Genre in Semantic Networks: A study of the Lexicon of News Articles

Yiwei Luo and Jose Krause Perin
{yiweil, jkperin}@stanford.edu

December 11, 2017

1 Introduction

Our project aims at understanding text genres within the domain of the news. Advances in computational methods and availability of digital corpora has ushered in a new age of empirically testing intuitions about genres and styles, in particular the automatic classification of a document to its genre. At the same time, identifying systematic patterns of difference between genres, both quantitative and qualitative, can bring insight to our understanding of discourse from a linguistic perspective, and perhaps the social motivations that underlie our speech patterns as well. From a more applied perspective, discovering the key differences between genres can be of enormous use to natural language processing (NLP) tasks that work with corpora from a range of genres by identifying biases associated with particular genres or potential artifacts of genre imbalance in datasets. Other tasks such as Part-of-Speech (POS) tagging, parsing, word sense disambiguation (WSD), and information retrieval can also be significantly improved when taking genre into account.

We study the semantic content and lexicon of individual news genres by mapping New York Times news data that is organized into various editorial "desks" (as categorized *a priori* by the newspaper) onto two types of network representations: one based in collocations of word forms and the other based on the hierarchical WordNet semantic graph. Such network representations allow us to investigate how documents behave not only with respect to their content, but the structure of their content as well. In particular, we discover the most distinctive lexical content of a news desk through the collocation network and how specialized a desk is through WordNet. Our use of network-based analyses provides readily interpretable answers, unlike other methods such as Latent Semantic Analysis or Topic Modeling.

The remainder of this paper is organized as follows: Section 2 gives an overview of prior NLP work relating to text genre. Section 3 details the data preprocessing and Section 4 describes how the data are represented as co-occurrence networks and WordNet subgraphs, respectively. Section 5 describes our evaluation metrics, Section 6 presents our findings, and Section 7 outlines directions for future work.

2 Relevant prior work

Though there has been much computational work in the realm of music and movie genres, for example, genre as a textual property has

been relatively unexplored and little work has been done on topic-constituted genres within a specific domain, such as the various editorial desks in the domain of the news.

Past NLP approaches to genre detection and characterization have focused on authorship attribution [1], style markers [1], [2], and genre classification at the more macro-level to labels such as research article, novel, poem, advertisement, court decision etc [3], or classification of web-pages into genres such as help/FAQ, news, link collection, and product info [4]. Moreover, the dominant approach to exploring genre has been to abstract away the propositional content of the text: in the words of Eissen and Stein[5], "[...] genre classification is orthogonal to a classification based on the documents contents." Pavlick and Nenkova [6] also note that "Most work on stylistic variation, however, has focused on larger units of text [...] and studies of style at the lexical level have been scant." The work of the latter analyzes statistical differences in specific lexical items across phrases, but their focus is on the task of inducing the formality/complexity of a text's style as opposed to its topic or genre-based properties.

And while there has been much work that leverages the WordNet hierarchy to compute how similar two word meanings are based on their distance in the WordNet graph (Pederson et al., 2004 [7], Richardson et al., 1994 [8], Varelas et al., 2005 [9]), but no work bringing together genre and inter-lexicon similarity, so this project makes a novel contribution in formalizing genre-specialization as an average of similarities across the genre's lexicon.

3 Data preprocessing

3.1 Data set

Our data set consists of the New York Times Annotated Corpus [10], which contains over 1.8 million articles from January 1, 1987 to June 19, 2007. For the first part of this project, we focused on semantic and lexical differences across editorial desks of the New York Times. There are altogether 52 different editorial desks, a few examples are Job Market, Sports, Science, Museums, and Travel. Naturally, there will be lexical differences across desks, since, for example, the vocabulary used in sports articles can be very different from science related articles. We will quantify these differences to measure which desks are more closely related to each other.

As it is typical in NLP problems, the data set has to be "normalized", which consists of removing words that participate in binding the text together, but do not carry much meaning. These so-called stopwords are articles, conjunctions, auxiliary verbs, and some frequently used words (e.g., about, might, indeed) that do not affect the context.

In addition to removing stopwords, the meaningful words of each article are lemmatized using the NLTK WordNet Lemmatizer (with additional Part-of-Speech tagging and tokenization steps) in order to collapse different morphological realizations of a word onto its shared base form.

As an example, the following excerpt of the article [11]

> "Representative John D. Dingell, a Michigan Democrat who with more

than 50 years' tenure is the senior member of the House, is not so sure about the idea of creating an independent group to enforce ethics rules,"

would reduce to the following lemmas

"representative john dingell michigan democrat year tenure senior member house sure idea create independent group enforce ethic rule."

Clearly, the lemmatized version of the excerpt preserves the meaning of the original text while making it more structured.

4 Network representations

We followed two different methodologies to represent the lemmatized text as a graph: co-occurrence statistics and WordNet mapping. These methodologies are detailed in the next subsections.

4.1 Co-occurrence statistics

Co-occurrence representation relies on the so-called "distributional hypothesis" that states that word semantics are implicit in the context. In other words, it is always possible to extract the meaning of a lemma from the context i.e., from its neighboring lemmas.

Based on this hypothesis, the lemmas are the nodes of the network. If two lemmas are neighbors in a sentence, they are connected by an undirected edge. The weight of this edge is equal to how often these two lemmas appear together regardless of order.

The co-occurrence matrix was computed using the method `CountVectorizer` from the NLP library of `sklearn`. The co-occurrence matrix is a sparse and symmetric matrix where its ij entry is the number of instances lemmas i and j were neighbors in a sentence. Thus, it is straightforward to build an undirected graph from the co-occurrence matrix.

4.2 WordNet

For the WordNet data, we use the English WordNet, version 3.0 [12], which contains 117,798 unique nouns belonging to 82,115 synsets, 11,529 unique verbs belonging to 13,767 synsets, 21,479 adjectives (18,156 synsets) and 4,481 adverbs (3,621 synsets). Many of these words have multiple senses and accordingly belong to multiple synsets.

A primary motivation for using WordNet is that a word's semantics, that is, the actual meaning of a word, can be represented. We use the Lesk algorithm from NLTK to automatically disambiguate a word form's meaning in context and label it with a WordNet synset. The co-occurrence approach does not apply word sense disambiguation (WSD) in creating the co-occurrence matrices; thus distinct senses of a homophonous form (for example, 'run') are not differentiated. After collecting the WordNet synsets that each article's collection of lemmas maps onto, we can perform graph analyses where each synset is a node of the WordNet semantic graph. Due to the architecture of WordNet, we consider

only the noun and verb lemmas in a desk's articles.

5 Evaluation metrics

5.1 Node degree and page rank

To evaluate the importance of lemmas in co-occurrence networks we use metrics such as node degree and page rank score, as discussed in class. Intuitively, in co-occurrence networks, high-degree nodes correspond to lemmas used frequently regardless of context. Conversely, low-degree nodes correspond to infrequent lemmas or lemmas with very specific meaning or function.

Similarly, a node page rank indicates the relevance of that lemma in a particular network representation. As shown in Section 6, we have observed that page rank score and node degree are strongly correlated in co-occurrence networks.

5.2 Evaluating specialization in WordNet representations

We use three similarity metrics taken from Pederson et al. (2004) [7] to compute an aggregate score for the mean distance of a synset to all other synsets in the article. We then averaged this score across all articles in a given desk to obtain a measure of a desk's level of specialization in the WordNet tree. Intuitively, more specialized desks consist mostly of lexical which are more similar to each other. The similarity metrics are described below.

5.2.1 Path similarity P_s:

$$P_s(u, v) = \frac{1}{l_s}, \tag{1}$$

where l_s is the length of the shortest path connecting node u and node v. The value of P_s ranges from 0 to 1 and a higher number indicates a higher degree of similarity (with 1 indicating identity of u and v).

5.2.2 Leacock-Chodorow Similarity[13] L_s:

$$L_s(u, v) = -\log\left(\frac{l_s}{2d}\right), \tag{2}$$

where l_s is again the shortest path length between u and v and d is the maximum depth of their least-common-hypernym.

5.2.3 Wu-Palmer Similarity[14] W_s:

$$W_s = \frac{2d}{d_u + d_v}, \tag{3}$$

where d is the depth of the least common hypernym of u and v and d_u, d_v are the depths of u and v, respectively.

5.2.4 Mean distance from root

Finally, we also compute the mean distance of all synsets in an article from the root synset, "entity." We expect that more specialized articles consist of synsets that are further from the root node. We use the WordNet function for path similarity included in the NLTK interface.

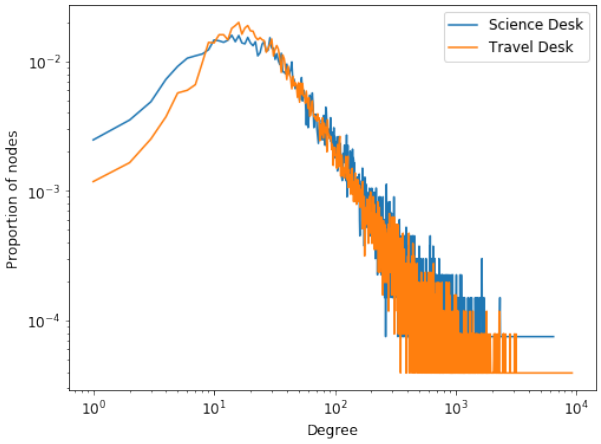

Figure 1: Normalized degree distribution of co-occurrence networks of Travel and Science Desks.

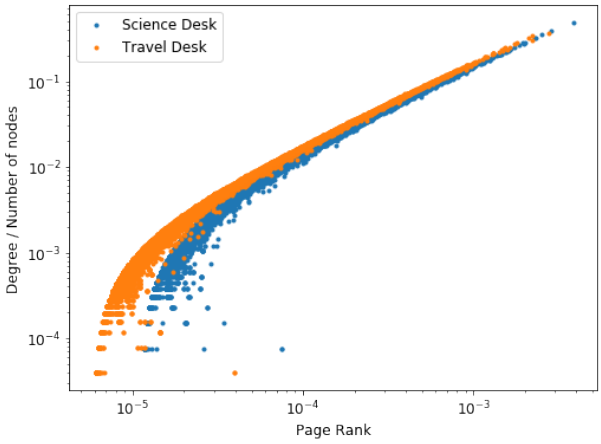

Figure 2: Normalized degree distribution of co-occurrence networks as a function of node page rank.

6 Findings

For the examples of this section we use articles from the Science and Travel desks of the New York Times in 2007. The analysis presented in this section can be readily extended to include other desks in the comparison.

Figure 1 shows the normalized degree distribution of co-occurrence networks of Travel and Science desks. The degree distributions of both networks exhibit the same behavior, which was verified in other desks as well. The first part of the graph corresponds to the small-degree nodes, which are mainly formed by uncommon words or artifacts that were not removed during data preprocessing. In the high-degree part of the curve (degree > 30), the degree distribution follows a power decay with approximately the same slope. The same slope was also verified in other desks. High-degree nodes represent commonly used words such as ('say', 'one', 'new', etc), which appear in most texts regardless of genre or content. However, there are also high-degree nodes that are exclusive of each desk. These high-degree nodes delineate the specific vocabulary used in a certain desk.

The co-occurrence networks exhibit a strong correlation between node degree and node page rank, as shown in Figure 2. Nodes with high degree have high page rank scores with very small variability, which results in the relatively thin lines in Figure 2. The slope in the log-log plot is the similar regardless of the desk. Noticeable differences between science and travel desks occur only for low-degree nodes. However, since these nodes have small page rank score, they cannot be used to draw meaningful conclusions. Hence, in the following subsections we focus on high-degree nodes and how they vary across desks.

Table 1: Lemmas with highest degree in each desk.

Science Desk		Travel Desk	
Lemma	Degree	Lemma	Degree
say	6436	www	9174
dr	5073	like	8498
one	4622	one	8144
year	4220	com	7657
like	4214	new	7148
make	3713	city	6852
new	3694	hotel	6302
university	3646	year	6187
time	3522	make	6026
would	3287	include	6010

Table 2: High-degree lemmas exclusive of each desk.

Science Desk		Travel Desk	
Lemma	Degree	Lemma	Degree
physicist	1034	euro	3446
genetic	890	cafe	1873
physic	860	boutique	1789
biology	760	chef	1769
chimp	650	theater	1687
analyze	602	tea	1622
polar	511	spa	1593
chimpanzee	499	grill	1490
atom	468	contemporary	1483
cern	466	lounge	1437

Table 3: Most common lemmas that co-occur with the lemma 'university'.

Science Desk		Travel Desk	
Lemma	Freq.	Lemma	Freq.
say	166	www	31
dr	92	film	24
study	67	new	22
science	66	street	22
professor	57	com	20
colleague	52	festival	17
researcher	52	one	13
year	47	student	13
make	43	york	13
one	43	april	12

In Section 6.1 we focus on vocabulary variations across desks, and in Section 6.2 we focus on word meaning variations across desks. In Section-6.3 we give the results of meaning specialization across desks.

6.1 Vocabulary variations across desks

As a first experiment, we compared vocabulary differences across desks. This is realized by looking at high-degree nodes that are not shared by the other desk. Table 1 shows the nodes with highest degree in each network regardless of whether they appear in the other network. Table 2 shows high-degree lemmas that are exclusive of each desk. Table 2 suggests a clear distinction of vocabulary between the Science and Travel desks.

6.2 Variations of word meaning across desks

In this second experiment we turn our attention to how a given lemma may exhibit different meaning in different desks. For a given lemma that is shared by both networks, we select its neighboring nodes (lem-

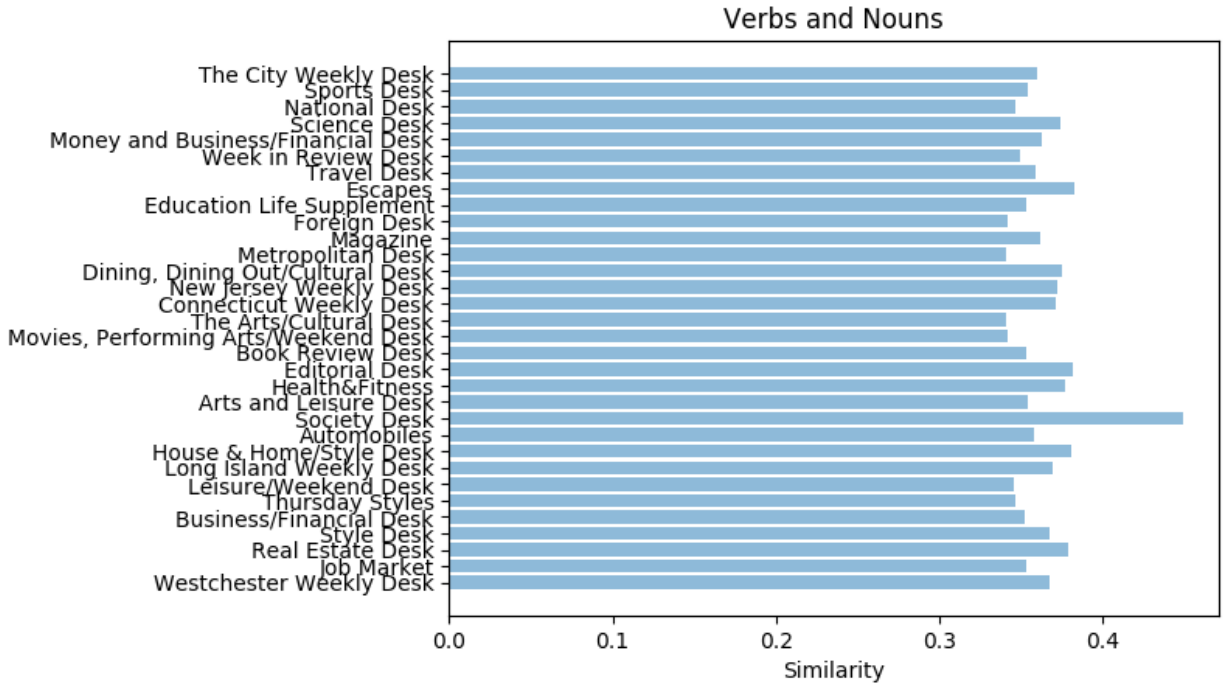

Figure 3: Average path similarity of verbs and nouns in each desk.

mas) with the highest edge weight. Recall that in the co-occurrence network, each edge has a weight that is equal to how often the two words appear together in a sentence. A high-weight edge indicates a commonly used term in a given field. Given the different nature of topics cover by the Science and Travel desks, it is natural to expect that certain words will carry very distinct meanings in each desk. As a example, Table 3 shows the most common lemmas that co-occur with the lemma 'university'.

Note that in the Science desk the lemma 'university' has the meaning of institution, while in the Travel desk the same lemma denotes a reference point or a touristic spot.

6.3 Desk specialization

Using the similarity metrics discussed in Section 5, we find a similarity score for each article by computing the average of the similarity for every pair of nouns and verbs. We then average across desks, summing the similarities of verbs and nouns. The results of the average path similarities by desk are shown in Figure 2. (Wu-Palmer similarity and Leacock-Chodorow similarity yield similar results and are shown in the Appendix.)

Though all desks have similarities that fall within the range of 0.3 and 0.4 (with Society Desk being an outlier), the differences in similarity scores are all significant with $p-$value

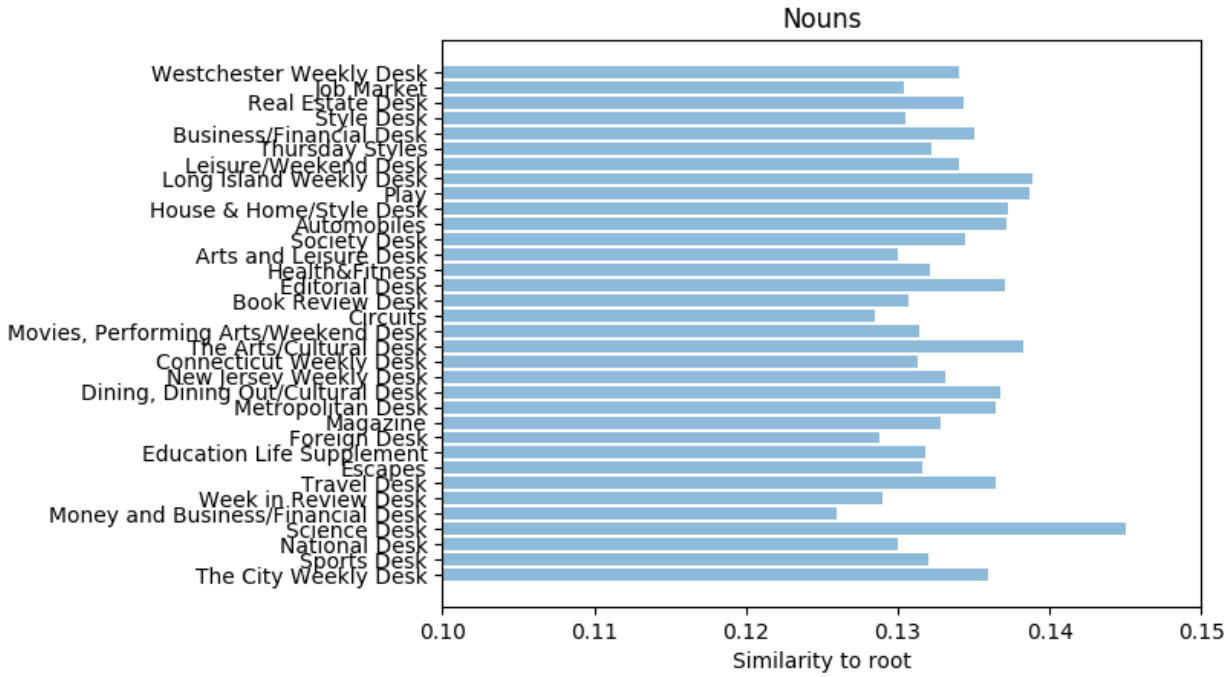

Figure 4: Average similarity of nouns in each desk to root.

< 0.5 on pair-wise t-tests.

It should not come as a surprise that Society Desk has the greatest similarity score– this desk consists of wedding announcements, so the lexicon is more limited, formulaic and specialized. The desks with the lowest similarity scores are Foreign Desk, Metropolitan Desk, and The Arts/Cultural Desk, which suggest that these desks have the lowest degree of specialization, in line with our intuitions given the size of their audiences and the breadth of their topics.

The results of desk similarity with root are shown in Figure 4. All similarities are significantly different with $p < 0.5$ on two-sided t-tests. Interestingly, the desk with the highest average score for similarity with root was the Science desk, and the top 5 articles ranking highest on this score came from the Science desk as well. Our spot-checks of the text of these articles suggests to us that scientific writing in the news is quite different from in a research journal, and technical terms are avoided in favor of more "laymen" paraphrases. Another reason for the high score of the Science desk may be that there are many direct or closely related hyponyms from the root 'entity' such as 'organism', 'matter', 'substance' and 'set', signaling the abstractness of the scientific genre.

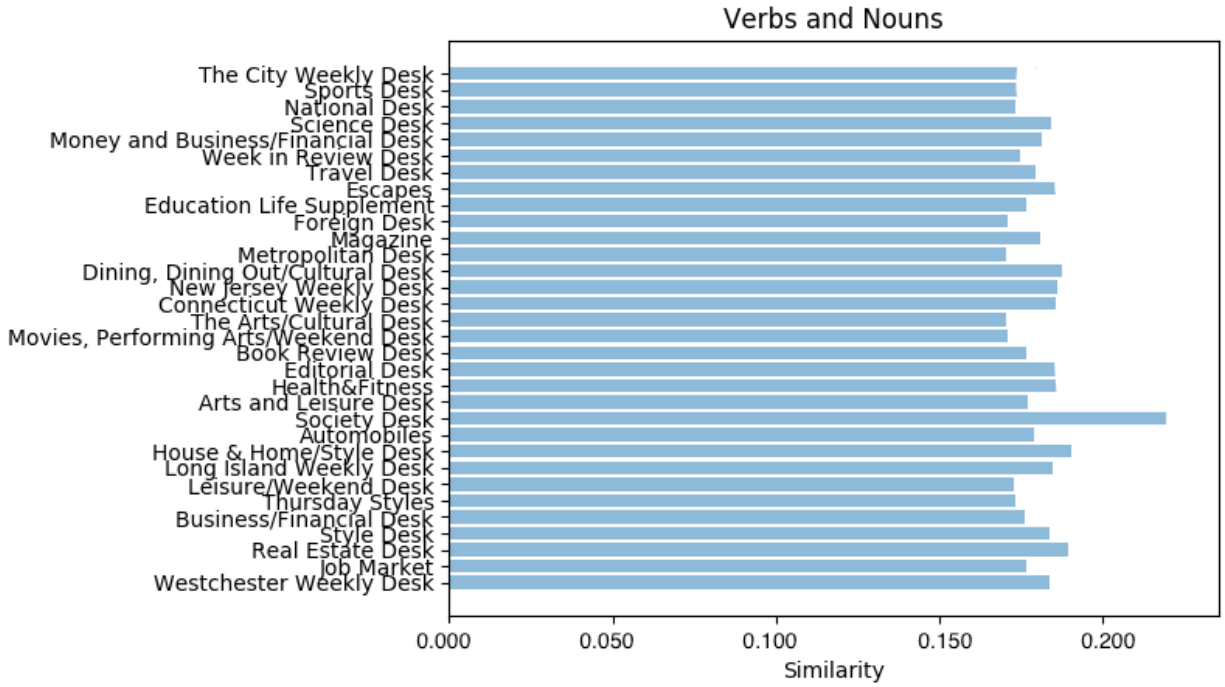

Figure 5: Average Wu-Palmer similarity of verbs and nouns in each desk.

7 Conclusion and future work

The goal of this project was to see if the graph structure of a text varies systematically with its genre. Using news articles as texts and news "desks" as genres, we focused on the graph properties of degree distribution, pagerank, and node similarity and found that there are significant correlations.

Specifically, by using the WordNet representation of an article's semantic graph and building on the intuition that more specialized texts contain lexical items that are closer to each other in the graph and farther from the root node, we find that certain genres are more specialized than others and seem to align with our intuitions of what a genre is about.

In the future, it would be interesting to leverage co-occurrence and semantic network features in automatic classification as a way of improving such tasks.

Future work may also explore how named entities can be judiciously incorporated into co-occurrence as well as WordNet representations. For example, in certain cases, named entities are a key ingredient to a desk's level of specialization (e.g., "Miuccia Prada" in the Style desk), but in other cases, the name of the individual is incidental to the desk's subject matter.

Appendix

Figure 5 shows the averaged Wu-Palmer similarities across desks.

Contributions

Yiwei: Analysis and results using WordNet.
Jose: Analysis and results using coocurrence networks.

References

[1] E. Stamatatos, N. Fakotakis, and G. Kokkinakis, "Automatic text categorization in terms of genre and author," *Computational linguistics*, vol. 26, no. 4, pp. 471–495, 2000.

[2] F. Khosmood and R. A. Levinson, "Automatic natural language style classification and transformation," in *BCS Corpus Profiling Workshop, London, UK*, sn, 2008.

[3] B. Kessler, G. Numberg, and H. Schütze, "Automatic detection of text genre," in *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, pp. 32–38, Association for Computational Linguistics, 1997.

[4] N. Dewdney, C. VanEss-Dykema, and R. MacMillan, "The form is the substance: Classification of genres in text," in *Proceedings of the workshop on Human Language Technology and Knowledge Management-Volume 2001*, p. 7, Association for Computational Linguistics, 2001.

[5] S. M. Zu Eissen and B. Stein, "Genre classification of web pages..," in *KI*, pp. 256–269, Springer, 2004.

[6] E. Pavlick and A. Nenkova, "Inducing lexical style properties for paraphrase and genre differentiation.," in *HLT-NAACL*, pp. 218–224, 2015.

[7] T. Pedersen, S. Patwardhan, and J. Michelizzi, "Wordnet:: Similarity: measuring the relatedness of concepts," in *Demonstration papers at HLT-NAACL 2004*, pp. 38–41, Association for Computational Linguistics, 2004.

[8] R. Richardson, A. F. Smeaton, and J. Murphy, "Using wordnet as a knowledge base for measuring semantic similarity between words," in *Proceedings of AICS conference*, pp. 1–15, 1994.

[9] G. Varelas, E. Voutsakis, P. Raftopoulou, E. G. Petrakis, and E. E. Milios, "Semantic similarity methods in wordnet and their application to information retrieval on the web," in *Proceedings of the 7th annual ACM international workshop on Web information and data management*, pp. 10–16, ACM, 2005.

[10] E. Sandhaus, "The new york times annotated corpus," *Linguistic Data Consortium, Philadelphia*, vol. 6, no. 12, p. e26752, 2008.

[11] C. Hulse, "As new congress nears, house democrats could be headed for own divide," January 2007. Online.

[12] G. Miller, C. Fellbaum, R. Tengi, P. Wakefield, H. Langone, and B. Haskell, *WordNet*. MIT Press Cambridge, 1998.

[13] C. Leacock and M. Chodorow, "Combining local context and wordnet similarity for word sense identification," *WordNet: An electronic lexical database*, vol. 49, no. 2, pp. 265–283, 1998.

[14] Z. Wu and M. Palmer, "Verbs semantics and lexical selection," in *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, pp. 133–138, Association for Computational Linguistics, 1994.