Differential Linguistic Features in U.S. Immigration Newspaper Articles: A Contrastive Corpus Analysis Using the Gramulator

Barbara E. Haertl and Philip M. McCarthy

Department of English & Institute for Intelligent Systems
The University of Memphis
Memphis, TN 38152, USA
{behaertl, pmmccrth}@memphis.edu

Abstract

We use the Gramulator to conduct a computational contrastive analysis on texts printed in states that border Mexico and Canada. The results suggest strong lexical differences between the regions. The primary factor in the southern border states is *security*, whereas factors in the northern border states are *citizenship* and *family*. The results reveal differing regional perspectives, offering insight as to how reporting conveys messages to different audiences.

Introduction

Immigration is a national issue in the United States; however, regional implications differ because of immigrants' varying effects on local economies. These implications are made manifest in the reportage of local newspapers, which, while ostensibly portraying "objective" language, may reveal the narrative of local perspectives on national issues.

Linguistic research on immigration is growing, but currently limited. For instance, Branton and Dunaway published several studies (e.g., 2009) on how proximity to the U.S.–Mexico border influences newspaper coverage. Their analyses focused on the volume and polarity of Californian newspaper articles. Burnett (2009) studied differences in reporting on municipal immigration ordinances. She found that a newspaper in Texas contained more attributes of the ordinance's opposition, whereas a newspaper in Pennsylvania focused mostly on the crime attribute. The importance of the immigration issue suggests that extensive research is needed from a greater breadth of sources and that sophisticated textual analysis tools are needed to reveal finer-grained details on the variations of language used.

The goal of our study is to determine whether the language of newspaper articles printed in southern and northern border states reflects linguistic differences in reporting on immigration. Through this study, we seek to better understand the lexical differences in language used to report on sensitive issues, such as immigration. This study is of interest to journalists, politicians, and sociologists. It is of particular interest to the analyses of objective reporting on contentious issues: offering insights as to which language is selected to convey which kind of messages to an intended audience.

Corpus

Our corpus comprises 752 texts, culled from newspapers of U.S. border states (approximately 75 texts per state). Because four states border Mexico, we selected four matching states (of the 11) that border Canada. To do so, we considered the following criteria for all 15 terrestrial border states: *total population*, *immigrant population*, *length of international border*, and *political leaning*. These data were input into a custom PERL script designed to match each Mexican border state with just one of the 11 Canadian border states. With each criterion equally weighted, the resulting matches were Arizona–Washington; Texas–Michigan; California–New York; and New Mexico–Minnesota. Although these pairings are based on limited criteria, they form a reasonable point of departure for the purposes of this study.

The first corpus (*southern border, SB*) comprises 417 texts, collected from newspapers in the four U.S. states that border Mexico. The second corpus (*northern border, NB*) comprises 335 texts, collected from newspapers in the four selected states that have terrestrial borders with Canada.

Representative state newspapers were selected according to circulation and availability of online articles. From each newspaper's Internet site, we searched for instances of the words *immigration* or *immigrant* in printed copy (no blogs or other online-only articles). All texts were from 2010 to reflect a representative zeitgeist. We did not consider texts from national news wires (e.g., AP, Reuters) because we could not ensure that such texts suitably reflected the state. Fact-based articles, editorials, and letters to the editor were all considered. We cleansed texts for author names, dates, and other aspects not relevant to the story.

The Tool: The Gramulator

Sophisticated textual analysis tools such as Coh-Metrix (Graesser et al. 2004) have revealed many rich and varied findings, but despite their success, tools that are based on built-in psychological and readability measures struggle to reveal the lexical characteristics that distinguish contrastive corpora (i.e., two highly related corpora). Thus, to conduct a computational contrastive corpus analysis, we use the textual analysis tool, the Gramulator (McCarthy, Watanabe, and Lamkin in press). The Gramulator's purpose is to help in the identification of linguistic features of correlative texts. The Gramulator accomplishes this feat by identifying *n*-grams that are typical of one corpus but untypical of the other. We refer to such a collection as an index, and examples from an index as differentials. In the current study, the index of SB relative to NB is denoted SB (NB). Similarly, NB (SB) is an index representing what is typical of NB while being untypical of SB. While the indices can be used to measure texts (much like Coh-Metrix), researchers can also use theory to organize the differentials into conceptual lexical features, which may in turn be used as an index.

Results

Using the Gramulator's *Sorter* module, two-thirds of the texts in each corpus were randomly placed into *training sets*. These training sets were analyzed by the Gramulator's main module to create indicative indices of each corpus: SB (NB) and NB (SB). The remaining one-third of each corpus was placed into *test sets*.

We conducted *t*-tests to validate the training-set–derived indices: SB_{train} (NB_{train}), NB_{train} (SB_{train}). Using the test-set data, the result for SB_{train} (NB_{train}) was in the predicted direction ($\mathbf{SB}_{test}$: M = .076, SD = .035; $\mathbf{NB}_{test}$: M = .060, SD = .034). The result reached a level of significance $t(1,220) = 3.597$, $p < .001$, $d = .483$. A similar result was found for NB_{train} (SB_{train}) ($\mathbf{NB}_{test}$: M = .065, SD = .048; $\mathbf{SB}_{test}$: M = .049, SD = .030). The result also reached a level of significance $t(1,246) = 3.227$, $p < .001$, $d = .412$. The results validate the indicative indices.

Table 1. Significance of bigrams using Fisher's exact test.

Southern Border		Northern Border	
p	Bigram	*p*	Bigram
< .001	border patrol	< .001	was born
< .001	border security	.007	citizenship for
.002	police department	.008	grew up
.002	patrol agents	.017	to citizenship
.004	border is	.035	green card
.008	an officer	.004	his mother
.016	public safety	.006	children and
.016	police and	.014	mother and
.016	of border	.023	wife and
.042	border in	.029	their families
		.029	a child

The training sets produced 442 individual differential bigrams in the SB corpus and 200 differential bigrams in the NB corpus. We used Fisher's exact test to determine which of these bigrams were statistically significant (see Table 1, above). Two unigrams were also found to be significant in the NB corpus: mother (p = .031) and citizenship (p = .048). All of these *n*-grams were then confirmed to be significant in the test-set data. The primary unifying factor in the SB corpus was the concept *border security*, whereas the primary factors in the NB corpus were the concepts of *citizenship* and *family*.

Discussion

This study provides evidence that the lexical content of immigration articles significantly differs between southern and northern border states. Specifically, the southern-border corpus has a stronger focus on border security, whereas the northern-border corpus has a stronger focus on citizenship and family. Future research will include comparing paired states and comparing articles, editorials, and letters to the editor within their respective groups. The Gramulator's ability to conduct a computational contrastive corpus analysis has helped to reveal different regional perspectives on immigration. Identifying the lexical features of these contrastive corpora is important because local opinions can be shaped by the media's influence. By focusing more on citizenship and family, northern-border newspapers limit their role in fostering the recent national commotion over border security. Thus, this initial study serves as a small but important step toward a better understanding of how the language of reporting conveys messages to an intended audience.

Acknowledgements

We thank Andrew Schick for the state-pairing program.

References

Branton, R. P., and Dunaway, J. 2009. Spatial Proximity to the U.S.–Mexico Border and Newspaper Coverage of Immigration Issues. *Political Research Quarterly* 62(2):289–302.

Burnett, A. L. 2009. Municipal Immigration Ordinances: An Analysis of Newspaper Coverage of a Controversial Issue. Master's Thesis, The University of Texas at Arlington.

Graesser, A. C.; McNamara, D. S.; Louwerse, M. M.; and Cai, Z. 2004. Coh-Metrix: Analysis of Text on Cohesion and Language. *Behavior Research Methods, Instruments, & Computers* 36(2): 193–202.

McCarthy, P. M.; Watanabe S.; and Lamkin, T. A. in press. The Gramulator: A Tool for the Identification of Indicative Linguistic Features. In P. M. McCarthy and C. Boonthum (Eds.), *Applied Natural Language Processing and Content Analysis: Identification, Investigation, and Resolution*. Hershey, PA: IGI Global.