The Role of RDF in the Empirical Search for Verb Classes

Hu Qinan
City University of Hong Kong
50002277@plink.cityu.edu.hk

I. Current Status of The Project

My research is concerned with the empirical search for verb classes. I am using the same BLIS legal corpus as the EBMT project.

II. Linguistic Theory

For a better understanding of my research methodology, let me address some of the fundamental issues first.

Semantics refers to the analysis of meaning and syntax refers to the analysis of sentence structure.

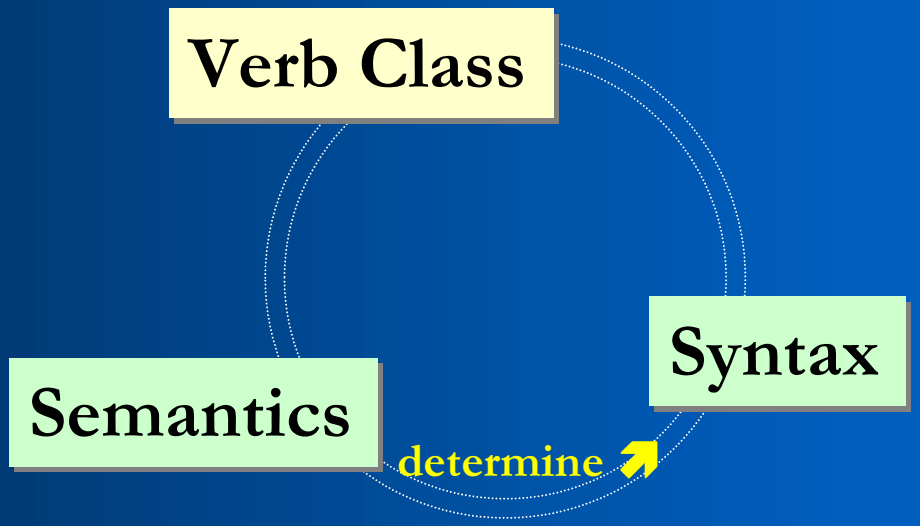
II. Linguistic Theory

As far as verbs are concerned, the study of semantics deals with predicate argument structure, theta role, *or* meaning component.

The syntactic behavior of verbs, on the other hand, deals with subcategorization and alternation.

SEMANTICS	SYNTAX
Predicate Argument Structure	Subcategorization
Theta Role	Alternation
Meaning Component	

II. Linguistic Theory



The basis for my investigation into verb classes arises from the following linguistic hypothesis:

The Syntactic behavior of verbs is semantically determined.

II. Linguistic Theory

Let us compare the following sentences involving the verb **break**:

VERB	break
SENTENCE	1. John broke the cup.
MIDDLE ALTERNATION	2. The cup breaks easily.
CONATIVE ALTERNATION	*3. John broke at the cup.

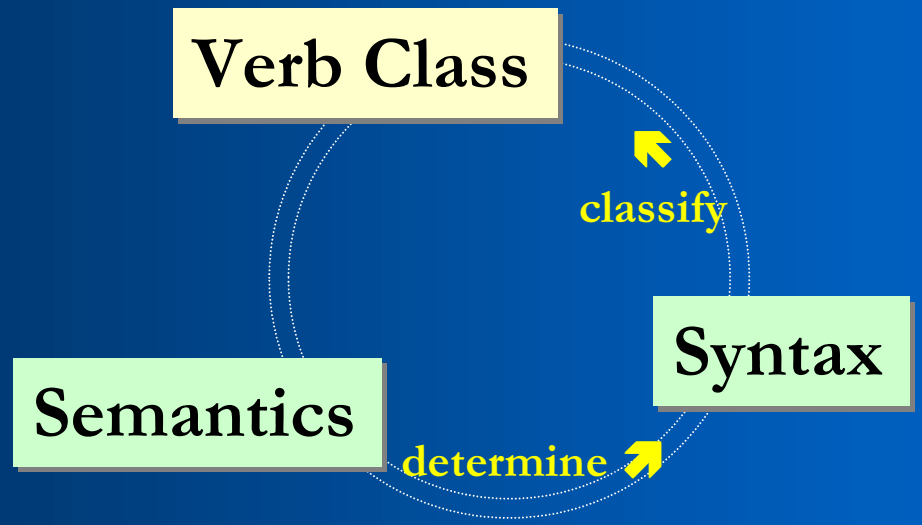
Native speakers of English know that only the 3rd sentence, which is a **conative alternation**, is invalid.

II. Linguistic Theory

	MEANING COMPONENT			ALTERNATION
	CHANGE OF STATE	CONTACT	MOTION	
✓	☑	☐	☐	Middle Alternation
x	☐	☑	☑	Conative Alternation

This table illustrates that an alternation of a verb is acceptable only if the corresponding meaning components are inherent to this verb.
Since the meaning of the verb break entails the concept of a change of state, it enters into the middle alternation. Alternatively, break does not necessarily involve a contact or a motion, thus it does not participate in conative alternation.

II. Linguistic Theory



Through further observation, one discovers that verbs can be grouped into classes according to their **syntactic behavior**.

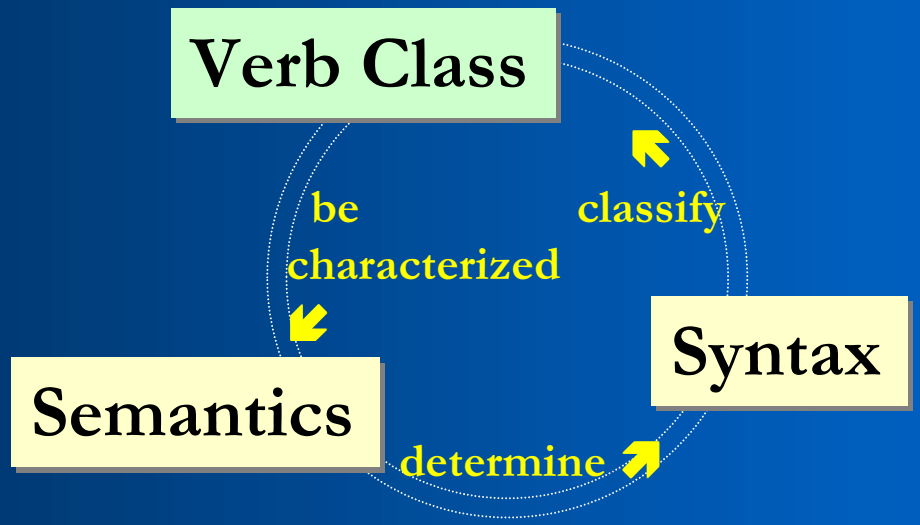
II. Linguistic Theory

For instance, using alternations as the criterion, verbs

break, crack, smash, split…

will fall into the same class, since they all participate in middle alternation, whereas none occurs in conative alternation.

II. Linguistic Theory



Interestingly, these syntactically generated verb classes can usually be semantically characterized.

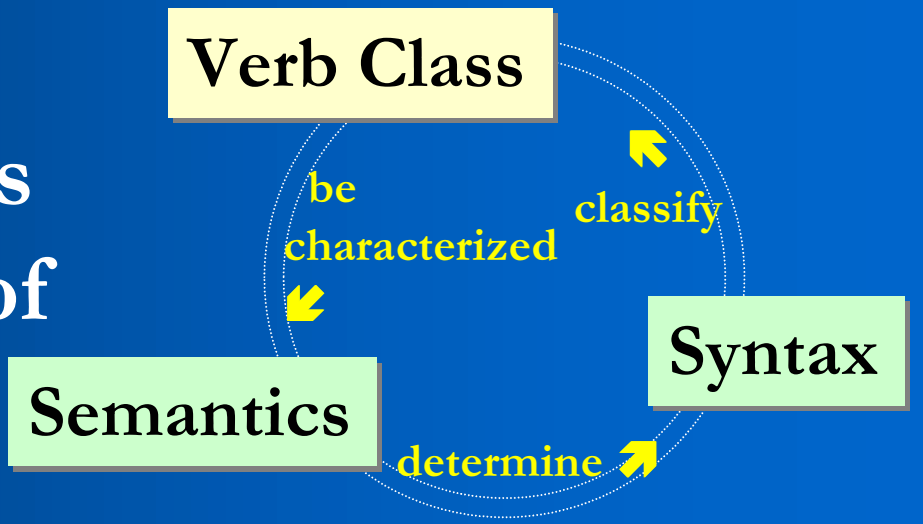
II. Linguistic Theory

break, crack, smash, split…

For instance, the above set of verbs shares a particular meaning component in common, which is that a change of state is involved.

II. Linguistic Theory

As an interface, the class of verbs undergoes certain semantic constraints by taking the form of corresponding syntactic behavior.



II. Linguistic Theory

Rather than considering each verb separately, understanding how verbs form certain class as a whole will help to systematically grasp the nature of a wide range of linguistic phenomenon.

For instance, some research findings on an individual verb can be generalized to a class of verbs according to shared semantic or syntactic properties.

III. Research Methodology

Verb classes are not fixed collections. Instead we need to consider how we might construct a knowledge-base from which verb classes could be dynamically generated.

III. Research Methodology

The major steps involved in the project include:

1. Morphological Analysis
2. Part-Of-Speech Tagging
3. Syntactic Analysis
4. Knowledge-base Construction
5. Verb Class Generation
6. Integration with the EBMT project

III. Research Methodology

- In morphological analysis, an inflected word is mapped into its root form. For instance, "walks" will be distinguished as the word "walk" plus a single present tense;
- In Part-Of-Speech tagging, words will be identified as noun, verb, adjective and so on;
- In syntactic analysis, we examine how words combine together to form phrase structures.

III. Research Methodology

There already exists software for adequately accomplishing many of the above tasks so the emphasis of my investigation is on the following steps related to the construction of a knowledge base.

IV. RDF Statements

Choosing RDF statements as the building blocks of the knowledge-base, all information will be represented in the form of a triple comprised of

{predicate, subject, object} where

predicate is a property;

subject is a resource;

object is a resource or a literal.

IV. RDF Statements

- **Reduction**

An exact mathematical model for RDF is the predicate logic. A theorem in predicate logic claims that any statement is able to be reduced into certain two-place predicates.

$$P(X, Y, Z) \Rightarrow R(Q(X, Y), Z)$$

IV. RDF Statements

- **Reduction**

Since the RDF statement is a two-place predicate in truth, it is capable to represent all kinds of information, no matter how complicated the original data structure of the information is.

Consider as a example the sentence:

"John walks to Manhattan."

IV. RDF Statements

- Reduction

The information generated in the morphological analysis and part-of-speech tagging can be expressed in RDF/XML as follows:

```
<rdf:description about =
"http://cpct86.cityu.edu.hk/verbclass/syntax/walks#1">
        <rootForm>walk</rootForm>
        <partOfSpeech>verb</partOfSpeech>
        <tense>present</tense>
        ...
</rdf:description>
```

IV. RDF Statements

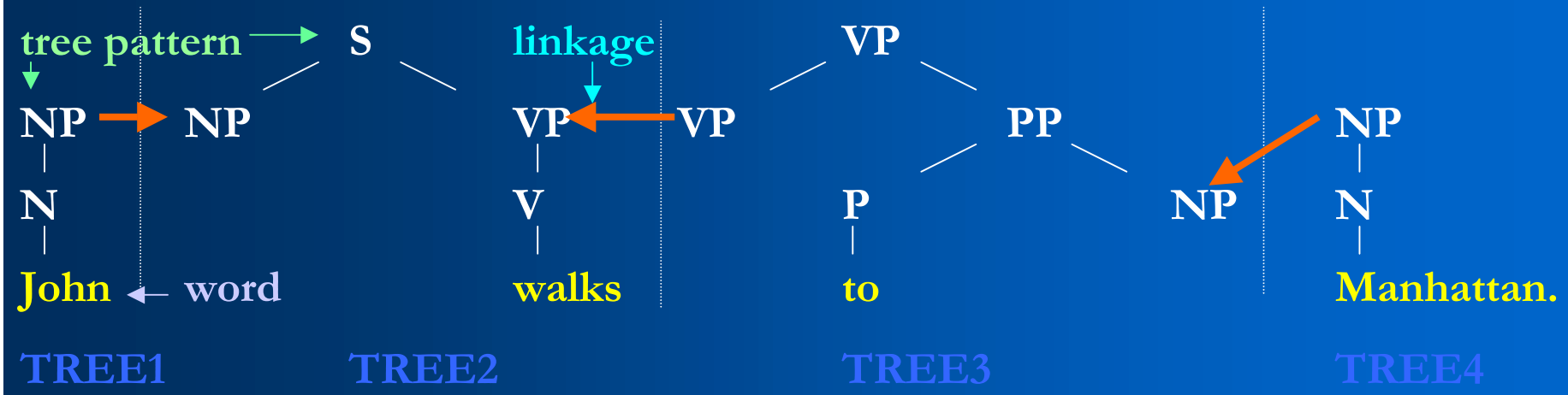
- **Reduction**

The next example involves the representation of the syntactic structures of sentences. It illustrates how complicated information is reduced into simple RDF statements.

The phrase structures of the above sentence can be represented in TAG (Tree Adjoining Grammar) as follows:

IV. RDF Statements

• Reduction



• As seen, this sentence is composed of four syntactic trees.

• Three categories of information are involved in each tree, that is inflected word, tree pattern, and linkages to other trees.

IV. RDF Statements

- **Reduction**

  The following RDF/XML snippet represents that the phrase structure of the preposition "to" in above sentence is derived from the inflected word "walks" using certain syntactic tree pattern.

```
<rdf:description about =
"http://cpct86.cityu.edu.hk/verbclass/syntax/to#4">
        <linkTo>http://cpct86.cityu.edu.hk/verbclass/syntax/walks#1</linkTo>
        <treePattern>http://cpct86.cityu.edu.hk/verbclass/tree/preposition#2</treePattern>
<rdf:description>
```

V. Knowledge-base Construction

Using RDF statements, a knowledge-base with the following characteristics can be constructed:

- Syntactic material of verbs should be used as the source data of the ontology.
- Different syntactic criterion should be allowed to apply to this knowledge-base, yielding various verb classes for specific purposes.

V. Knowledge-base Construction

- Due to the inherent flexibility of language, minor violations of the classification criterion may be expected.
- The system output should be expressed not only in the description of class properties, but also class interrelationships.

VI. Query and Inference

Using the query & inference capability of RDF, the following information on verb classes can be obtained:

- By querying the system for verbs which satisfy certain syntactic alternations, verb classes can be generated.

VI. Query and Inference

- By querying the system for all the syntactic patterns associated with a particular verb, inference can be drawn to obtain all the alternations this verb participated.

- Using the co-occurrence rate of the arguments of verbs to query the verb classes generated, verb sense identification may be achieved.

VII. Integration with EBMT

The last step of my investigation is to integrate this module with other components in the EBMT project for further learning of language patterns for *example acquisition.*

VII. Integration with EBMT

RDF statements not only provide a simple common framework for interior information of the knowledge-base, it also gives an open interface for the addition of new homogeneous information, upon which the expansion of the knowledge-base is no longer a bottleneck.