Recognition of alternation paraphrases: a robust and exhaustive symbolic approach

Marilisa Amoia
Dept of Computational Linguistics
University of the Saarland
Saarbrücken Germany
amoia@coli.uni-sb.de

Claire Gardent
CNRS/Loria
Campus Scientifique BP 239
54506 Vandoeuvre-les-Nancy, France
claire.gardent@loria.fr

Abstract

In this paper we show how by incorporating linguistic knowledge in a shallow parser like XIP, it is possible to build a robust semantic parser which can cope with paraphrastic constructions involving alternations and/or lexical synonymy. The robustness of the parser is dependent on the amount of linguistic knowledge at disposal. For the moment, we have incorporated the VerbNet verb lexicon in our parser, so that it can recognise verb synonymy and meaning preserving alternations of declarative sentences. In the future, we want to extend the parser taking into consideration grammatical variations of declarative sentences and more complex cases of paraphrastic constructions, such as nominalisations and noun phrase-adjective substitution.

Topics: use of language resources for reasoning in question-answering.

1 Introduction

A salient feature of natural language is that it allows paraphrases that is, it allows different verbalisations of the same content. Thus although the various verbalisations in (1) may have different pragmatic or communicative values (with respect for instance to topicalisation, presuppositions or focus/ground partitioning), they all share a core semantic content, the content approximated by a traditional montagovian compositional semantics.

(1) a. This key opens the safe.
The safe opens with this key.

b. The water fills the jug.
The jug fills with water.

c. The laboratory merges with the firm.
The laboratory and the firm merge.

d. Jean hit the wall with a stick.
Jean hit the stick on the wall.

e. I give books to John.
I give John books.

Linguists have long noticed the pervasiveness of paraphrases in natural language and attempted to caracterise it. Thus for instance Chomsky's transformations capture the relation between one core meaning (a deep structure in Chomsky's terms) and several surface realisations (for instance, between the passive and the active form of the same sentence) while [Mel'čuk, 1988] presents sixty paraphrastic rules designed to account for paraphrastic relations between sentences.

More recently, work in information extraction (IE) and question answering (QA) has triggered a renewed research interest in paraphrases as IE and QA systems typically need to be able to recognise distinct verbalisations of the same content. Because of the large, open domain corpora these systems deal with, coverage and robustness are key issues and much on the work on paraphrases in that domain is based on automatic learning techniques.

In this paper, we investigate an alternative research direction and present a symbolic treatment of paraphrases which (for the moment) is restricted to alternations and/or lexical synonymy. For this type of paraphrases, we present a *robust* and *wide coverage* system which assigns two such paraphrases one and the same semantic representation.

Robustness is achieved by using the Xerox Incremental Parser (henceforth XIP) which is based on layered grammars (grammars are ordered and the rules in each grammar can refer to the representation produced by the preceding grammars). The XIP system is robust in that it always delivers an output although the parse produced may be partial and underspecified in case a full analysis cannot be performed.

To achieve coverage, we integrate in the XIP grammars the detailed knowledge of alternations and of lexical synonymy encoded in large scale existing linguistic resources namely, VerbNet and WordNet.

The paper is structured as follows. We start (section 2) by presenting the linguistic resources used namely, VerbNet and WordNet. We then describe the type of semantic representations produced by our approach and illustrate the coverage achieved with some examples. In section 3, we show how to extend XIP to integrate VerbNet and WordNet information so as to assign paraphrases the same semantic representation. Section 4 presents an evaluation of the system based on a set of annotated examples extracted from VerbNet. Section 5 compares the approach presented here with related work

Members	*cascade(1), climb(4), crawl(), cut(), drop(), go(7), meander(1), plunge(), run(3), straggle(2), stretch(1), sweep(5), tumble(), turn(), twist(), wander(4), weave(4), wind(1 2)*
Thematic Roles and Selectional restrictions	Location[+concrete] Theme[+elongated]
Frames	Intransitive (+ path PP) "The river runs through the valley" Theme V Prep[+path] Location Prep(during(E),Theme,Location) exist(during(E),Theme)
	Locative Inversion "Through the valley meanders the river" Prep[+path] Location V Theme Prep(during(E),Theme,Location) exist(during(E),Theme)
	There-insertion "There meanders through the valley a river" there V Prep[+path] Location Theme Prep(during(E),Theme,Location) exist(during(E),Theme)
	There-insertion "There meanders a river through the valley" there V Theme Prep[+path] Location Prep(during(E),Theme,Location) exist(during(E),Theme)

Figure 1: VerbNet representation of the *meander-47.7* class

and 7 concludes with pointers for future work.

2 Lexical resources: VerbNet and WordNet

VerbNet is a broad-coverage domain-independent verb-lexicon (Kipper et al. 2000a) which encodes syntactic and semantic information for about 4000 english verbs. The verbs are organised in classes which refine Levin's classes (Levin, 1993) and capture generalisations about the regular association between syntactic and semantic verb properties. More specifically, a VerbNet class has the following components:

- The set of english verbs belonging to that class, each verb being annotated with the WordNet meaning(s) relevant to that class
- The set of theta roles which can be mapped to the arguments of these verbs
- Selectional restrictions on the arguments
- A set of frames consisting of an indentifier, an example, a syntactic description and a decompositional semantics common to all verbs in that class
- The set of superclasses of that class (the frames of these upper class are then inherited) if any.

Figure 1 pictures the VerbNet frame for the class *meander-47.7*.

For the treatment of paraphrases, the information contained in VerbNet is useful for several reasons. First, VerbNet documents the alternations of each verb such as those given in 1 and those illustrated by Figure 1 e.g., :

(2) 1. The river meanders through the valley
2. Through the valley meanders the river
3. There meanders through the valley a river
4. There meanders a river through the valley

By constructing the group of arity preserving alternations a verb participates in, it is then possible to identify all meaning preserving alternations a verb can occur in. Further since VerbNet associates with each verb class a thematic grid and a decompositional semantics, it becomes possible to develop a parser which, based on this knowledge (knowledge of the alternations of a verb and of its semantic representation) can build identical semantic representations for alternations paraphrases. Thus for instance, using the thematic role information associated in VerbNet with the *meander-47.7* class, all sentences in 2 can be assigned the same basic semantic representation :

River(R) & Valley(V) & Meander(E) & Location(E,V) & Theme(E,R)

For paraphrase recognition this is enough. For deeper semantic treatment involving inference for instance, the decompositional semantics might also be useful.

Another feature of VerbNet which makes it attractive for the treatment of paraphrases is its linking with WordNet. As indicated above, the verbs of a VerbNet class are annotated with the WorNet meaning(s) relevant to that class so that for instance, the verb *meander* in the *meander-47.7* class is the verb with meaning 1 in WordNet. Now because WordNet records the synonyms of a given word usage, this linking between VerbNet and WordNet also gives us access to synonymic paraphrases. In the case at hand for instance, the set of synonyms retrieved from WordNet for meaning 1 of *meander* is *weave, wind, thread, meander, wander*. By integrating this information in a parser lexicon and combining it with the knowledge of alternations given by VerbNet, we can thus obtain a parser which assigns one and the same semantic representation to the following sentences.

(3) 1. The river meanders through the valley
2. Through the valley meanders the river
3. There meanders through the valley a river
4. There meanders a river through the valley 5. The river weaves through the valley
6. Through the valley weaves the river
7. There weaves through the valley a river

8. There weaves a river through the valley
9. The river winds through the valley
10. Through the valley winds the river
11. There winds through the valley a river
12. There winds a river through the valley
13. The river threads through the valley
14. Through the valley threads the river
15. There threads through the valley a river
16. There threads a river through the valley
17. The river wanders through the valley
18. Through the valley wanders the river
19. There wanders through the valley a river
20. There wanders a river through the valley

3 Extending XIP to recognise VerbNet alternation paraphrases

In what follows, we show how the knowledge encoded in VerbNet can automatically be integrated in a robust parser thereby supporting the recognition of the set of alternation and/or lexically synonymic paraphrases covered by VerbNet. We start by presenting the parser used namely, the Xerox Incremental Parser (XIP). We then show how we extend it to deal with paraphrases.

3.1 XIP

Robustness that is, the ability to process real-world textual data is a important desiderata of natural language processing systems both from a theoretical and from a practical point of view. Theoretically it is important because testing theories with non artificial data is an requirement of the scientific enterprise and practically, because it is necessary in real world applications. As nicely summarised in [Ait-Mokhtar *et al.*, 2002], three main types of approaches to robustness can be distinguished: those based on deep grammars and using special mechanisms in order to recover an analysis when parsing fails or to rank analyses in case of overgeneration; those based on probabilistic parsing approaches and those adopting a shallow approach to parsing usually based on finite state techniques.

The XIP parser belongs to the third type of approach and garantees robustness by adopting incrementality: the input sequence is processed by a layered grammar, each grammar layer being applied sequentially. As the input is processed, it is either enriched or left unchanged – the output is the input sequence as annotated by the sequential application of the rules from the different layers. By ordering the grammar rules appropriately, data which is either infrequent or incorrect (e.g., sentences violating verb/subject agreement) can therefore be handled. It suffices to place the rules handling that data last. Since the data does not conform to the rules governing the most frequent data (which are placed early in the grammar layers), these rules do not fire and the rules governing the infrequent or incorrect data can apply.

Based on finite state techniques, the parser is also reasonably efficient running at a speed of 1 300 words per second on a Pentium II 50 where processing time includes tokenization, morphological analysis, part of speech disambiguation, chunking and dependency parsing.

We use XIP version 3.10 (2000-2001) as developed at the Xerox Research Europe. This version includes the NTM tokenizer and morphological analyser based on finite states technology, the HMM statistical POS tagger and a grammar for English which includes two types of subgrammars namely, chunking and dependency grammars.

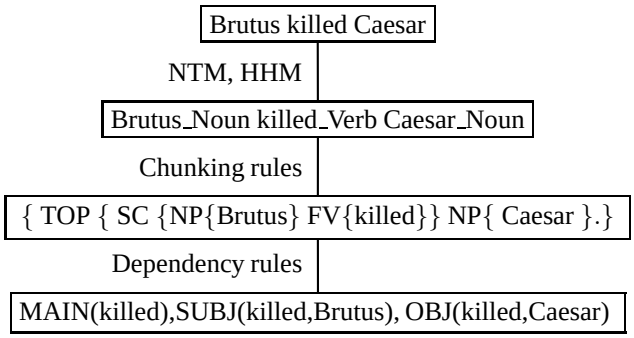

Figure 2: XIP representation of the sentence *Brutus killed Caesar.*

The **chunking grammar** describes constituency structure and consists of layered groups of chunking rules which are either ID/LP rules applying to partially ordered bags of nodes or sequence rules applying to ordered subsequences. The following rule, for example, builds an NP chunk if a sequence is found which consists of a determiner (Det) followed immediately or not by a noun.

$$15 > NP = Det, ?*, Noun.$$

Chunking rules are grouped into layers, each layer applying to the output of the preceding layers thus allowing for the production of chunking trees with depth more than one.

The **dependency grammar** supports the specification of (functional, thematic, semantic, anaphoric, etc.) relations between words or chunks and is based on the (layered) specification of groups of dependency rules of the form:

```
|<subtree_pattern>|
  if < conditions >
     <dependency_term >
```

`subtree_pattern` is a tree matching expression describing structural properties of part of the input tree, `conditions` is any Boolean expression built up from dependency terms, linear order statements and the conjunction, disjunction and negation operators and `dependency_term` is a term of the form `name<flist>(a1, ..., aN)` with `name` the name of a dependency relation, `flist` a list of features associated with that dependency relation and `a1, ..., aN` the relation arguments. Both `subtree_pattern` and `conditions` are optional.

The dependency rule given in Figure 3 for example, recognises a VCOMP dependency between two words #1 and #2 if #1 is the head of finite verb chunk (FV) that has trans value +

$| SC(?*, FV[trans:+]\{?*, \#1[last:+]\}\}, NP[time:]\{?*, \#2[last:+]\} | vcomp[dir:+](\#1, \#2)$

Figure 3: Example Xip dependency rule

(i.e., is transitive) and the FV is within an SC (clause chunk) followed by an NP chunk with a negative time feature the head of which is #2. The vcomp dependency is assigned the feature dir (for direct verb complement).

3.2 Incorporating Verbnet into XIP

To support a robust and large scale treatement of alternation paraphrases, we extended XIP with VerbNet information and with a semantic construction module that assigns alternation paraphrases one and the same semantic representation. Briefly, the idea is to integrate VerbNet information into a XIP lexicon and to then specify dependency rules which use this information together with the VerbNet set of lexico-syntactic patterns in order to assign a given input sequence the thematic grid assigned to that sequence by VerbNet. In what follows, we start by presenting the lexicon. We then describe the semantic construction module we added to XIP.

3.3 The verb lexicon

To integrate VerbNet information into XIP, we specified a lexicon which associates each verb with its VerbNet classes and with the WordNet Synset identifier corresponding to the relevant usage of that verb in that VerbNet class. For instance, the verb *meander* is assigned the following lexical entry:

```
meander:verb+=[meander-47.7, pred=c01828635].
```

The *VerbNet class* will be used both to guide syntactic parsing and to support semantic construction. Thus as we shall see in the following section, only those dependency rules whose antecedent mention the semantic class of the input verb will be triggered. Further, the VerbNet semantic class is used in the rule to specify the syntax/semantic interface that is, the pairing between syntactic and semantic arguments.

The *WordNet synset information* on the other hand serves to group together synonyms. That is, all verbs in a VerbNet class which belongs to the same WordNet synset will be assigned the same semantic representation. So for instance, the verbs *meander, wander, weave, wind, thread* in the VerbNet class *meander-47.7* will all be assigned a semantic information identical to that assigned to "meander".

The VerbNet class and synset assignment was made automatically on the basis of both VerbNet and WordNet information. At present, the lexicon contains 4225 verbs corresponding to 2779 WordNet synsets and 352 VerbNet verb classes. However since word sense disambiguation is not integrated in XIP, we only consider the most frequent meaning of a verb. In future, we intend to bypass this limitation by tagging the input verbs with meaning identifiers. Another more principled but less reliable way to remedy this shortcoming would consist in integrating a verb sense disambiguation module into XIP.

3.4 Semantic construction

To assign identical semantic representations to alternation paraphrases, we augment the XIP grammar with a set of *thematic grid dependency rules*. These rules assumes as input the output of the existing XIP parser that is, a representation of the input including both constituency and grammatical functions (subject, object, etc.) information. Based on this information, a thematic grid (dependency) rule identifies a given VerbNet pattern (syntactic frame and verb semantic class) and specifies a mapping between syntactic and thematic argument.

Let us illustrate this with a simple example. Suppose the sentence to be parsed is:

(4) The river meanders through the valley

As indicated in section 2, the VerbNet syntactic and semantic information associated with that usage of the verb *meander* is:

VerbNet class	*meander47-7*
Syntax	Theme V Prep[+path] Location

where the syntactic description abbreviates the following specification: the canonical subject is a theme and a prepositional object introduced by a path denoting preposition denotes the location of the event. In the XIP framework, such a specification is captured by the following (simplified) dependency rule:

```
if( (  MAIN(#1[coil9_6])
    || MAIN(#1[coil9_61])
    || MAIN(#1[escape51_1])
    || MAIN(#1[escape51_11])
    || MAIN(#1[escape51_12])
    || MAIN(#1[escape51_121])
    || MAIN(#1[meander47_7])
    || MAIN(#1[substance_emission43_4])
    || MAIN(#1[vehicle51_4_1])
    || MAIN(#1[vehicle51_4_11])
    || MAIN(#1[waltz51_5])
    )
     & VDOMAIN[passive:~](#1,#11)
     & SUBJ-SEM(#1,#2) & ~OBJ(#1,?)
     & VMOD[post](#1,#4) & PREPD(#3,#4)
     & (#3[vnpath])
  )
EVENT(#1),Theme(#1,#2),Location(#1,#4).
```

In words: if the main verb is associated (via lexical lookup) with one of the listed VerbNet classes (and in particular with the *meander47_7* class), if this verb is not in the passive mode, has no object but has a subject and a postposed verb modifier introduced by a path denoting preposition, then the semantic representation produced is EVENT(#1), Theme(#1,#2),Location(#1,#4) where #1, #2 and #4 are the nodes associated with the main verb, the subject and the modifier head respectively.

As this rule applies to the input sequence (4), the following representation is output where indeed the correct thematic representation has been produced.

```
EVENT(pred:C01870464_MEANDER47_7:+)
LOCATION(pred:C01870464_MEANDER47_7:+,valley)
THEME(pred:C01870464_MEANDER47_7:+,river)
```

More generally, the extended XIP grammar counts 425 thematic grid dependency rules. These rules are ordered by specificity, the most specific rules occurring first and the least specific last. For instance, the rules for ditransitives will be tested before the rules for transitives which again will be tested before the rules for intransitives. Since within one grammar layer only the first applicable rule is used, this ensures that the syntactic configuration captured by the rule that is executed is indeed the most appropriate (even though a rule describing an intransitive configuration correctly describes the syntactic configuration of a transitive sentence, the rule describing the transitive configuration describes it configuration more specifically and thus more accurately). This rule ordering also allows an appropriate treatment of the difference between adjuncts and subcategorised PPs. Being more specific, the rules describing verbs taking prepositional arguments will be tested before the general rules describing the combination of verbs with adjuncts and so will be preferred in case they can apply. Here is an illustrating example. Suppose we have the two sentences given in 5.

(5) a. Sharon shivered from fear.
EVENT(C01834682),CAUSE(C01834682,fear), EXPERIENCER(C01834682,Ann)
b. Sharon breakfasted in the garden.
EVENT(C01149559), AGENT(C01149559,Ann)

In the first sentence, the PP is described in VerbNet as an element of the subcategorisation frame of the verb *shiver* which is mapped to the CAUSE role. In contrast, in the second sentence the PP is treated as an adjunct and is not assigned a thematic role. By placing the rule describing the "shiver" configuration before the rule describing that of adjuncts, we can ensure that both sentences be assigned the correct thematic grid. In case the input is (5a), the "shiver" rule is first to apply thereby licensing the construction of the given semantic representation. For (5a) on the other hand, the "shiver" rule does not apply (because the verb "breakfast" does not belong to the same VerbNet class as "shiver") but the adjunct rule does which fails to assign the PP a thematic role[1].

To define the specificity ordering over the thematic rules, we first generalised the syntactic frames to 68 more general templates by ignoring prepositional and selectional information. For instance, the VerbNet syntactic frames `[NP, V NP Prep(of) NP]` and `[NP, V, Prep(with) NP]` were both abstracted to the more general template `[NP,V,Prep, NP]`. The resulting set of templates was then organised in a hierarchy (cf. figure **??**) which was then used to automatically order the XIP thematic rules.

[1]Of course the locative PP *the garden* should be assigned a semantic representation too and be related eg by a locative relation to the described event. We do not discuss this here as we are only concerned with correctly describing the thematic roles of the arguments of a verb as defi ned by VerbNet.

3.5 Postprocessing

To cover unknown input and more specifically verbs whose VerbNet class is not given in the lexicon, we introduce an additional postprocessing step which performs a default thematic grid assignment on the basis of the 68 abstract rule templates used to specify rule ordering. Specifically, these very general rule templates are used to specify 68 general rules describing general subcategorisation frames and assigning a default role to each of the arguments identified through those frames. For instance, suppose that the input sentence is 6 and that the VerbNet class for "stand" is not given in the lexicon. In such a case, the general rule specifying a syntactic configuration of the form `PP[`loc] V NP+, will assign the locative PP an `arg2` role and the NP an `arg1` role thereby producing the given semantic representation.

(6) a. "On the pedestal stood a statue
EVENT(stood_PRED:C02654415_PUT_SPATIAL9_21:+)
ARG1(stood_PRED:C02654415_PUT_SPATIAL9_21:+,statue)
ARG2(stood_PRED:C02654415_PUT_SPATIAL9_21:+,pedestal)

More generally, the postprocessing step will assign default underspecified thematic roles to maximal projection phrases occurring in the input on the basis of surface syntax information. Note that the use of underspecified thematic roles renders the obtained semantic representations similar to those assumed by PropBank [Kingsbury *et al.*, 2002].

4 Evaluation

To evaluate the extended XIP parser, we extracted from Verbnet the 1012 example sentences it contains together with their thematic role annotation. For instance, given the VerbNet representation of the *meander-47.7* class, we extracted the four following annotated examples:

(7) a. "The river runs through the valley"
Theme V Prep[+path] Location
b. "Through the valley meanders the river"
Prep[+path] Location V Theme
c. "There meanders through the valley a river"
there V Prep[+path] Location Theme
d. "There meanders a river through the valley"
there V Theme Prep[+path] Location

We then applied the parser to the resulting set of sentences and automatically compared the thematic grid output by the extended XIP parser with the thematic grid described by the VerbNet annotation. We obtained the following results:

- 71% of the sentences were assigned the correct representation (i.e. the same roles assignment as in VerbNet),
- 15% of the sentences were assigned the correct syntactic pattern but the wrong theta roles because selectional restrictions could not be checked

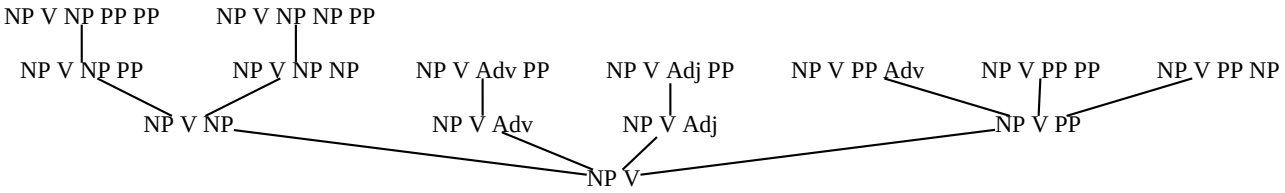

Figure 4: Hierarchy of syntactic patterns in Xip

- 4% of the sentences were assigned a default pattern because either the tagger was not able to recognize the class appartenance of the verb, or the verb class assignment in the lexicon did not allow the syntactic pattern illustrated by the given sentence.
- 10% of the sentences could not be mapped onto a VerbNet pattern because the constituency information delivered by XIP was insufficient.

The problem with selectional restrictions can be illustrated by the following example. In Verbnet the verb "buy" is assigned (among others) the following two frames :

(8) a. a. Basic Transitive
"Carmen bought a dress"
Agent V Theme

b. b. Sum of Money Subject Alternation (Asset Subject)
"$50 won't even buy a dress at Bloomingdale's"
Asset V Theme

Thus, without knowledge about the ontological type of the arguments it is impossible for the parser to decide whether the subject should be assigned an AGENT or an ASSET thematic role. In other words, for 15% of the VerbNet data ontological knowledge is required in order to correctly determine the thematic grid of the input sentence. Such knowledge could be integrated into XIP by resorting e.g. to the WordNet hierarchy which as we saw in section 2 is linked to the verb usages described by Verbnet.

Another 10% of the failure cases is due to incorrect parsing and could be reduced by improving the XIP grammar. The remaining 4% requires improving both the tagger (in case, the verb is not tagged appropriately) and the VerbNet description (in case a syntactic pattern is missing for a given verb usage).

In sum, the evaluation shows that the robust parser developed deals appropriately with 71% of the VerbNet data and that there is reasons to hope that it can be further improved by incorporating selectional restrictions and by improving the basic constituency grammar.

5 Related work

Recent years have witnessed a strong interest in paraphrase recognition. In what follows, we compare our approach with the statistical approaches used in information extraction (IE) and question answering (QA) as well as with two related symbolic approaches, one based on Verbnet and XTAG and the other, based, like ours, on XIP.

5.1 Statistical approaches

Because of the large, open domain corpora IE and QA systems deal with, coverage and robustness are key issues and much on the work on paraphrases in that domain is based on automatic learning techniques. For instance, [Lin and Pantel, 2001] acquire two-argument templates (inference rules) from corpora using an extended version of the distributional analysis in which paths in dependency trees that have similar arguments are taken to be close in meaning. Similarly, [Barzilay and Lee, 2003] and [Shinyanma *et al.*, 2002] learn sentence level paraphrase templates from a corpus of news articles stemming from different news source. And [Glickman and Dagan, 2003] use clustering and similarity measures to identify similar contexts in a single corpus and extract verbal paraphrases from these contexts.

The differences between such approaches and the approach presented here are those generally occurring between symbolic and statistical approaches. First, the coverage differs. Whereas the approach presented here concentrates on alternation paraphrases and verbal synonymy, statistical approaches either take a very general way of identifying paraphrases grouping together e.g., synonyms, hyperonyms and sibblings [Barzilay and McKeown, 2001] or are specialised to a few chosen relations occurring frequently in a given domain [Ravichandran and Hovy, 2002]. Second, the resources used differ in that statistical treatments of paraphrases rely on aligned related corpora, on corpora treating of the same topic or on extremely large corpora (the web). By contrast, our approach is based on existing symbolic resources namely VerbNet and WordNet. Third, the output of the two approaches differ in that statistical approaches typically yield a "paraphrase lexicon" that is, a list of paraphrases which is independently put to work in a given application by some string manipulation procedure. In contrast, our output is a parser designed to handle alternation paraphrases.

5.2 Combining VerbNet with XTAG

In [Ryant and Kipper, 2004], Verbnet is related to XTAG, a lexicalised Tree Adjoining Grammar of English which covers the possible transformations of canonical frames. Specifically, a mapping is specified between VerbNet frames and XTAG tree families whereby each VerbNet frame is mapped to a corresponding XTAG tree family. Since an XTAG tree family specifies the possible transformations (active, passive, extrapositions, etc.) of a given syntactic frame, this mapping in effect extends the coverage of VerbNet beyond canonical frames to all transformations of these canonical frames.

In comparison, the approach proposed here only deals (so far) with the canonical version of the VerbNet frames. It is therefore syntactically more limited and we intend to extend the current version of our system to take into account grammatical variations. On the other hand, the two approaches are based on different grammar frameworks (XTAG versus layered grammars) and different parsing techniques (suppertagging and parsing for XTAG versus cascaded finite state automata for XIP) thus yielding interesting differences. For instance, the layered grammars supported by XIP are not subject to the very strict argument/modifier enforced by TAG elementary/auxiliary trees. As a result, the treatment of alternations involving adjuncts (such as *Induced Action* in "Tom jumped the horse over the fence") is unproblematic in XIP whilst certain frames in VerbNet (namely the Transitive(+here/there) construction) simply have no XTAG mapping. Another difference concerns the disambiguation module used. Whilst in XIP a unique output is determined based on rule ordering (the first applicable rule is applied), in XTAG disambiguation is done using supertagging that is, using n-grams. It is worth noting though, that whilst an evaluation of the XTAG parser on the data examined in this paper is in principle possible, no such evaluation has yet been carried out so that a comparison of the two approaches cannot be made. It would be interesting to see how they compare and in particular to compare the results both of the two disambiguation modules and of the two grammars used.

5.3 An alternative XIP based approach

[Hagège and Roux, 2003] also presents a XIP based treatment of paraphrases which uses XIP dependency rules to construct a "normalised" representation of functional dependencies. For instance, given a passive sentence, the subject will be labelled as a normalised object and the indirect object as a normalised subject. The approach is both more general and less exhaustive than ours. It is more general in that it covers more types of paraphrases than ours including in particular, syntactic variations, alternations and morphoderivational equivalences. On the other hand, alternations are handled in a restricted way in that only certain types of alternations are covered. By contrast, the approach we propose only deals with alternation paraphrases but does so in an exhaustive manner by importing in the grammar the complete knowledge encoded in VerbNet together with some of the lexical synonymic knowledge encoded in WordNet. As discussed in section 6, we intend to extend our approach to syntactic variations and morphoderivational equivalences so that the final model should have a coverage similar to that proposed by [Hagège and Roux, 2003] but integrates the available linguistic knowledge in a more systematic fashion.

6 Perspectives

The work presented here is a first step towards a robust symbolic treatment of paraphrases. To improve coverage however, much remains to be done.

For a start, the approach needs to be extended to non canonical variants. Indeed the evaluation is currently restricted to canonical alternation variants that is, sentence without extraposition or pronominalisation for instance. Because the basic XIP grammar does produce functional dependencies (subject, object, etc.) and because the thematic rules used for semantic construction rely in part on such dependencies, it is likely that the treatment of alternations presented here straightforwardly extend to non canonical variants. However this needs to be tested on a systematic testsuite. To this end we are currently exploring techniques for automatically building non canonical variants of the VerbNet sentences.

Another needed extension concerns the treatment of other types of paraphrases such as those produced using intercategorial synonymy (9a), morphoderivational variants (e.g., nominalisation, 9b), converse constructions (9c) and antonyms (9d).

(9) a. John stopped smoking/Jean no longer smokes.

b. The cost of the cruise is high/The cruise costs a lot.

c. John lent a book to Marie / Marie borrowed a book from John.

d. John is slow/John is not fast.

For morphoderivational variants, linguistic resources such as Celex (HTTP://WWW.RU.NL/CELEX/) exists which could be integrated into XIP in a manner similar to the integration of VerbNet information. For antonyms, WordNet could be drawn upon to provide a treatment similar to that of lexical synonyms. For converses and intercategorial synonymy, it is less clear however where the linguistic resources would come from.

Finally, a research trend has recently emerged which aims at determining whether one sentence is more general than another or in other words, whether one sentence is entailed by another. Because the VerbNet classes are often semantically homogeneous classes, the approach proposed here provides a handle on entailment between verbs. Thus, consider again the verbs belonging to the *meander-47.7* VerbNet class :

cascade(1), climb(4), crawl(), cut(), drop(), go(7), meander(1), plunge(), run(3), straggle(2), stretch(1), sweep(5), tumble(), turn(), twist(), wander(4), weave(4), wind(1 2)

As in Levin's work, the VerbNet grouping of verbs into classes is based on syntactic criteria: verbs sharing the same set of alternations are grouped together. However the driving idea is that shared syntactic properties reflect shared semantic properties so that in effect, verbs belonging to the same VerbNet class are likely to be semantically similar. This is particularly clear in the *meander-47.7* class presented above where the verbs of that class are very close semantically. More specifically, the *meander-47.7* class contains motion verbs which differ from each other with respect to speed, path of motion or selectional restrictions.

Now, since each verb in a VerbNet class is labelled with the intended WordNet usage, it is possible to use the WordNet hierarchy to order the set of verbs included in a VerbNet class with respect to hyperonymy and troponymy. By using the concept hierarchy thus defined, entailment between verbs can then be checked. We are currently investigating this research direction, examining how such information is best integrated into XIP and how the resulting system can be evaluated.

7 Conclusion

In this paper, we have shown how to integrate the linguistic knowledge of alternations encoded in VerbNet and (some of) the verbal lexical synonymy information encoded in WordNet into a robust parser thus yielding a parser that can assign alternation paraphrases one and the same semantic representation. A first evaluation shows that the parser has an accuracy of 71%.

Current and future work concentrates on (i) improving the current results by improving the grammar and integrating selectional restrictions, (ii) extending the paraphrastic coverage by considering additional paraphrasing mechanisms such as morphoderivational variants, cross categorial synonyms and antonyms and (iii) evaluating the system on real text corpora.

An additional line of research concerns the usability of the existing parser for automatically tagging real text either with VerbNet thematic grid information using the detailed thematic grid dependency rules or with less specific PropBank thematic grid information by resorting to the less detailed rule templates used in the postprocessing step.

Acknowledgments

We would like to thank the Xerox company for making XIP available to us. We would also like to thank the Contrat Plan Etat Région : Ingénierie des Langues, du Document et de l'Information Scientifique, Technique et Culturelle for partially funding the research presented in this paper.

References

[Ait-Mokhtar *et al.*, 2002] S. Ait-Mokhtar, J P. Chanod, and C. Roux. Robustness beyond shallowness: incremental deep parsing. *Natural Language Engineering*, 8(2/3):121–144, 2002.

[Barzilay and Lee, 2003] R. Barzilay and L. Lee. Learning to paraphrase: an unsupervised approahc using mutliple-sequence alignment. In *Proceedings of NAACL-HLT*, 2003.

[Barzilay and McKeown, 2001] Regina Barzilay and Kathleen McKeown. Extracting paraphrases from a parallel corpus. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, Toulouse, 2001.

[Glickman and Dagan, 2003] O. Glickman and I. Dagan. Identifying lexical paraphrases from a single corpus: a case study for verbs. In *Proceedings of Recent Advances in Natural Language Processing*, 2003.

[Hagège and Roux, 2003] C. Hagège and C. Roux. Entre syntaxe et sémantique : normalisation de la sortie de l'analyse syntaxique en vue de l'amèlioration de l'extraction d'information à partir de textes. In *TALN*, Batz-sur-Mer,France, 2003.

[Kingsbury *et al.*, 2002] Paul Kingsbury, Martha Palmer, and Mitch Marcus. Adding semantic annotation to the penn treebank. In *In Proceedings of the Human Language Technology Conference*, San Diego, California, 2002.

[Lin and Pantel, 2001] Dekang Lin and Patrick Pantel. Discovery of inference rules for question answering. *Natural Language Engineering*, 2001.

[Mel'čuk, 1988] I. Mel'čuk. Paraphrase et lexique dans la thorie linguistique sens-texte. *Lexique*, 6:13–54, 1988.

[Ravichandran and Hovy, 2002] Deepak Ravichandran and Eduard Hovy. Learning surface text patterns for a question answering system. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 41–47, Philadelphia, 2002.

[Ryant and Kipper, 2004] Neville Ryant and Karin Kipper. Assigning xtag trees to verbnet. In *Proceedings of the 7th International Workshop on Tree Adjoining Grammar and Related Formalisms (TAG+7)K*, Vancouver, British Columbia, Canada, 2004.

[Shinyanma *et al.*, 2002] Y. Shinyanma, S. Sekine, K. Sudo, and R. Grishman. Automatic paraphrase acquisition from news articles. In *Proceedings of HLT*, 2002.