ABSTRACT
Discourse analysis and natural language processing appear to share fundamental priorities: the detailed and systematic examination of text. However, authors have argued that discourse analysis does not have a strong tradition in the field of applied natural language processing, mainly because of the computational challenges associated with the nature and characteristics of discursive resources [5]. In this study, I assess the current affordances and limitations of a specific set of NLP tools in supporting discourse analysis, in order to identify possible areas for further research and development. I conduct manual discourse analysis on a pool of 15 newspaper articles and identify recurring discursive strategies. I then explore the extent to which these strategies are supported by Stanford CoreNLP [11], a suite of tools that is considered to be foundational and comprehensive. My findings show that of the 12 recurring discursive strategies, four are fully supported by this tool; five are partly supported; and three are not. I conclude that many of the dimensions that are not addressed can nevertheless be supported by a combination of machine learning techniques and by other NLP-based resources, for example those that draw on rhetorical structure theory (RST) [9]. I conclude by suggesting areas for future exploration.

Categories and Subject Descriptors
1.2.7 [Natural Language Processing]: Discourse.

General Terms
Human Factors, Languages.

Keywords
Discourse, discourse analysis, natural language processing.

1. DISCOURSE ANALYSIS AND NATURAL LANGUAGE PROCESSING
Discourse analysis is a methodology in the social sciences concerned with exploring how social reality is “created, supported, and contested through the production, dissemination, and consumption of texts” [4]. The notion of “texts” has been broadly defined as encompassing a wide variety of social products such as visuals, practices, and material works such as architecture and comic strips, but has often been understood to refer to language-based social products. Discourse analysis is linked to social constructionist assumptions. It assumes that the categories by which we understand and organize the social world are not objective, natural, permanent, or inevitable, but subjectively created and precarious in that they can be recreated or changed over time [3]. This is significant, because it is assumed that discourses (and correspondingly, the categories and meanings that they are associated with) are seen to have concrete implications on reality. For example, discourses on efficiency that emerged during the industrial revolution continue to fundamentally shape strategic and day-to-day practices not just business institutions [10], but also educational and health care institutions. Drawing on this observation, a change in discourse can arguably lead to changes in specific dimensions of social reality.

An important dimension in doing discourse analysis is understanding discursive strategies, tools or techniques that people consciously or unconsciously mobilize in their production of texts. Studies have shown that there is a diverse range of discursive strategies that can be mobilized in different social situations [2], in order to achieve specific outcomes. A discursive strategy involving the use of active (versus passive) verbs, for example, can have the impact of making a subject more powerful, whether in protagonist or antagonist capacity. A discursive strategy involving “naming” subjects in different ways (“John Smith, associate professor and political analyst” versus “John Smith, recovering alcoholic”) will be associated with different outcomes and hence will be mobilized for different purposes. The use of modifiers like verbs and adverbs can also construct associated persons, places, objects, and events in different ways.

Research activities that harness discourse analysis as a methodology can be challenging in that they can involve large volumes of text. A question that arises, therefore, as to how (if at all) technological tools can be harnessed to aid the process of discourse analysis.

In the field of computer science, one area that has been linked to discourse analysis is the field of natural language processing (NLP). Kent and McCarthy [5] have explored the relationship between the two domains. They posit that while discourse analysis and NLP appear to be natural extensions of one another (because of the seemingly strong relations between “language” and “discourse” as well as between “processing” and “analysis”), they nevertheless argue that the field of NLP does not have a strong tradition in discourse analysis. They attribute this to the nature of discourse, specifically how challenging it is to capture computationally. Unlike content analysis, which can be characterized by counting the occurrences of specific terms [3], discourse analysis is concerned with heavily qualitative interests, including why something was said this way instead of that way; what was not said; and why it was not said.

It is important to note, however, that while NLP tools and techniques cannot do “everything” that discourse analysts want, there are still many things that these tools and techniques can facilitate. This paper therefore seeks to address the research
question, “What are the affordances and limitations of existing NLP tools and techniques in supporting discourse analysis?”

2. METHODS

To address the research question, I performed manual discourse analysis on a predefined pool of text in order to identify recurring discursive strategies, and proceeded to assess the extent to which these strategies could be supported by a collection of NLP tools. The pool of text consisted of 15 newspaper articles on a single news event. Newspaper articles have been understood to have powerful and significant effects on society, and their close examination has therefore been the subject of much research over a long period [1]. In this case, the news event involved the death of a seven-year-old female due to a celebratory stray bullet on December 31, 2012.

Data analysis was initiated by first examining a single article and identifying discursive strategies that were mobilized within it. Examples of discursive strategies include the use of active verbs; the mobilization of sources through direct quotes from people in authority; and topic-shifting. Analysis was then performed on succeeding articles, hence a list of discursive strategies was constructed iteratively. A total of 12 discursive strategies were identified. Once the complete set of strategies had stabilized, these strategies were then classified using a scheme proposed by Richardson [7].

Once this categorization was completed, I then explored the extent to which these strategies could be supported by a suite of tools called Stanford CoreNLP [11], an open source software that covers a broad range of NLP tools (identifying the structural parts of sentences, identifying entities alluded to by proper nouns, pointing out parts of speech, etc). It is mobilized as a test tool here because it is comprehensive (in that it offers an array of functions) as well as foundational (that is, higher level analyses can build on capabilities captured by Stanford CoreNLP) [1].

3. FINDINGS

Having identified a set of discursive strategies, I then used Richardson’s [7] scheme to cluster related strategies together. I analyzed that the first set consisted of six strategies that were all associated with the way that propositions were structured: word choices for people and events; the mobilization of predicates; transitivity choices (who does what to whom); modality choices (the extent to which opinion or judgments were being mobilized); and the presence or absence of presuppositions. These are presented, with examples, in the first two columns of Table 1. The second set consisted of six strategies that were all associated with the way that propositions were combined. Most of these strategies were associated with shaping arguments: causes leading to effects; problems leading to solutions; single events leading to generalizations. This is not surprising, since journalism has been understood by some to be an argumentative discourse genre [8]. One strategy had to do with topics shifting through a set of propositions. It is also clear, therefore, that Table 1 is linked to micro-textual analysis, while Table 2 is concerned with macro-textual analysis.

It is important to note that the list of discursive strategies discussed here is by no means comprehensive; Gee [2], for example, presents a total of 27 tools for doing discourse analysis. Nevertheless, examining the nature of these strategies discerned in this study is a useful starting point given that their characteristics might be shared with other strategies not considered here, hence lessons about their existing NLP support might be generalizable.

It is tempting to conclude that micro-level discourse strategies should generally be “easier” to support computationally than macro-level discourse strategies. Based on this study, this is true only in the sense that there are more strategies that are fully or partly supported by NLP tools in Table 1. It cannot be taken to mean, though, that all micro-strategies are supported. A discursive strategy involving implicit presuppositions, for example, is a micro-strategy involving “hidden” meanings that cannot be directly traced to sentence form and content, and hence has been analyzed to have no current NLP support. Instead, a macro-strategy like topic shifting can be supported by basic NLP tools. Details on how NLP tools can be mobilized to support each strategy can be seen in column 3 of Tables 1 and 2.

3.1 Readily-supported discursive strategies

It appears that tools like Stanford CoreNLP [11] can provide robust support in cases where there is a close link between form and content in a proposition – that is, cases wherein “the meaning of what is written is related to the way that it is written” [7]. A lesson learned here is that if a discursive strategy can be discerned by parsing a sentence into identifiable structural elements (hence enabling one to use one element as a reference point for another), or if a strategy can be signaled by key words, then it can readily be supported by Stanford CoreNLP. For instance, the use of modifiers, an example of the “predication” strategy [C], involves mobilizing adjectives, adverbs, or other nouns in order to characterize a subject in specific ways. This discursive strategy can readily be discerned simply by identifying subjects in a sentence and setting up the tool to identify all modifiers for the subject. The strategy of naming sources by using formal titles [A] can readily be supported by establishing key words. The mobilization of one’s opinion and judgment [E] can likewise be tracked through words like “should”, “ought to”, and “must”. A shift in topic [G] can be established by identifying the focal center of a proposition and seeing if it is pronominalized in succeeding propositions. In summary, therefore, Discursive Strategies [A], [C], [E], and [G] are readily supported by Stanford CoreNLP [11].

3.2 Partly-supported and unsupported discursive strategies; possible alternatives

The process of supporting discursive strategies becomes more challenging when key words (cues) are not mobilized or are replaced by other words, or when a proposition (or set of propositions) structure is so complex that tools are unable to use one element as a reference to another element.

The issue of transitivity [D], for example, includes, but is not limited to, the issue of whether one is using an active or passive verb. This choice can readily be discerned. However, the use of active voice is often just one of a family of strategies meant to attribute increased agency to a subject in ways such that s/he is seen as “accomplishing something”. Alternative strategies may be much more subtle and hence not be so clearly signaled. In the sample texts given, this is carried out by using phrases like authorities “started inspecting” or an officer having “narrow[ed] down suspect list from 45 to 32.” In both cases, both segments talk about how the actor seems to have completed something.
even though a close examination shows that this something is just a part of a larger task (the actor “started inspecting” but has not finished inspecting). This is considered an insight by discourse analysts, but in both cases, there is a lack of a clear and consistent signal or rule that can be used to capture such examples.

To address this issue, one possible source of additional support might be in the area of machine learning. A machine learning system, given a large enough volume of data, might be trained to recognize patterns that indicate that a subject is being given increased powers and agency. The advantage of machine learning in this case is that rules do not have to be tightly defined; it can function on the basis of heuristics. This is consistent with the area of discourse analysis, which is governed more by trends than by rules.

Table 1. Discursive strategies linked to structuring propositions.

<table>
<thead>
<tr>
<th>Discursive strategy</th>
<th>Example from article</th>
<th>Supported by Stanford CoreNLP?</th>
</tr>
</thead>
<tbody>
<tr>
<td>[A] Naming sources in specific ways (through the use of formal titles associated with authority, or in association with others)</td>
<td>“Superintendent Jackie Candelario…told” “Senator Panfilo Lacson said…” “Sotto said in a statement…” “Senator Loren Legarda, the Roman Catholic Church, and non-government organization Gunless Society have all issued statements…” “Lacson joined the many who have been calling for stricter gun control….”</td>
<td>Yes. (1) Identify sources as subjects that are linked to verbs like “said”, “told”, “is quoted as saying”. (2) In cases where a source is given more credibility through the use of a title, set up the tool to identify titles. (3) In cases where multiple people are named or referred to, to strengthen a case, set up the tool to identify chains of proper nouns or collective nouns.</td>
</tr>
<tr>
<td>[B] Naming an event</td>
<td>Shooting is “bloodbath”, “carnage in Cavite” Total gun ban was [sic] just “a knee jerk reaction”</td>
<td>Partly. Identify two nouns linked by a linking verb, in effect saying “A is B” (“total gun ban” is “knee jerk reaction”). Limitation: the outright substitution of one word for another cannot be identified unless one article is compared with other articles (ie., “bloodbath” and “carnage” instead of “shooting incident” will not be flagged unless two articles are compared and these differences surface).</td>
</tr>
<tr>
<td>[C] Predication: use of modifiers to characterize people, events, places, objects</td>
<td>“a string of violent deaths” “seven year old girl Stephanie Nicole Ella” (in other articles, she is also described as a good student, a lover of the bible, as someone who was celebrating just a few meters from her home)</td>
<td>Yes. Determine adjectives, adverbs, other modifier phrases linked to a specific subject</td>
</tr>
<tr>
<td>[D] Transitivity: Participants being portrayed as engaged in issue resolution.</td>
<td>Caloocan cop “lists” Authorities “started inspecting” [Superintendent Jackie Candelario] “narrows down the [suspect] list from 45 to 32”</td>
<td>Partly. Identify use of active voice of verbs based on sentence structure. Limitation: The classification of “active” versus “passive” voice is not nuanced enough to capture more subtly strategies: in this case, verbs are mobilized to frame certain things as “finished” when in fact the bigger task is still ongoing. For example, it is noted that authorities “started inspecting”, not “authorities inspected.” It is reported that gun owners “requested” to hand in their guns, not that “gun owners have handed in their guns.”</td>
</tr>
<tr>
<td>[E] Modality: expressing one’s opinion, sometimes as fact</td>
<td>“…government should increase of the budget of the Philippine Drug Enforcement Agency (PDEA)… instead of imposing a total gun ban.”</td>
<td>Yes. Look for key words such as “should, ought to, may, can, will”</td>
</tr>
<tr>
<td>[F] Presuppositions</td>
<td>“Senator Panfilo Lacson said permits to carry firearms outside residence (PTCFOR) should no longer be issued to civilians so that their guns could only be used for self-defense” (presupposes that guns restricted to residences will only be used for self-defense)</td>
<td>Not supported</td>
</tr>
</tbody>
</table>
A second concern has to do with difficulties that arise when one is unable to use propositional parts as ways to reference other parts. Discursive Strategy [B], for instance, involves directly naming an event, in this case a shooting incident labeled as “carnage” or a “bloodbath”. Stanford CoreNLP should readily be able to detect this if the proposition is structured as “A is B” (or “The shooting was a bloodbath”), but a direct reference to the event as carnage will not be detected.

In this case, a solution can be mobilized as follows, and it involves using external texts as the reference point. A tool called a Gramulator [6] is capable of comparing multiple documents covering similar issues. The Gramulator compares sister corpora (two highly related sets of data) and identifies differentials that characterize one corpus but is considered “antithetical” of the second corpus. One could arguably enter similar articles on the same news event and establish which key words characterize Article A, and which key words characterize Article B. In this sense, it supports the discourse analysis practice of asking “why this way and not that way”, by sensitizing researchers to similar propositions can be framed in alternative ways.

Another set of discursive strategies with limitations in support are the argument strategies: Discourse Strategy [H] cause and effect, [I] expansion and elaboration; [J] problem-solution; and [K] dichotomization. Support is limited in two ways. First, these strategies are generally detected through visible cues (cause and effect by words like “because”, “so”, “then”, “therefore”). Limitation: none of the sample sentences use these words.

In the examples given, it appears there is a dichotomy between “civilian guns for self-defense only” and “no to total gun ban” are presented as opposing sides “Civilians will be allowed to own firearms” pit against “if we outlaw guns, only the outlaws will have guns” “government should increase of the budget of the Philippine Drug Enforcement Agency (PDEA)… instead of imposing a total gun ban.”

In this case, a solution can be mobilized as follows, and it involves using external texts as the reference point. A tool called a Gramulator [6] is capable of comparing multiple documents covering similar issues. The Gramulator compares sister corpora (two highly related sets of data) and identifies differentials that characterize one corpus but is considered “antithetical” of the second corpus. One could arguably enter similar articles on the same news event and establish which key words characterize Article A, and which key words characterize Article B. In this sense, it supports the discourse analysis practice of asking “why this way and not that way”, by sensitizing researchers to similar propositions can be framed in alternative ways.

Another set of discursive strategies with limitations in support are the argument strategies: Discourse Strategy [H] cause and effect, [I] expansion and elaboration; [J] problem-solution; and [K] dichotomization. Support is limited in two ways. First, these strategies are generally detected through visible cues (cause and effect by words like “because”; dichotomization by words like “either/ or”). This means that the absence of cue words would make these strategies difficult to detect. A second point is that subtleties in the strategies might not be captured. In the examples given, it appears there is a dichotomy between “civilian guns for

<table>
<thead>
<tr>
<th>Discursive strategy</th>
<th>Example from article</th>
<th>Supported by Stanford CoreNLP?</th>
</tr>
</thead>
<tbody>
<tr>
<td>[A] Topic shifting, for example from heading to most of the body</td>
<td>Headline is “Caloocan cops list 32 potential suspects in Nicole Ella’s death” but is only discussed in first 3 sentences out of 21</td>
<td>Yes. (1) Establish the discourse center for a proposition and assess the extent to which it is referred to in succeeding propositions. When it ceases to be referred to, the topic has changed. (2) For more obvious shifts, look for cue words that are linked to digression: “in other news”, “on another note”</td>
</tr>
<tr>
<td>[B] Argumentation via cause and effect</td>
<td>“The death of Ella has sparked public criticism over lax gun control” [With stricter gun control]…” “A lot of lives can be saved and a big positive impact on the country’s peace and order will be felt by the citizenry,” “Senator Vicente Sotto III believes a total gun ban in the Philippines will not prevent the recent shooting rampage in Cavite from happening again”</td>
<td>Partly. (1) Look for cue words associated with causality: “because”, “so”, “then”, “therefore”. Limitation: none of the sample sentences use these words.</td>
</tr>
<tr>
<td>[C] Argumentation via expansion</td>
<td>“The death of Ella has sparked public criticism over lax gun control in the Philippines and the shooting rampage last Friday in Cavite fueled pressure for more assertive action”</td>
<td>Partly. Look for cue words that mean elaboration or expansion: “and”, “in addition”, “further”. Limitation: such words can mean other things.</td>
</tr>
<tr>
<td>[D] Argumentation via problem-solution (or problem-non-solution)</td>
<td>“…total gun ban in the Philippines will not prevent the recent shooting rampage in Cavite”, “…government should increase of the budget of the Philippine Drug Enforcement Agency (PDEA)… instead of imposing a total gun ban.”</td>
<td>Not supported.</td>
</tr>
<tr>
<td>[E] Dichotomization: presenting two sides as opposites when in fact they are not</td>
<td>“Civilian guns for self-defense only” and “no to total gun ban” are presented as opposing sides “Civilians will be allowed to own firearms” pit against “if we outlaw guns, only the outlaws will have guns” “government should increase of the budget of the Philippine Drug Enforcement Agency (PDEA)… instead of imposing a total gun ban.”</td>
<td>Partly. Look for key words that suggest parallel construction and contrast: “on the other hand”, “in contrast”. Limitation: may be blurred if contrasting points are dispersed instead of presented point by point</td>
</tr>
<tr>
<td>[F] Generalizations from single cases</td>
<td>Shooting of Ella → need for gun ban Shooting in Cavite → solution is not gun ban</td>
<td>Not supported.</td>
</tr>
</tbody>
</table>
self-defense only” and “no to total gun ban”, but a close reading of arguments will show that one is not a direct opposite of the other.

A set of resources worth exploring in greater detail would be Rhetorical Structure Theory [9]. RST work can further support discourse analysis in two ways. First, it is founded on the premise that many structures and relations within and between texts do not have signals. Second, in exploring texts, RST researchers have been able to identify what appears to be a more finely grained continuum of possible relations between them. For example, the issue of “dichotomization” discussed earlier in light of more limited NLP tools, suggests that both sides of the gun control debate constitute an “either/or” scenario exploring two sets of text in equal footing. With RST, we can be sensitized to the fact that such cases can be explored in various ways: not just as a neutral contrast, but possibly as a concession, or as an antithesis. The example given in [K], for example, would be more accurately captured as an antithesis.

There were three discursive strategies that were assessed as having no support under the more basic tools. These included [J] problem-solving; [L] generalizing from a single case; and [F] presuppositions. While further work has to be done to explore the capabilities of RST resources to support these issues, a preliminary assessment is made here.

The issue of [J] problem-solving appears to be supported in two ways under RST: through its exploration of relations of “means” and relations of “solutionhood”. The strategy [L] of generalizing from a single case is not directly addressed, however, RST work has identified a relation called elaboration, wherein one can move from a generalization to a single case. It may be possible to use this to detect [L] (in effect, L in reverse) [9].

4. CONCLUSION

In this study I have explored the extent to which Stanford CoreNLP [11] supports a set of 12 discursive strategies identified across a predefined pool of newspaper articles. My conclusion is that there is considerable support for discourse analysis given the capabilities of existing NLP resources. Even “foundational” tools like Stanford CoreNLP [11] provide support (complete or at least in part) for most of the strategies identified here. In establishing why some strategies are easier to support (they use cue words or sentence structures), we can then move on to other types of discursive strategies, assess if they are also characterized by cue words or by signals related to structure, and if so conclude that these strategies can be supported as well. The limitations of foundational NLP tools can be filled by other resources like the Gramulator [6], RST [9], or machine learning techniques.

There are two possible areas of research that can be identified based on this study. The first has to do with discursive strategies related to presuppositions. One may argue that presuppositions cannot be explored computationally because they are hidden and implicit. However, it has been suggested that presuppositions can be signaled in at least three ways [7]. If this is the case, then they can be potentially supported by NLP tools.

A second point has to do with the strategy of making a generalization from a single case. From the point of view of logic this may at times be classified as a logical fallacy called a hasty generalization. This opens up the possibility of exploring the extent to which NLP techniques are, or at least can be, used to discern logical fallacies in texts.

5. ACKNOWLEDGMENTS

My thanks to Chris Cayaba, Raymond Timog, and Rhea Valbuena for their insights on the discourse analysis carried out here, as well as to Charibeth Cheng for her technical input into this paper.

6. REFERENCES