Exploring Cultural Differences in Language Usage: The Case of Negation

Svetlozara Stoytcheva
Graduate School of Library & Information Science
University of Illinois at Urbana-Champaign
501 E. Daniel Street,
Champaign, IL 61820
stytchv2@illinois.edu

Dov Cohen
Department of Psychology
University of Illinois at Urbana-Champaign
603 E. Daniel Street,
Champaign, IL 61820
dovcohen@illinois.edu

Catherine Blake
Graduate School of Library & Information Science
University of Illinois at Urbana-Champaign
501 E. Daniel Street,
Champaign, IL 61820
clblake@illinois.edu

ABSTRACT

Prior research suggests that speakers of Asian languages are more likely to use negation than English speakers. Our goal in this work is to explore this theory using empirical data from news stories. Specifically, we used natural language processing to compare negation usage in two newspapers: the *New York Times* and *Xinhua News* (English Edition). Overall, negation represents 0.55% of typed dependencies in the *New York Times* (versus 0.18% in *Xinhua News*). Additionally, 9.28% of sentences and 86.56% of articles in the *New York Times* contain one or more instances of negation (compared to 3.33% of sentences and 24.94% of articles in *Xinhua News*). In contrast to the prevalent theory, negation is approximately three times more common in the *New York Times* than in *Xinhua News* (English Edition).

Keywords

Text mining, natural language processing, language & culture.

INTRODUCTION

Prior research in social psychology indicates language use is strongly influenced by cultural backgrounds. This can be a function of different cultural syndromes (Triandis, 1994), epistemologies (Nisbett, 2003), or semantic and grammatical features of a language. There is some reason to suspect this would be the case based on Asian epistemologies that emphasize absence and negation more than Western epistemologies do (Jullien, 2007), social norms in a face culture that emphasize indirect rather than direct communication (Triandis, 1994; Leung & Cohen, 2011), a prevention-focused orientation (emphasizing failures to be avoided rather than successes to be achieved (Lee, Aaker, & Gardner, 2000), and the less expansive vocabulary of most Asian languages (as compared to English).

ASIS&T 2014 November 1-4, 2014, Seattle, WA, USA.

Previous research in Dr. Cohen's lab has confirmed that – in the domain of morality – Asian language speakers prefer that ethical injunctions be expressed in terms of what one should not do (rather than what one should do). This suggests that negation is more common in texts produced by speakers of Asian languages.

One goal of this poster is to further explore these theories using empirical data from new stories. Rather than hand coding each story for the presence of linguistic features, which is both time- and labor-intensive, we introduce computational techniques that scale.

This poster outlines the application of natural language processing methods to study the use of negation in a large corpus of newspaper articles. We measured how frequently negation occurs in approximately three years' worth of articles from the *New York Times* and *Xinhua News* (English Edition). Based on the prior work, our hypothesis was that negation would be more common in articles from *Xinhua News* than in articles from the *New York Times*. However, the opposite turns out to be true: negation is roughly three times more frequent in the *New York Times*.

METHODS

The data used in this research comes from the AQUAINT Corpus of English News Texts, which contains full-text articles from the *New York Times*, the *AP Newswire*, and *Xinhua News* (English Edition). The period of coverage is 1998-2000 (LDC, 2013). For another example of how this data has been used for natural language processing research, see Llorens et al. (2012).

All of the available articles in the collection were considered in this analysis. After preprocessing to tokenize the articles into sentences, the texts were parsed using

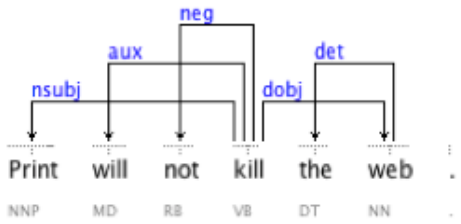

Figure 1 Example of a typed dependency structure.

version 3.2 of the Stanford Natural Language Parser.[1] For a given sentence, the parser identifies the part of speech of each word in the sentence, as well as the grammatical relationships between words in the sentence. These grammatical relationships are called typed dependencies.

Negation is one kind of typed dependency recognized by the parser; it is identified by the presence of a negation modifier: not, never, or the contraction nt or n't. Figure 1 shows an example of a typed dependency tree generated by the Stanford Parser for the sentence: "Print will not kill the web." De Marneffe et al. (2006) describe the rules governing the parser's dependency recognition and explain the difference between the Stanford Parser's typed dependency schema and those used by other parsers. The grammar rules used to program each parser affect its output and none of the parsers are 100% accurate in their assignment of dependency relationships. However, the Stanford parser has been used successfully in a variety of projects. For an example, see Lucic and Blake (2012). The output of the typed dependency parser links each negation modifier to the term being negated. We chose to use the Stanford Parser–rather than simply counting the frequency of negation words–in order to maintain this relationship between negation modifiers and the context in which they appear. We plan to take advantage of that contextual information in future research.

The original sentences from the AQUAINT Corpus and the results of the Stanford Parser were loaded into an Oracle SQL Developer (version 3.2.20.09) database. Several unique identifiers were assigned to link the parser output and the original sentences. Finally, the records in the database were queried using Structured Query Language (SQL).

RESULTS

Our analysis consists of several comparisons of the two texts. We considered some general attributes of the texts, including number of articles, average article length, and average dependencies per sentence. We counted the number of discrete instances of negation, as well as the number of distinct sentences and articles that contained one or more negations. Overall, we found that negation occurs approximately three times more frequently in the *New York Times*.

[1] See http://nlp.stanford.edu/software/lex-parser.shtml

Newspaper	Number of sentences	Number of articles
New York Times	16,923,893	313,291
Xinhua News (English Edition)	3,431,663	290,243

Table 1 Total number of records in the data set.

Newspaper	Number of dependencies	Average Dependencies per sentence
New York Times	302,387,123	17.87
Xinhua News (English Edition)	66,121,130	19.27

Overview of the Data

Table 1 shows the total number of sentences and articles available in the AQUAINT Corpus. Despite the availability of a comparable number of articles from each newspaper, considerably more data is available from the *New York Times*. On average, *New York Times* articles are about five times longer than articles from *Xinhua News* (approx. 54 sentences per article vs. 11 sentences per article).

Table 2 shows the total number of typed dependencies identified by the Stanford Parser for each newspaper, as well as the estimated ratio of dependencies per sentence. Again, each typed dependency represents a unique grammatical relationship in the text. As the average numbers of dependencies per sentence for each dataset are almost identical, we can infer that the grammatical complexity of the two texts is comparable. Thus differences in negation usage between the two collections cannot be attributed to sentence complexity.

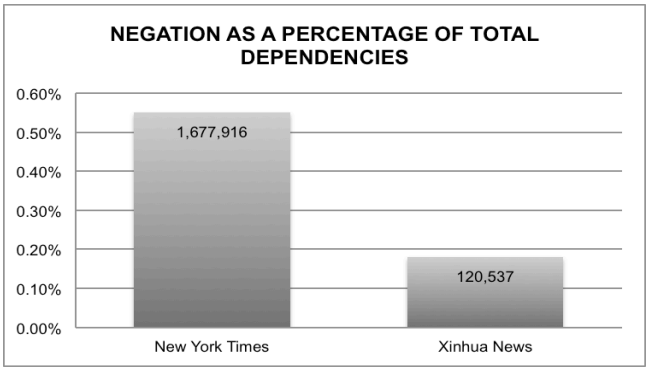

Figure 2 Negation as a percentage of total dependencies.

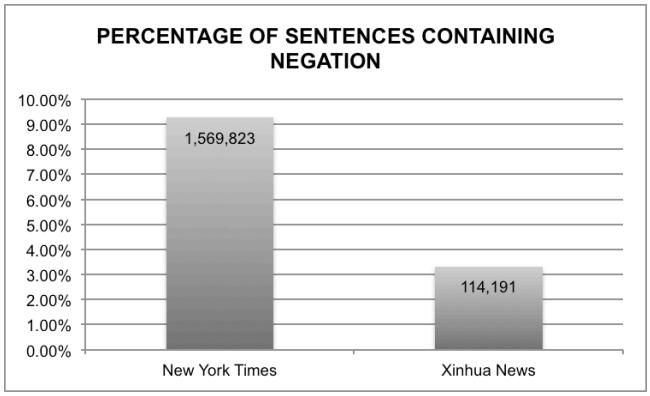

Figure 3 Percentage of sentences containing at least one negation.

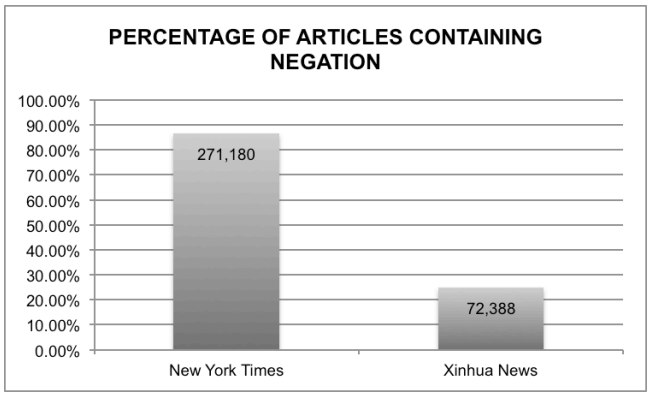

Figure 4 Percentage of articles that contain at least one negation.

Overall Negation Frequency

Figure 2 illustrates the comparative frequency of overall negations between the American and Chinese texts and shows the number of negation typed dependencies as a perctange of total dependencies. Articles in the *New York Times* contain 1,677,916 unique instances of negation (0.55% of total dependencies) and articles in the *Xinhua News* containe 120,537 unique instances of negation (0.18% of total dependencies). In both cases, negation represents less than 1% of all grammatical relationships in the text, which tells us that, in general, negation is not commonly used in news stories. However, these results show that authors of articles in the *New York Times* are roughly 3 times more likely to use negation than *Xinhua News* authors.

Sentences and Articles Containing Negation

Considering negation as a percentage of total typed dependencies captures each unique instance of negation in the text and highlights that negation is a relatively uncommon grammatical phenomenon in this data set. However, considering the number of distinct sentences and articles in which negation appears is also informative. The percentage of sentences from each newspaper containing one or more negations is shown in Figure 3. In the *New York Times* data, 1,569,823 distinct sentences (9.28% of total sentences) contain at least one negation, while 114,191 distinct sentences (3.33% of total sentences) from *Xinhua News* contain at least one negation. At the sentence level, negation is approximately 2.8 times more common in the *New York Times*.

Figure 4 shows the number of articles in each newspaper that contain one or more instances of negation. 86.56% of *New York Times* articles (271,180 distinct articles) and 24.94% of *Xinhua News* articles (72,388 distinct articles) contain at least one instance of negation. Negation is approximately 3.5 times more likely to occur in an article from the *New York Times* than in one from *Xinhua News*. This finding is consistent with the relative frequency of sentences containing negation and total instances of negation in the data. Overall, negation is significantly more common in the *New York Times* than in *Xinhua News*.

CONCLUSION

Our findings are limited to the newspaper texts on which they are based and the texts considered in this study are just a sample of texts published in the United States and China. Another limitation of this research is that it only considers the English-language edition of *Xinhua News*. Additional insights would be gained by analyzing the Chinese-language edition. This edition is not included in the AQUAINT corpus, however a version of the Stanford Dependency Parser for Chinese is available. For more details see Chang et al. (2009).

Despite these limitations, the research methods outlined above can be applied to additional text collections in order to gain a better understanding of the way people from different cultural backgrounds use language. This project is a starting point for future work, including examining (through a combination of close reading and other methods) how negation functions in these texts and in what contexts it appears. Additionally, in light of prior work in Dr. Cohen's lab, we plan to investigate the negation of terms associated with ethical injunctions.

Prior research in the domain of social psychology has shown that speakers of Asian languages are more likely to use negation than English speakers. As such, we anticipated finding more negation in articles from a Chinese newspaper (*Xinhua News* [English Edition]) than from an American newspaper (the *New York Times*). Negation was identified by parsing a large collection of newspaper articles from the AQUAINT Corpus using version 3.2 of the Stanford Natural Language Parser. Negation was shown to be roughly three times more common in the *New York Times*.

The prevalent theory suggests that Chinese authors would use negation more frequently than American authors. Our empirical results show that this theory does not hold in news texts. Further research remains to be done to determine how uses of negation differ in other social contexts (formal vs. informal) or other genres of communication, such as speech, fiction, or social media.

ACKNOWLEDGMENTS

We would like to thank Craig Evans, Senior Research Programmer at the Graduate School of Library & Information Science, University of Illinois at Urbana-Champaign for his assistance with data processing and analysis.

This research is made possible in part by a grant from the U.S. Institute of Museum and Library Services (IMLS), Laura Bush 21st Century Librarian Program Grant Number RE-05-12-0054-12 Socio-technical Data Analytics (SODA).

REFERENCES

Chang, P., Tseng, H., Jurafsky, D. & Manning, C. 2009. Discriminative reordering with Chinese grammatical relations features. *Proceeding SSST '09 Proceedings of the Third Workshop on Syntax and Structure in Statistical Translation*, 51-59. http://nlp.stanford.edu/pubs/ssst09-chang.pdf

De Marneffe, M. C., MacCartney, B., & Manning, C. 2006. Generating typed dependency parses from phrase structure parses. *Proceedings of the LREC*, 6, 449-454. http://nlp.stanford.edu/pubs/LREC06_dependencies.pdf

Jullien, F. (2007). *The impossible nude: Chinese art and western aesthetics*. Chicago: University of Chicago Press.

LDC. (2013). The AQUAINT Corpus of English News Texts. *Linguistic Data Consortium Catalog*. Retrieved from: http://catalog.ldc.upenn.edu/LDC2002T31

Lee, A. Y., Aaker, J. L., & Gardner, W. L. (2000). The pleasures and pains of distinct self-construals: the role of interdependence in regulatory focus. *Journal of personality and social psychology*, 78(6), 1122-1134. doi: 10.1037/0022-3514.78.6.1122

Leung, A. K.-Y., & Cohen, D. (2011). Within- and between-culture variation: Individual differences and the cultural logics of honor, face, and dignity cultures. *Journal of personality and social psychology*, 100, 507–526. doi: 10.1037/a0022151

Llorens, H., Saquete, E. and Navarro-Colorado, B. (2012), Automatic system for identifying and categorizing temporal relations in natural language. *International journal of intelligent systems,* 27: 680–703. doi: 10.1002/int.21542

Lucic, A. and Blake, C. 2012. *Characterizing Authorship Style Using Linguistic Features*. Long paper presented at the Digital Humanities conference, June 16-22, Hamburg, Germany. Retrieved from: http://www.dh2012.uni-hamburg.de/conference/programme/abstracts/characterizing-authorship-style-using-linguistic-features/

Nisbett, R.E. (2003). *The geography of thought: How Asians and Westerners think differently… and why*. New York: Free Press.

Triandis, H.C. (1994). *Culture and social behavior*. New York: McGraw-Hill.