, predominant for pitch perception). We do not take advantage of acoustic/prosodic cues in this work. Our text is also artificial discourse, as opposed to natural speech; therefore these characters' lines may rather express how writers choose to convey sociolinguistic attributes of their characters.

In terms of features, we have explored observations from previous studies. For instance, common lexical items have been shown successful, with males tending to use more obscenities, especially when talking to other males (Boulis and Ostendorf, 2005), and females tending to use more third-person pronouns. Phrases also tended to be more useful than unigrams, though whether the commonly-used words tend to be content-bearing remains a question according to Boulis and Ostendorf (2005). Tackling another form of text, Kao and Jurafksy (2012) examined the statistical properties of 20th century acknowledged versus amateur poets in terms of style and content substance, finding, for example, that lexical affluence and properties coherent with imagism, as an aesthetic theorized ideal, distinguished contemporary professionals' poetics, while sound phenomena played a lesser role, and amateurs preferred the use of more explicit negative vocabulary than professionals. In our study, we focus on data collection, corpus analysis, and exploratory experimentation with classification algorithms.

3 Data

The texts used were freely available transcriptions from the Gutenberg Archive. English translations of public-domain plays by August Strindberg and Henrik Ibsen were collected from the archive, from various translators and years of release. As noted above, these plays are often performed in English, and we assume that the translations will convey relevant linguistic cues, as influenced by

	Strindberg	Ibsen	**Total**
# of plays	11	12	**23**
# of characters	65	93	**158**
# of lines	6555	12306	**18861**

Table 1: Distribution of plays, characters, and lines between Strindberg and Ibsen in the dataset.

Character	Gender	Age	Class
Christine	Female	Young	Upper
Jean	Male	Young	Lower
Miss Julia	Female	Young	Lower

Table 2: Example annotations from *Miss Julia*.

authors, as well as translators. We assume that the translators intended to replicate as closely as possible the voice of the original author, as this is generally the function of literary translation, but we recognize the potential for loss of information.

The texts were minimally pre-processed (such as removing licensing and introduction text), leaving only the written lines-to-be-spoken of the characters. Each character's lines were automatically extracted and aggregated using a Python script. Characters should have a significant number of lines (equal to or greater than fifteen in his or her respective play) to be considered.[1] We also record metadata per character, such as the play title, the play translator, and the URL of the original play text on Gutenberg. The basic characteristics of the resulting dataset are shown in Table 1.

In terms of annotation, characters from each play were annotated by a third party and assigned characteristics primarily according to the plot descriptions on Wikipedia of their respective plays of origin. The characteristics considered were gender (male vs. female), age (young vs. old), and socioeconomic standing or class (upper-middle class vs. lower class). For example, for *age*, characters with children are considered *old*, and those children are considered *young*. A childless character whose peers have children or who has experienced life-changing events typically associated with age (e.g. widows/widowers) is also *old*, unless separately noted otherwise. The *gender* annotations were validated by a project-independent person

[1]The only exception to this rule is Mrs. X from Strindberg's *The Stronger*. She has only 11 separate "lines", but also has the only speaking part for the entire play, which is a single act of substantial length. We also note that while an ad hoc threshold for lines was used, future work could explore principled ways to set it.

Attribute	Annotation	Strindberg	Ibsen
Gender	Male / Female	42 / 23	61 / 32
Age	Old / Young	46 / 19	61 / 32
Class	Upper / Lower	57 / 8	83 / 10

Table 3: Character attribute distributions for *gender*, *age*, and *class* for each author.

in Scandinavia (Swedish native speaker) based on her knowledge of Scandinavian naming conventions. Example character annotations for Strindberg's well-known naturalistic play *Miss Julia* (or *Miss Julie*) are shown in Table 2. As seen in Table 3, the imbalance of *class* labels presents the greatest problem for our model. Baselines of 88% and 89% *upper class* for Strindberg and Ibsen, respectively, indicate that there may be less information to be extracted for *class*.

4 Models

Here we describe the design and performance of computational models for predicting a character's *gender*, *age*, and *class* for Strindberg and Ibsen, yielding six models in total. Logistic regression, implemented in Python with the `scikit-learn` machine learning library (Pedregosa et al., 2011), is used for all classification models.

4.1 Feature Extraction

Many features were examined, some inspired by previous analyses in the literature, such as type-token ratio, subordinate clauses, and *wh*-questions, as well as some exploratory features, such as honorific terms of address. A full list of the features examined is shown in Table 4. All features were automatically extracted using Python. We use *honorifics* here to mean common formal terms of address during the time period (*sir, madam, lord, Mr., Mrs.,* etc.). It seems intuitive that such terms may be used differently based on *class* or possibly *age* (e.g. lower class using more higher terms of address when speaking to their superiors). We use *family* words to mean anything that indicates a familial relationship (*father, daughter, nephew,* etc.). The use of such words may be affected by gender roles (Hota et al., 2006). Part-of-speech tagging was accomplished using the Natural Language Toolkit (NLTK) (Bird et al., 2009).

Linguistic features	
Family words	For/with
Honorifics	Modals
Pronouns 1st	Personal pronouns
Pronouns 2nd	Nouns singular
Pronouns 3rd	Nouns plural
Pronouns all	Verbs past
Wh- questions	Verbs past part.
Type-token ratio	Verbs sing. pres. non-3rd
Determiners	Mean line length
Adjectives	Number of lines
Prepositions	% short lines ($\leq$5 words)

Table 4: List of linguistic features examined for the models. All features, with the exception of the last three in the right column, were measured once as raw counts and once as the fraction of the overall words for a given character.

4.2 Cross-Author Validation

We compared translations of Strindberg and Ibsen's use of language to convey sociolinguistic attributes. This was done for each of the three attributes of interest (*gender*, *age*, and *class*) by training one model for each author, then using it to classify the other author's characters. We accomplish this by defining a *cross-author validation* procedure, a variation of the standard k-fold cross-validation procedure in which the trained model in each fold is used to predict both its own test set and the test set of the other author. This procedure is explained visually in Figure 1. The procedure is especially interesting as these two authors were contemporaries and dealt with topics of social commentary in their works, although from their own perspectives.

The results of cross-author validation are shown in Table 5 as a matrix where the row is the author used for training, the column is the author used for testing, and the value inside a cell is the average accuracy over all iterations of cross-author validation. Majority class baselines are also shown. As expected, the models for each author's texts were better at predicting themselves than the other author, with a couple of exceptions. For *age*, the Strindberg-trained model was still able to improve on Ibsen's baseline, but not vice versa. One possible explanation could be that common features between their depictions of *age* might be more useful for one author than the other. Another interesting exception is in the *class* models

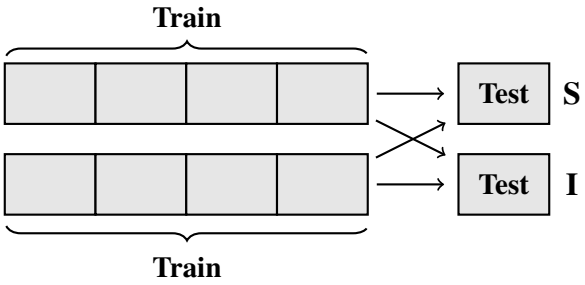

Figure 1: Example of one fold of *cross-author validation* for Strindberg (S) and Ibsen (I). Arrows indicate testing. Each author has its own 5-fold cross-validation, but in each fold, the trained model is tested on both its own test set and the test set of the other author.

	Gender		Age		Class	
	S	I	S	I	S	I
Strindberg (S)	**68**	60	**74**	**70**	**89**	**90**
Ibsen (I)	61	**67**	70	**74**	**91**	**90**
Baseline	65	66	71	66	88	89

Table 5: Results of cross-author validation (see Figure 1). Rows are the author used for training, columns are the author used for testing, and the value in the cell is the average accuracy over 500 iterations of 5-fold cross-validation. Accuracies above majority class baselines are shown in bold.

for both authors, which performed slightly above high baselines for the opposite authors as well as their own. While *class* improvements are recognizably marginal (and not claimed to be significant), these results might indicate that the two authors' translated texts are using similar characteristics to convey social class of their characters. It is important to note that the baselines for *class* were extremely high, making prediction of this attribute more difficult. At least in the intra-author testing, the *gender* and *age* models were generally able to improve accuracy over their respective baselines more so than the *class* models, with *age* being the best overall.

4.3 Comparison of Useful Features

Since the experimentation used a linear model (logistic regression), we can inspect the coefficients/weights of a trained classifier to determine which features contributed particularly to the classification. The absolute value of a coefficient indicates how influential its feature is, and the sign (+/-) of the coefficient indicates which class the feature is associated with. During cross-author

	Strindberg		Ibsen	
Gender	**Pronouns 3rd**	**Female**	**Pronouns 3rd**	**Female**
	Honorifics	Female	Family words	Female
	Determiners	Male	Modals	Male
Age	Nouns singular	Old	**Family words**	**Young**
	Family words	**Young**	Verbs sing. pres. non-3rd	Young
	Modals	Young	Prepositions	Old
Class	**For/with**	**Lower**	**For/with**	**Lower**
	Verbs past part.	Upper	**Honorifics**	**Lower**
	Honorifics	**Lower**	Nouns singular	Lower

Table 6: Most useful features for *gender*, *age*, and *class* for each author, determined by examining the coefficients of classifiers that performed above baseline during cross-author testing. The pairs in the table consist of a linguistic feature and the label indicated by more frequent use of that feature (e.g. for Strindberg, third-person pronoun usage contributed to predicting gender, with greater usage indicating a female character). Features marked in bold are shared between authors for a given attribute.

validation, if the trained classifier for a given fold performed above the baseline of its own test set, then we record its three most heavily weighted features. At the end, we have a tally of which features most often appeared in the top three features for such better-performing classifiers. We can use this to compare which features were more consistently involved for each author and attribute pair, as shown in Table 6.

Some of the useful features are more intuitive than others. For example, as mentioned in an earlier section, it seems reasonable that family words may relate to depictions of gender roles of the time period in which the plays were written, with women being expected to take on social roles more confined to the home. This appears to be true for Ibsen, but not for Strindberg. We also see family words suggesting young characters for both authors' texts. It seems intuitive that authors may have chosen to depict children as spending more time around family members, including using family terminology as terms of address. The use of honorifics is also as predicted earlier in the paper: lower class characters use more higher terms of address, presumably when interacting with their superiors. Another interesting result is the frequency of third-person pronouns being the most useful predictor of gender, indicating female characters for both authors. Possibly, women may have spoken more about other people than men did in these texts.

Some other results are not as easy to explain. For example, the use of the prepositions *for* and *with* was consistently the most useful predictor of lower class characters (which could explain why the models performed comparably on opposite authors in Table 5). An interesting result was the more frequent use of singular, present tense, non-third person verbs among young characters in the Ibsen texts. This suggests that young characters used more verbs centered around *I* and *you* in the present tense. One possible explanation is that children were depicted as being more involved in their own personal world, speaking less about people they were not directly interacting with in a given moment.

5 Conclusion

We have presented a dataset of translated plays by August Strindberg and Henrik Ibsen, along with computational models for predicting the sociolinguistic attributes of gender, age, and social class of characters using the aggregation of their textual lines-to-be-spoken. We compared the performance and important features of the models in both intra- and inter-author testing using a cross-author validation procedure, finding that models generally performed above challenging baselines for their own authors, but less so for the other, as one would expect. The exception was the social class variable, which was consistently slightly above baseline regardless of the author used for testing. While this could indicate that the translated Strindberg and Ibsen texts conveyed social class using similar linguistic cues, this remains a topic for future exploration, given the class imbalance for that attribute. We also examine some indicative features for each attribute and author pair, identifying similarities and differences be-

tween the depictions in each set of texts. This analysis supported the trends seen in the cross-author testing.

Future work would include exploring other authors and literary genres, or extending the scope to non-literary domains. When expanding this initial work to larger datasets, there is an opportunity to better understand the intricacies of performance through other metrics (e.g. precision, recall). There is certainly much opportunity to expand sociolinguistic features on fictional texts and to explore other potentially simpler or more advanced modeling frameworks. Alternatives for assigning annotation of sociolinguistic variables, such as socioeconomic standing, also deserve further attention. Additionally, it would be interesting to verify the preservation of linguistic/sociolinguistic cues in translation by repeating this work using different translations of the same texts.

Acknowledgements

We thank the Swedish Institute (http://eng.si.se) for partially supporting this work. We also thank the reviewers for valuable comments that were considered in the revision of this paper.

References

Paul Baker. 2010. *Sociolinguistics and Corpus Linguistics*. Edinburgh University Press, Edinburgh.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python – Analyzing Text with the Natural Language Toolkit*. O'Reilly Media, Sebastopol.

Constantinos Boulis and Mari Ostendorf. 2005. A quantitative analysis of lexical differences between genders in telephone conversations. In *Proceedings of the 43rd Annual Meeting of the ACL*, pages 435–442, Ann Arbor, MI, USA, June.

Nikesh Garera and David Yarowsky. 2009. Modeling latent biographic attributes in conversational genres. In *Proceedings of the 47th Annual Meeting of the ACL and 4th IJCNLP of the AFNLP*, pages 719–718, Suntec, Singapore, August.

Sobhan Raj Hota, Shlomo Argamon, and Rebecca Chung. 2006. Gender in Shakespeare: Automatic stylistics gender character classification using syntactic, lexical and lemma features. In *Digital Humanities and Computer Science (DHCS 2006)*.

Justine Kao and Dan Jurafsky. 2012. A computational analysis of style, affect, and imagery in contemporary poetry. In *Workshop on Computational Linguistics for Literature*, pages 8–17, Montréal, Canada, June 8.

William Labov. 1972. *Sociolinguistic Patterns*. University of Pennsylvania Press, Philadelphia, PA.

Hans Lindquist. 2009. *Corpus Linguistics and the Description of English*. Edinburgh University Press, Edinburgh.

Rajend Mesthrie, Joan Swann, Anna Deumert, and William Leap. 2009. *Introducing Sociolinguistics (2nd ed.)*. Jon Benjamins, Amsterdam.

Fabian Pedregosa, Gael Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Edouard Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Sara Rosenthal and Kathleen McKeown. 2011. Age prediction in blogs: A study of style, content, and online behavior in pre- and post-social media generations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, pages 763–772, Portland, Oregon, June 19-24.