The Importance of Nouns in Text Processing

Lauren M. Stuart, Julia M. Taylor, and Victor Raskin ({lstuart, jtaylor1, vraskin}@purdue.edu)
Purdue University, 656 Oval Drive
West Lafayette, IN 47907-2086 USA

Abstract

In the area of computational processing of natural language texts, advances toward simpler yet more accurate models of meaning are desirable. As syntax is a major component of semantic analysis, we explore how a long-term institutional bias towards the verb as the main determiner of syntactic (and semantic) structure may underserve some kinds of information. We introduce an analysis paradigm that restores the noun to some importance in syntactic analysis. A noun-driven syntax representation has been developed and evaluated, and implications of its use in further processing and in better modeling of natural language meaning are investigated.

Keywords: Linguistics, Language Understanding, Human-Computer Interaction, Representation, Syntax, Semantics

Introduction

Text interpretation by computers is highly desirable and arguably necessary as we continue to produce and analyze text. One major benefit to the improvement of natural language understanding (NLU) for text is more intuitive natural interaction with highly structured or, conversely, loosely associated, large stores of information.

Text processing may proceed sequentially, on the assumption that only full (or major) analysis of the surface-ward structure yields the next deepest structure, where deepest structure is some formulation of the text's meaning, possibly applicable to other meanings of other texts, and the surface structure is the written or spoken input for a computer. This linear process encourages the development of incremental processing modules; that is, given some intermediate representation of something going on in the text, the module will produce a further refined model according to its internal rules and heuristics. For an example, take a phonetic processor that processes speech data and outputs a series of symbols for use in phonological and morphological analysis.

The process does not have to be linear; an alternative approach may parallelize the different analyses to some degree, even to the maximum possible (for instance, we cannot process any syntactic data if it has not yet been furnished). In this approach, iterations of processing may "clear up" a map of the sentence's interpretation (meaning) incrementally. Easy or simple rule applications start the process and such selections provide feedback for further selections in those areas of the map that are not yet clear, for some threshold of "clear" that is dependent upon the form and eventual use of the data.

Linear or not, any processing of a text from its surface form to some model of its meaning relies on various stages of language processing. We wish to explore how a bias in one of these stages, and its correction, affects processing in another.

Sentence Processing

As a sentence is analyzed, much importance is given to the verb(s): modal verbs modify the main verb; noun phrases participate as subjects or objects; any noun phrase not directly related to the verb may be a complement of a preposition, which is itself associated with the verb or a noun phrase, or of some other clause or phrase whose meaning props up the meaning of the verb (the event said to be encoded in this particular sentence) and whose place or expression in the syntactic structure of the sentence is as much (or more) dictated by the verb as the head of the phrase. Nouns generally remain building blocks of arguments to be fed into a verb.

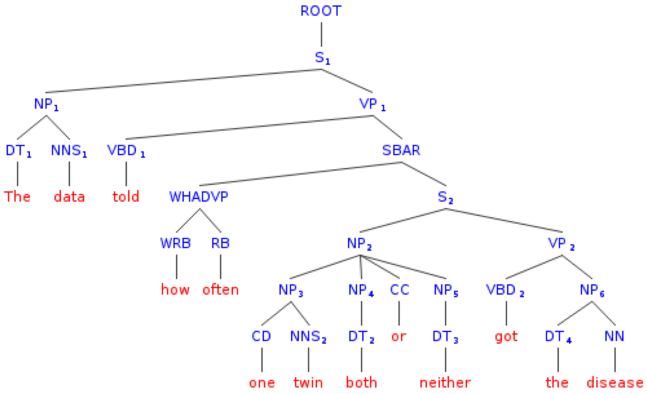


Figure 1: Representation of a phrase structure parse from Berkeley Parser

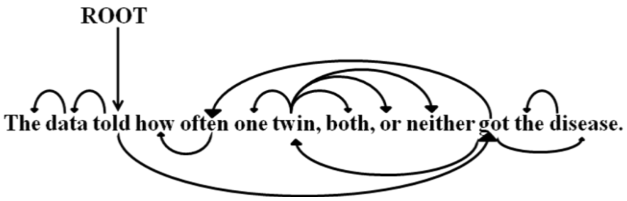

ROOT
The data told how often one twin, both, or neither got the disease.

Figure 2: Representation of a dependency parse from Stanford Typed Dependencies

In syntactic representations, this privileging of the verb is typically expressed as a shorter distance between the root of the tree and the main verb; some representations go so far as to shorten the paths to the other verbs in the sentence. See Figure 1 for a constituency tree representation, and note the comparative height of the verb and its nouns. Then see

Figure 2 for a dependency tree, noting how the removal of some phrase structure also removes some steps to access any particular word, but verbs are still comparatively close to the root and can be accessed from one another. This reveals an intuition that a sentence is primarily "about" its verb and that composition of multiple sentences is "about" lining up the verbs together.

In the process of computational understanding of natural language, a computer may be given a syntax tree from which to construct a meaningful model of the events and objects that the sentence describes (with "meaningful", here, determined by the eventual use to which the model will be put: are we looking for frequencies of events? a reconstruction of the actions of one particular object? similarities in locations, origins, or attributes of events or objects?). Then, to maintain the appropriate interpretation of meaningfulness, the distance of a particular word or word category from the root of the sentence must correlate to its need, or incorporation, in the construction of this model. Given syntax trees that are verb-centered, it is most efficient to construct verb-centered semantic representations. Processing may not need to be sequential (as in, phonological then morphological then syntactic then semantic), but we will leave that possibility alone for right now.

A mapping may be observed between simple sentences and logical expressions in first-order predicate logic: *The images show a landscape* can be formulated as a function *show()* with the arguments *images* and *landscape*. The proposition *show(images, landscape)* is held to be equivalent to the sentence – that is, if it is tested for a binary truth value, it evaluates to "true" in all the situations in which the sentence from which it comes would be considered true. More complex statements can be generated using such rules as well, e.g., *show(images, landscape) & is_on(landscape, Mars).* By mapping natural language sentences to this restricted logical form, we arrive at a sort of semantic notation that is easier, somehow, for a computer to use. Its close resemblance to the syntax of many programming languages suggests that, if only we can translate all sentences into such expressions, we can execute the program obtained by concatenation (in accordance with rules for coordination, negation, etc.) of these expressions and thereby arrive at some truth value for the sentences taken together. We may not just want an answer (true or false) but a model of the meaning in the text; tweaking the execution of these formulae may allow us to build that model.

However, first-order predicate logic is not entirely adequate. Luuk (2009) extends the mapping to a less strict system and theorizes about the possible evolution of argument-like concepts (nouns) before predicate-like concepts (verbs, among others).

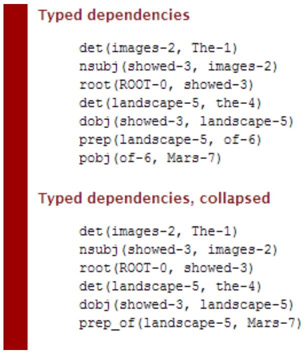
Still, all of this analysis is predicated on the idea that the verb (or the event it describes) is the central element of analysis, from which all other considerations flow. However, there are some natural and regular instances in which the verb is no help, or possibly even absent. Take a copular sentence: *The Curiosity is the Mars rover*. While the existential senses of *is* are large and have many implications, they are only manageable when knocked down to the scope of *is the Mars rover*. In practice, the verb in this sentence is demoted to purely technical predicate status and the predicative nominal elevated in its place. Now we are to analyze a noun phrase as a predicate; there is plenty of precedent, as we can talk about noun-expressed events taking arguments structured similarly to those their verb-expressed counterparts accept. Compare *We celebrated the launch today* with *The celebration of the launch was today*.

However, the question must be asked: why is our analysis so verb-centered, and to such a degree that we must postulate verbs, i.e., essentially to create dummy verbs, where there are none?

Towards a Noun-Driven Paradigm

We propose an alternative, perhaps complementary analysis paradigm: center the noun. Such a paradigm might include analysis of concepts as informational objects – for events to be frames – and events as actions somehow intrinsic or controlled by the objects they involve. This paradigm may open up a world of gains in processing different flavors of information sources, particularly those that have been traditionally managed by computers, with different degrees of naturalness in the language used to interact with them. A noun-driven paradigm may then boost ease of interaction with these sources via natural language understanding by simply not introducing an unneeded event structure for analysis.

To this effect, we have proposed a noun-driven syntax representation (Stuart, 2012). It inherits from the class of dependency grammars by formulating syntax rules as directed binary relationships between nodes. For instance, a preposition may be eliminated entirely and encoded in the syntactic tree as a directed relation, carrying the meaning of the preposition, between the elements it used to connect. For an example, see Figure 3.



Typed dependencies

```
det(images-2, The-1)
nsubj(showed-3, images-2)
root(ROOT-0, showed-3)
det(landscape-5, the-4)
dobj(showed-3, landscape-5)
prep(landscape-5, of-6)
pobj(of-6, Mars-7)
```

Typed dependencies, collapsed

```
det(images-2, The-1)
nsubj(showed-3, images-2)
root(ROOT-0, showed-3)
det(landscape-5, the-4)
dobj(showed-3, landscape-5)
prep_of(landscape-5, Mars-7)
```

Figure 3: Prepositional collapse in Stanford Typed Dependencies Output

There may be tradeoffs between dependency grammars and constituency grammars – Nivre (2006) considered the tradeoffs favorable – but some may be more important in the noun-driven paradigm (for instance: should it perhaps be the noun-phrase-driven paradigm?) and only further investigation will reveal these.

A noun-driven representation has been developed, starting in Stuart, et al. (2012a) and Stuart, et al. (2012b). The structure links all nouns from the root, so a parallelized meaning-scaffolding program may have several starting points and begin to converge upon an intermediate structure (towards a model of meaning) as it traverses the nodes held in common between noun-rooted subsets. The corresponding parallelization applied to verb-driven dependency grammar representations does not result in the same gains: there are typically two noun phrases to every verb phrase in English (Baker, 2005). The number of nouns relative to verbs only gets larger when we consider noun chains ("crater rim"—see also Taylor et al 2010, 2012) and prepositional phrases (the objects of which are always noun phrases).

The number of prepositional phrases also contributes to the complexity of syntactic analysis. Some prepositional phrases have ambiguity in attachment; some may attach somewhere in syntactic analysis but be restricted from that attachment during more meaning-directed evaluation. Consider *The images show a landscape on Mars* vs. *The images show a landscape of Mars*. In the first, *on* may attach to both *show* (the event of showing could occur on Mars) and to *landscape* (the landscape is specifically one on Mars). In the second, *of* may only attach to *landscape*, but can sometimes attach to other verbs (as a particle, for instance, in *to think of*), so the evaluation of the attachment as valid may take several steps of analysis. Multiple prepositions can also attach to the same elements; as they all will have noun-phrase objects, the centering of the noun phrase in analysis hikes the prepositional phrase up in the hierarchy of importance as well.

Implications for Semantic Processing

The "object-oriented" nature of noun-driven syntax may also align sufficiently well with object-oriented semantic languages to collaborate easily in parallelized processing, allowing groups of objects at certain "stages" to be swapped out with "higher-stage" interpretations or representations of them.

Take the sentence introduced before: "The images show a landscape" and its partial processing, as shown in Figures 2 and 3. In the latter, "show" has been tagged as a noun rather than a verb. Possibly these are candidates with the highest confidence due to internal simplicity, some rules about sentence formation, the topic of the information, or from previously-mentioned information: "the images" likely refers to some set of images that have already been introduced. In a pass that has produced some semantic or intermediate representations for parts of the map——the intermediate conclusions made in the other parts of the sentence contribute to analysis of the other parts. Analogously, in a greedy meaning algorithm, the subsets of the sentence which are simplest to compute or represent drive the interpretation of the rest of the sentence.

As in many syntactic analysis algorithms, "steps" of processing can be reverted (or previous states saved) to enable backtracking or the output of multiple best candidates if appropriate. The processing shown in Figures 4 and 5 presupposes object-oriented semantic-processing modeling. For those systems with event-oriented semantic processing, verb-driven syntactic approaches (linear or parallel) are just as useful.

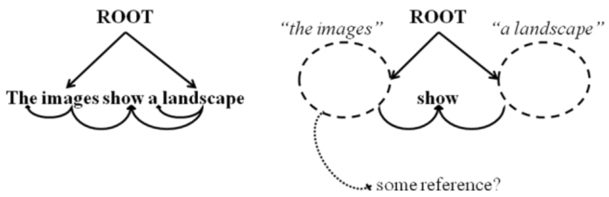


Figure 4: Processing steps when "show" tagged as a verb

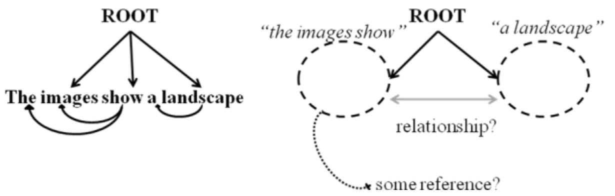


Figure 5: Processing steps when "show" tagged as a noun

Regardless of whether the semantic modeling used in process is object-oriented, event-oriented, or some hybrid, we find it prudent to also consider the possible object- or event-biased information sources. Suppose we have an encyclopedia article about the geographical features of Mars. While many sentences will perhaps explain the dates and methods of discovery of certain facts about Mars and its geography, we can expect a large amount of discussion on the features themselves, their attributes, positions, and the larger classes to which they belong. In the example sentence *A peak sits to the south of the crater rim,* we receive information that there is some peak (perhaps in a larger range of mountains) somewhere south of a crater that has been (we hope) previously introduced. This is position information; the verb is *sit* but if it were for some reason left out of the sentence or replaced with *is*, *is located*, or *is situated*, we would still get the gist.

Consider, then, a general article about geography, the content of which an NLU system will have to make use of in order to understand the instantiations (in the Mars geography article) of types and classes described in the general article. We can expect some events here: various geographers may have contributed to the development of some concepts. However, we argue that the most useful part of the article (again, for understanding something about the features of Mars) is that part which defines the types, attributes, and relative locations of geographical features. This area of the article is likely to be strongly noun-centered: positions, lists of items, and weak verbs may dominate. (*A volcanic crater is a circular depression in the ground.*) This is not to say that the verb-driven paradigm is useless in a very noun-dominated information space. However, a verb-driven syntactic analysis may waste some resources by trying to identify and promote predicates. As well, a verb-driven semantic analysis may find that most of its "events" are existential and not temporally-bounded instances of some action.

Now consider some more rigidly-defined information spaces. The most rigid might be the company relational database, which strictly encodes lists of relationships between objects – Li et al. (2008) developed an engineering ontology and lexicon for processing natural language search queries, in a supply chain application, to just such a database, which also included information in unstructured (that is, natural language) descriptive documents. The user queries (and therefore the ontology) emphasize parts' shapes, materials, purposes, and origins. Such domains are fundamentally concerned with objects and how they are related to each other. The "translation" from machine-readable language to natural language must rely heavily on nouns, adjectives, quantities, and the relationships and attributes of semantic concepts representing "things" rather than "happenings". The basic semantic structure of the information is already there: the database schema gives us relations by which to connect the objects (event instances, object-parts, etc.) that are its subject.

Take now, for example, a computer program written in some programming language. Improvement of NLU also underwrites the improvement of translation between natural language and programming languages (for instance, in specifying and evaluating privacy constraints, as in Brodie et al. (2006)). The language itself does not need to be conceived of as object-oriented (Java, C++) because a reductive view of a computer program is that of an informational object which takes informational objects as inputs and gives more information objects as outputs. The transformations that these input objects undergo are events, the transformations may be affected partially or wholly by some qualities of the objects to which they are applied. This is a strength of inheritance and polymorphism in those object-oriented languages: one prototypical "event" can be expressed in terms of the classes of objects to which it is applied, and many events are specific, intrinsic, and unique to a (class of) informational object (data structure). Thus, "translation" of the event relies ultimately upon the objects involved; that is, *Object.do()* is specified at least in part by the implementation of the class *Object*. Even in a language that is not specifically object-oriented, the specific actions undertaken in the execution of the method *process(thing)* depend on the content or nature of the object *thing*.

Finally, consider the sentence "Airborne geomagnetic surveys showed a strange pattern of symmetrical magnetic reversals on opposite sides." Our main verb here is "show", but the important events (the act of surveying, and the occurrence of magnetic reversal) appear in noun form, and some important attributes appear as adjectives and a prepositional phrase. Verb-centered processing (purely syntactic or as a step towards semantic processing) prioritizes "show"; it stands in for a fuller explanation that investigating scientists learned about the reversals by reviewing data from the surveys, and thus does have some importance (for instance in auditing the assertion "Many magnetic reversals have occurred", as one question could be "How do we know that?"). In an article about the scientific process, its many forms, and its contribution to knowledge, this is a salient detail. However, in reading an article about the geomagnetic history of the earth, we may be much more interested in the apparent occurrence of magnetic reversals, and the source of the data analyzed in order to reach that conclusion.

Experimental Evaluation

Stuart et al. (2012b) began evaluating the performance of the noun-driven syntax in a small query context: assuming that most queries to syntax trees take the form of traversing the tree to or from a certain node or a node of a certain category, the node's depth is used as a rough measure of accessibility for further analysis. The initial experiment was carried out by hand and in comparison with outputs from Stanford Dependencies, Stanford PCFG, and Berkeley parsers. The dataset consisted of only 30 sentences; the noun-driven syntax representation performed at least as well, or better, than the dependency grammar trees, and both much better than the phrase-structure trees. As well, the dependency and noun-driven trees were evaluated from a parallelized perspective.

A larger experiment used 600 sentences, chosen from six different articles, each of which fell under one of three categories. The categories – "noun-heavy", "verb-heavy", and "neither" – attempted to capture a meaning-motivated difference in syntax between information sources, as well as to test intra-category variance to a degree. The performance of the noun-driven syntax was compared to that of a developed hybrid phrase structure representation – the latter considerably "flattened" phrase structures. Direct comparison with a dependency parser's output could not be obtained for this larger experiment due to technical difficulties, but it is planned for the future.

The noun-driven representation was generated by a phrase-structure parser integrated with Ontological Semantics Technology (OST), a natural language understanding framework under current development (Taylor et al., 2012). The parser used a partial lexicon – a set of word entries with associated syntactic information, intended eventually to include semantic, morphological, and phonological information useful for word sense disambiguation and construction of semantic meaning representations. The parser used a modified chart-parsing algorithm, similar to that presented in Allen (1987) but organized around heads of phrases rather than building phrase structure from left to right. The parser generated parses in phrase-structure representations then converted those to noun-driven trees. An example representation of the output appears below in Figure 6; compare with Figures 1 and 2.

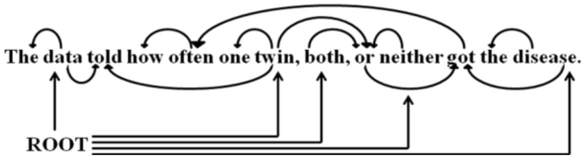

The data told how often one twin, both, or neither got the disease.
ROOT

Figure 6: Representation of a noun-driven dependency parse

A testing program counted the depths for each word in the word categories of interest (determiner, noun, verb, adverb, preposition, adjective). One metric for appropriateness of the categories is in measuring the ratio of nouns to verbs; this data is shown in Table 1. A "fingerprint" for each of the subcategories (counts by word class) appears in Figure 4; note some similarity within each of the categories. However, these findings are complicated by unknown effects of style (the two "noun-heavy" articles were from Wikipedia, the two "neither" articles from the New York Times website, and the last two from Safety.gov and a recipe book, respectively), though there may be an association between style and subject matter that does uphold the categorization. As well, these results do not distinguish between parses with differing levels of correctness or acceptability; work in progress does mark parses on a spectrum from non-grammaticality to candidacy as the parse most compatible (with some semantic processing) with the rest of the article.

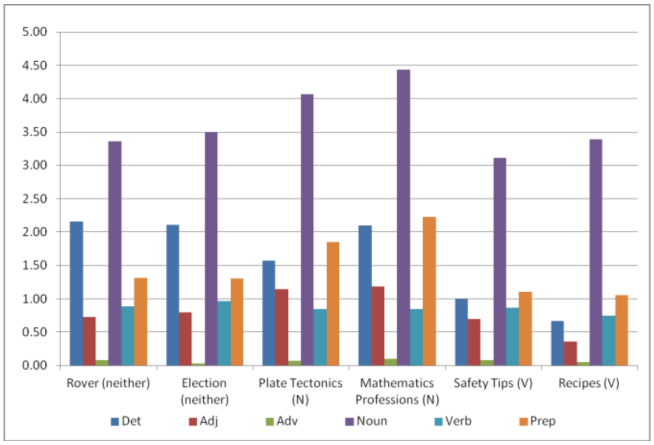


Figure 7: Word class counts per parse, by subcategory

Table 1: Noun/Verb Ratios

Sentence subcategory(article subject)	N/V
Neutral 1 (Mars Rover)	3.82
Neutral 2 (2012 Election)	3.65
Noun-heavy 1 (Plate Tectonics)	4.82
Noun-heavy 2 (Mathematics professions)	5.23
Verb-heavy 1 (Safety Tips)	3.59
Verb-heavy 2 (Recipes)	4.56

A sample of 64 sentences, randomly selected from the 600, was tested for parser correctness and accuracy – the distinction arises from a difference between transforming a string of grammar symbols into all possible syntactically correct parses of the symbols, regardless of their content, and obtaining the correct parse, as the syntax parser has no ability to determine which of the correct strings of symbols is actually completely grammatically correct. This is a result of an inexpressive tag set and a lack of semantic parsing integration. For the 64-set, in 6 cases were the correct parses not included in the output; this was due to a lexicon gap, verb particles not correctly accounted for, or an oversight in vetting the data set for conformance to the limited grammar template that the parser was designed.

For the 600 sentences, some measures were taken of possible syntactic ambiguity: if for one sentence the parser turned out more than one parse and could be counted on (according to an interpretation of the outcome of the 64-set results, which is not entirely dependable) to be correct, if enthusiastic, in its determination of good syntactic parses, then the sentence was determined to carry syntactic ambiguity. Of the 600, 241 sentences were given only 1 parse; an equal number had 2. 16 sentences had 10 or more different parses turned out; some of these were due to prepositional phrase attachment ambiguity, and some to possibilities that, for instance, the form of a verb in the 3rd person-present-singular is identical to that of a plural noun, or vice versa.

Table 2: Depth Counts for Nouns

	Noun-Driven			Phrase-Structure		
Group	Average	Min	Max	Average	Min	Max
All	1	1	1	3.38	2	9
Neutral 1	1	1	1	3.29	2	9
Neutral 2	1	1	1	3.34	2	8
Noun-heavy 1	1	1	1	3.48	2	9
Noun-heavy 2	1	1	1	3.66	2	9
Verb-heavy 1	1	1	1	3.27	2	8
Verb-heavy 2	1	1	1	2.98	2	7

Table 3: Depth Counts for Verbs

	Noun-Driven			Phrase-Structure		
Group	Average	Min	Max	Average	Min	Max
All	2	2	2	2	2	2
Neutral 1	2	2	2	2	2	2
Neutral 2	2	2	2	2	2	2
Noun-heavy 1	2	2	2	2	2	2
Noun-heavy 2	2	2	2	2	2	2
Verb-heavy 1	2	2	2	2	2	2
Verb-heavy 2	2	2	2	2	2	2

Table 4: Depth Counts for Prepositions

	Noun-Driven			Phrase-Structure		
Group	Average	Min	Max	Average	Min	Max
All	2.38	2	3	3.60	2	8
Neutral 1	2.40	2	3	3.58	2	8
Neutral 2	2.47	2	3	3.68	2	7
Noun-heavy 1	2.34	2	3	3.69	2	8
Noun-heavy 2	2.33	2	3	3.59	2	8
Verb-heavy 1	2.56	2	3	3.64	2	7
Verb-heavy 2	2.32	2	3	3.26	2	6

Tables 2 and 3 show results for depth counts between the two syntax representations evaluated, over the 600-sentence set, for the two main categories, noun and verb. We include the depth counts for prepositions as well (Table 4) because of the complexity that prepositional phrases add to syntactic structure.

Examination of these tables reveals that, even when compared with a “flatter” (verb-driven) phrase structure syntax representation, the noun-driven representation does at least as well, if not better. Notice as well that in the phrase-structure trees, nouns have a wider range of “float” because, while a single noun may be the subject and thus be found in a shallower position, prepositional phrases and noun-chaining bury nouns further.

Preliminary results from further evaluation also reveal data that may be usable for characteristic profiles of prepositional attachment. As well, we may investigate whether the addition of some phrase structure features to the dependency-like representation would provide better information for prepositional attachment and other local operations influenced by concepts or structure use.

Conclusion

If computational processing of information is (even sometimes) object-centered, then an object-centered approach aligns with it. Given that we have started at the syntax level, and that most objects (as well as some events) are typically expressed as nouns, the noun-driven syntax representation, and an eventual development of a parsing approach, begins the building of a noun/object-centered paradigm for the analysis of natural language text.

There are information spaces that are not so noun-biased, or even further verb-biased – with the full field of verb-driven syntactic and semantic analysis, these will not be left behind. A dual analysis, using both perspectives, may produce some gains in efficiency and effectiveness.

Acknowledgments

This research has been supported in part by the National Science Foundation CISE/TC Research Grant #1012208.

References

Allen, James. 1987. *Natural language understanding (2nd edn).*Addison-Wesley.

Baker, Mark. 2005. "Lexical Categories: Verbs, Nouns, and Adjectives." Cambridge: Cambridge University Press.

Brodie, C. A., Karat, C., & Karat, J. (2006). An empirical study of natural language parsing of privacy policy rules using the SPARCLE policy workbench. SOUPS '06 Proceedings of the second symposium on Usable privacy and security, 1.

Klein, Dan and Christopher D. Manning. 2003. Accurate Unlexicalized Parsing. Proceedings of the 41st Meeting of the Association for Computational Linguistics 423-430.

Luuk, Erkki. 2009. The noun/verb and predicate/argument structures. *Lingua* 119 (2009): 1707–1727.

de Marneffe, Marie-Catherine and Christopher D. Manning. 2008. The Stanford typed dependencies representation. In Proceedings of the workshop on Cross-Framework and Cross-Domain Parser Evaluation 1-8. Manchester, UK.

de Marneffe, Marie-Catherine, Bill MacCartney, and Christopher D. Manning. 2006. Generating Typed Dependency Parses from Phrase Structure Parses. 5th International Conference on Language Resources and Evaluation (LREC 2006) 449-454.

Li, Zhanjun, Victor Raskin, and Karthik Ramani. 2008. Developing Engineering Ontology for Information Retrieval. Journal of Computing and Information Science in Engineering 8(1):011003-13.

Nivre, J. 2005. Dependency Grammar and Dependency Parsing. MSI report 05133. Växjö University: School of Mathematics and Systems Engineering, Växjö, Sweden

Raskin, V. 1983. *A Concise History of Linguistic Semantics Pt 1*. Department of English and Program in Linguistics, Purdue University, West Lafayette, IN.

Stuart, Lauren M. 2012. *Computational Evaluation of a Noun-Driven Syntax Representation.* Unpublished Master’s Thesis, Linguistics Program, Purdue University, West Lafayette, IN.

Stuart, Lauren M., Julia M. Taylor, and Victor Raskin. 2012a. Towards Noun-Driven Parsing. In Proceedings of 2012 IEEE International Conference on Systems, Man, and Cybernetics (SMC) 1984-1989. Seoul, South Korea.

Stuart, Lauren M., Julia Taylor, and Victor Raskin 2012b. “Noun-Driven Syntactic Parsing for Natural Language Interfaces to Object-Centered Information Stores.” In Proceedings of Society for Design and Process Science Conference (SDPS). Berlin, Germany.

Taylor, Julia M., Victor Raskin, Christian F. Hempelmann, and Max S. Petrenko 2010. “Multiple Noun Expression Analysis: An Implementation of Ontological Semantic Technology, *Computational Linguistics – Applications Workshop*, Wisla, Poland.

Taylor, Julia M., Victor Raskin, and Lauren M. Stuart (2012). Machine Human Understanding: Syntax and Semantics Revisited. IEEE-SMC Conference Seoul, S. Korea.