Using inheritance and coreness sets to improve a verb lexicon harvested from FrameNet

Mark McConville and Myroslava O. Dzikovska

School of Informatics, University of Edinburgh
2 Buccleuch Place, Edinburgh EH8 9LW, Scotland
{Mark.McConville,M.Dzikovska}@ed.ac.uk

Abstract

We investigate two aspects of the annotation scheme underlying the FrameNet semantically annotated corpus — the inheritance relation on semantic types with its corresponding links between semantic roles of increasing granularity, and the specification of coreness sets of related semantic roles — against the background of our ongoing effort to harvest a lexicon of verb entries for deep parsing. We conclude that these aspects of the FrameNet annotation scheme *do* prove useful for reducing the complexity and ambiguity of verb entries, allowing for semantic roles of lower granularity for purposes of deep parsing, but need to be applied more systematically to make the lexicon usable in a practical parsing system.

1 Introduction

Semantically annotated corpora and wide-coverage semantic lexicons are an important resource for building NLP systems. They have been used to train shallow semantic parsers (Gildea and Jurafsky, 2002), provide paraphrases in question answering (Kaisser and Webber, 2007), and extend lexicons for deep parsing (Crabbé et al., 2006). All these applications use a 'frame-based' representation to express sentence semantics, where the semantic *type* corresponding to the meaning of a verb is related to its dependents by means of semantic *roles*. An essential task in building this representation is to make a connection between the surface form of the utterance and its semantics, usually by linking between syntactic and semantic structure.

Linking syntactic and semantic structure can be facilitated by a computational lexicon that describes possible mappings. McConville and Dzikovska (2007) report on an attempt to harvest a verb lexicon for deep linguistic processing from the FrameNet 1.3 semantically annotated corpus. We demonstrated that harvesting verb entries directly from annotations, as is done in the lexical entry files currently distributed with FrameNet, results in a number of subcategorisation frames which are unsuitable for inclusion in a computational lexicon used by a deep parser. We proposed a set of filtering rules to reduce the number of spurious subcategorisation frames generated by syntactic phenomena not directly captured in the FrameNet annotation.

In this paper we evaluate how this lexicon can be further improved by using two other aspects of the linguistic annotation underlying the corpus — the organisation of the semantic types (a.k.a. 'frames') and roles ('frame elements') into a hierarchy, and the specification of certain 'coreness sets' of related roles. The FrameNet ontology is very expressive and richly structured, with the aim of simplifying a number of reasoning tasks. However, we argue that FrameNet's level of role name granularity creates problems from the perspective of parsing, since it is traditionally assumed that verbs subcategorise for a relatively small number of arguments.

We first of all demonstrate that it is possible to use role inheritance to reduce the size of the role set (and hence the lexicon as a whole) without losing information, thus restricting the granularity of the semantic roles used in the output representation. We then describe an attempt to apply the coreness sets defined in the FrameNet ontology to eliminate ambiguity in lexical entries, making the FrameNet-based lexicon easier to use in a parsing system. We conclude that the FrameNet annotation scheme provides for useful mechanisms for reducing the complexity and ambiguity of verb entries, but needs to be applied more systematically to make the lexicon usable in a practical parsing system.

Section 2 provides some necessary background. **Section 3** discusses our investigations into the use of semantic role inheritance to reduce the vocabulary of roles invoked by arguments in verb entries. **Section 4** then turns to the topic of coreness sets in FrameNet, and the extent to which they can be used to eliminate redundancy in the harvested lexicon. Finally, **Section 5** discusses how our algorithms could be used in the future to benefit applications other than deep parsing.

2 Background

Regardless of the particular grammar formalism which they presuppose, lexicons used for parsing and semantic interpretation contain representations that map syntactic structure (a subcategorisation frame or a set of syntactic roles) to semantic structure (a predicate name and a set of arguments). For example, a lexical entry for the verb *move* would specify that: (a) the verb invokes a predicate which we might call 'motion'; (b) it subcategorises for a noun phrase subject which denotes the 'theme' (i.e. the object undergoing movement); and (c) it also subcategorises for a prepositional phrase complement headed by the preposition *to* which denotes the 'goal' (i.e. endpoint of the trajectory). This kind of information can be harvested automatically from semantically annotated corpora such as FrameNet (Baker et al., 1998), PropBank (Palmer et al., 2005) or OntoNotes (Hovy et al., 2006). The ultimate goal of our

project is to create a wide-coverage lexicon yielding representations that can be connected to the reasoning engine of a dialogue system. Thus, we chose FrameNet as our source for extracting lexical entries, since it includes an ontology which has already proved useful for information retrieval and question answering tasks (Surdeanu et al., 2003; Kaisser and Webber, 2007).

The FrameNet annotation scheme allows one to harvest a lexicon by reading the subcategorisation frames and their corresponding role assignments directly off the annotated sentences. The resulting lexicon contains 2,770 verb entries, each specifying a semantic type, an orthographic form, and a set of subcategorisation frames. Subcategorisation frames are sets of arguments, each of which specifies a syntactic role, syntactic category and semantic role.[1] Here is an example lexical entry for the verb *fry*, derived from an annotated sentence like *Matilde fried the catfish*:

$$\begin{bmatrix} \text{ORTH} & \langle \textit{fry} \rangle \\ \text{CAT} & \text{V} \\ \text{TYPE} & \texttt{Apply\_heat} \\ \text{ARGS} & \left\langle \begin{bmatrix} \text{ROLE} & \texttt{Ext} \\ \text{CAT} & \text{NP} \\ \text{ROLE} & \texttt{Cook} \end{bmatrix}, \begin{bmatrix} \text{ROLE} & \texttt{Obj} \\ \text{CAT} & \text{NP} \\ \text{ROLE} & \texttt{Food} \end{bmatrix} \right\rangle \end{bmatrix}$$

The subcategorisation frame lists two arguments, one for each annotated dependent in the sentence.

While collecting such entries is straightforward on the surface, not all of them would be usable with a deep parser. To begin with, all entries have to correspond to "canonical" syntactic subcategorisation frames - i.e. to indicative mood, direct word order entries, and include only syntactic complements but not modifiers. Entries for other constructions, such as passives and clefts, are normally derived by syntactic transformations and are not included in the lexicon. We addressed these issues previously (McConville and Dzikovska, 2007; McConville and Dzikovska, 2008), developing methods to remove such spurious entries from the lexicon.

Secondly, we need to consider how well the syntax-semantics mappings harvested from the corpus fit with the representations traditionally used for parsing. We observed that the representations in the extracted entries manifest at least one significant difference in this respect. While there is no easily definable "canonical" representation for semantic roles, deep parsers, generally speaking, assume that the target semantic representation utilises a relatively small set of roles. There are several reasons for this. Firstly, restricting the vocabulary of semantic roles is convenient from a representational perspective — many existing lexicons are hierarchical (Copestake and Flickinger, 2000; McConville, 2006), and having a large number of distinct roles may make the lexicon less compact because it offers fewer opportunities for re-use through inheritance. Secondly, it has been proposed that the syntactic and semantic behaviour of verbs is correlated (Levin, 1993), and can be mediated through a small set of 'thematic roles', as for example encoded in the VerbNet lexicon (Kipper et al., 2000).

Finally, disambiguating between a large number of roles may require world knowledge and pragmatic information which is difficult to obtain and integrate in a domain-independent way. For example, the FrameNet semantic type `Closure` defines two distinct roles which can be denoted by the direct object of a transitive verb: `Container_portal` (e.g. *John closed **the tent flap***), and `Containing_object` (e.g. *Mary buttoned **her coat***). Human annotators are able to distinguish these roles based on common sense knowledge, and whilst it is true that such distinctions may be important for certain reasoning tasks, a deep parser would find this kind of ambiguity extremely difficult to resolve. Thus, a more compact roleset may be necessary to reduce the ambiguity in parsing and semantic interpretation.[2]

The importance of having a relatively small set of basic semantic roles has not been lost on the creators of FrameNet. Indeed, a lot of recent effort (between versions 1.1 and 1.3) has gone into organising the semantic types in the FrameNet ontology into an inheritance hierarchy and, in particular, into linking the fine-grained roles of child types with the more generic roles of their parent types. In addition, a number of 'coreness sets' of semantic roles have been specified, the idea being that only one member of a coreness set need be explicitly invoked in a well-formed, non-elliptical sentence, and hence that these roles are equivalent in some way. In the rest of this paper we describe how we used inheritance and coreness sets to eliminate redundancy in both the vocabulary of semantic roles and in the verb entries themselves.

As our general evaluation metric, we take the reduction in the number of individual roles and the reduction in the number of subcategorisation frames per verb entry in the lexicon. For comparison, we looked at two other lexicons: VerbNet, a lexicon of English verbs that aims to have a complete coverage of syntactic alternations for each verb covered, and the TRIPS lexicon (Allen et al., 2007) — a multi-domain lexicon used with a wide-coverage deep grammar. These lexicons were developed independently, but share the aim of explicitly representing the connections between syntax and semantics, with VerbNet focusing more on complete coverage, and TRIPS focusing on practical parsing applications that require syntactic and semantic disambiguation. Thus, while there is no way of determining the 'ideal' number of roles per se, comparison with these lexicons can give us some insight in the complexity or redundancy of the FrameNet-based lexicon compared to lexicons intended for parsing.[3]

The initial lexicon harvested from FrameNet (McConville

[1] We extracted this lexicon independently, but FrameNet contains an analogous set of lexical entries as part of the distribution, which we could have used as a starting point in the same way.

[2] Additional information can be brought in at a post-processing stage, linking the more generic semantic representation with more specific knowledge representation (Dzikovska et al., 2007).

[3] The various lexicons are not completely independent, in the sense that TRIPS contains an ontology of concepts inspired by an early version of FrameNet (Dzikovska et al., 2004), and it contains entries extracted from VerbNet (Crabbé et al., 2006). However, all entries were hand-edited to ensure that they conform to the independently developed lexicon design.

and Dzikovska, 2007) contains 9,180 subcategorisation frames, invoking 362 distinct semantic types, and arguments invoking 441 distinct semantic role labels, an average of 1.2 semantic role labels per semantic type. In comparison with other deep verb lexicons, this ratio of roles to types is quite high. The TRIPS lexicon contains verb entries invoking 284 distinct semantic types and arguments invoking 48 distinct semantic roles, yielding a ratio of 0.17 roles per semantic type. Similarly, the VerbNet lexicon has 395 verb classes, with arguments instantiating just 33 distinct semantic/thematic roles, giving a ratio of 0.084 roles per verb class. In addition, the FrameNet-based lexicon contains 3.3 subcategorisation frames per verb entry, compared to 2.8 in VerbNet and 1.3 in TRIPS.[4]

3 Using inheritance to reduce the role set

We first consider how the inheritance relation encoded in the FrameNet ontology can be used to reduce the size of the vocabulary of semantic roles.

The FrameNet ontology of semantic types is organised into an inheritance hierarchy, where child types are connected to their parents by means of an `Inheritance` relation. For example, this relation partitions the `Motion` semantic type (encoding events involving a theme traversing a path) into a number of more specific subtypes such as `Self_motion` (the theme is a living being, acting under its own volition), `Fluidic_motion` (the theme is a fluid), etc. All the semantic roles associated with a parent type must be implemented by some role of each child type. For example, two of the roles associated with `Motion` are `Source` (start of the trajectory) and `Goal` (end of the trajectory). These roles are implemented directly by all child types of `Motion` using roles of the same name. On the other hand the `Theme` role associated with the `Motion` type is implemented by different roles in subtypes: in `Self_motion` it is implemented by `Self_mover`, in `Fluidic_motion` by `Fluid`, and so on. In addition, child types can introduce new roles which are *not* linked to roles of parent types.

The existence of this inheritance relation and its associated links between parent and child roles has important implications for the vocabulary of semantic roles in the lexicon we harvested from FrameNet. For example, the transitive verb *dismiss* invokes the FrameNet semantic type `Firing`, and its subject and object instantiate the associated semantic roles `Employer` and `Employee` respectively, hence the following subcategorisation frame:

(1) `Sbj:Employer Obj:Employee`

However, the semantic type `Firing` is subsumed by the parent type `Intentionally_affect` in the FrameNet ontology, with the `Employer` role linked to the superrole `Agent` and the `Employee` role linked to the `Patient` superrole. Thus, an alternative way of representing the transitive subcategorisation frame for *dismiss*, using the information contained in the inheritance hierarchy, is:

(2) `Sbj:Agent Obj:Patient`

Note that the semantic roles specified in this lexicon are much more generic, and are similar to the kinds of role names used in the VerbNet and TRIPS lexicons.

The aim of the first part of our project was to investigate the extent to which we can use information about supertypes and 'superroles' in the FrameNet 1.3 ontology to decrease the number of distinct semantic roles invoked by arguments in the harvested lexicon, thus creating a less redundant verb lexicon for deep parsing.

3.1 Methodology

We went through each argument of each subcategorisation frame of each verb entry in the harvested lexicon and, where the entry's semantic type was linked to some parent type in the FrameNet ontology and the argument's semantic role was linked to some role of the parent type, we replaced the original role with the superrole. We repeated this until we reached the root type in the ontology, which in this case involved five cycles (i.e. the maximum depth of the relevant part of the inheritance hierarchy is 5). In the cases where a role is linked to two or more distinct superroles (because of multiple inheritance in the FrameNet ontology), we included all of them.

3.2 Results

The results are presented in Table 1 in the 'full lexicon' column. Each row represents a level of recursion, i.e. '0' means that no supertypes are taken into account, '1' means that we move one level up the hierarchy etc. The first column represents the number of distinct semantic role labels across the entire lexicon at each cycle, and the second column represents represents the number of distinct types of subcategorisation frame in the lexicon (where a subcategorisation frame is abstracted to a set of semantic roles).

Thus, taking the lexicon we harvested from FrameNet as a whole, we can reduce the number of distinct semantic role labels by 21%, from 441 to 347. The five most common roles which are the beneficiaries of this process are presented in Table 2.

Note that the number of distinct role labels, 347, still appears to be very high in comparison with the selection found in other deep verb lexicons like TRIPS and VerbNet. In addition, Table 2 demonstrates that, although the three most popular roles to be introduced are the generic roles `Theme`, `Patient` and `Agent`, familiar from both the VerbNet and TRIPS lexicons and from mainstream theories of thematic roles, there are still some overly specific roles in evidence, for example `Communicator` and `Sought_entity`,

We hypothesised that the very small reduction in the number of semantic roles is a function of the incomplete nature of the inheritance relation in the FrameNet ontology. Recall that the FrameNet 1.3 ontology contains 362 verbal types. However, a large proportion of these, 145, are 'orphan types', in the (strong) sense that they are not linked to any other type in the ontology, neither as child nor as

[4]Note that the TRIPS figure is significantly lower in part because the TRIPS lexicon has been built based on the subcategorisation frames attested in spoken dialogue corpora, so it does not contain many frames that are included in VerbNet but only rarely appear in speech and dialogue.

	full lexicon		restricted lexicon	
cycle	roles	frames	roles	frames
0	441	1256	289	807
1	364	1129	196	653
2	348	1083	177	596
3	347	1083	176	596
4	347	1083	176	596
5	347	1083		

Table 1: Results of the inheritance experiments

full lexicon		restricted lexicon	
frequency	role	frequency	role
1254	`Theme`	1843	`Agent`
1150	`Patient`	1486	`Theme`
777	`Agent`	1189	`Patient`
709	`Communicator`	827	`Communicator`
225	`Sought_entity`	591	`Goal`

Table 2: Most common role labels in the resulting lexicon

parent. In order to determine whether the disappointingly small reduction in distinct semantic roles as we climb the hierarchy is a result of the existence of these orphan types, we eliminated all verb entries from the harvested lexicon which invoke one of the 145 orphan types, and repeated the process.

Our restricted lexicon now contains 1,729 verb entries invoking 217 distinct semantic types. There are 6,253 subcategorisation frames distributed across these entries. The results of substituting more general roles for more specific ones, according to the inheritance relation underpinning the FrameNet 1.3 ontology, are presented in the 'restricted lexicon' half of Table 1.

The five most common roles which are now the beneficiaries of this process are presented on the right hand side of Table 2.

Thus, assuming the subset of the FrameNet-harvested lexicon which only includes types which are incorporated into the inheritance relation underpinning the FrameNet 1.3 ontology, we can reduce the number of distinct semantic role labels by 39%, from 289 to 176. This is significantly higher than the 21% reduction we managed using the full lexicon, thus supporting our hypothesis that the more 'connected' the FrameNet inheritance relation is, the more useful it will be in allowing us to harvest a deep verb lexicon with a manageable set of semantic roles. The fact that only 975 of the 2,770 verb entries in the harvested lexicon have semantic types which are rooted in either the `State` or `Event` supertypes shows that the FrameNet ontology still has a way to go in this respect.

4 Using coreness sets to filter subcategorisation frames

As discussed in the introduction, after filtering out modifiers and frames derived from non-canonical usages of target verbs, the lexicon we harvested from FrameNet contained 9,180 subcategorisation frames, distributed among 2,770 verb entries.

One interesting feature of the FrameNet ontology which we have not considered until now involves the specification of certain kinds of dependency between the semantic roles associated with a particular semantic type. For example, in certain semantic types, a particular subset of the semantic roles may be grouped together in a 'coreness' set, only one of which need be expressed in order to produce a complete, non-elliptical sentence. The most prevalent example of this involves the following semantic roles within the `Motion` semantic type and its subtypes:

- `Source` (e.g. *from Cairo*)
- `Goal` (*to Khartoum*)
- `Path` (*down the Nile*)
- `Area` (*around the country*)
- `Direction` (*towards Alexandria*)

The fact that these five roles are grouped together into a coreness set, captures the fact that they are in some sense equivalent, or that they instantiate the same underlying role, that of "trajectory".

The existence of coreness sets has implications for lexical concision. For example, the harvested lexicon contains 115 entries invoking the `Self_motion` semantic type, and these entries involve *eleven* distinct types of subcategorisation frame (ignoring syntactic categories) with the `Self_mover` role as subject and these 'trajectory' roles as oblique dependents, for example:

(3) `Sbj:Mover Dep:Source`
`Sbj:Mover Dep:Goal`
`Sbj:Mover Dep:Source Dep:Goal`
...

However, if we assume that the trajectory roles are actually just alternative realisations of the same underlying semantic role, then we can condense all these frames into just the one, where the Kleene star denotes an unbounded number of instances of the specified argument type:

(4) `Sbj:Theme Dep:Trajectory*`

The FrameNet 1.3 ontology specifies 210 coreness sets for 174 verbal semantic types. Each coreness set brings together an average of 2.5 semantic roles. The aim of the second part of our project was thus to investigate to what extent we can use the coreness sets defined in the ontology to consolidate the harvested lexicon, in terms of reducing the number of subcategorisation frames that need to be specified.

4.1 Methodology

We proceeded in two stages. First of all, we went through every argument of every subcategorisation frame of every verb entry and, where the argument's semantic role was part of some relevant coreness set, we replaced the semantic role name with the coreness set name. Then we went through every verb entry and eliminated duplicate frames, assuming that two frames are identical if and only if they have the same arguments, and that two arguments are identical just in case they have the same syntactic role, syntactic category and semantic role/coreness set.

4.2 Results

The first stage of the procedure, where we replaced semantic role labels with relevant coreness sets, affected 1,542 of the 2,770 verb entries in the lexicon, and 5,954 of the subcategorisation frames found in these entries. After eliminating duplicate subcategorisation frames, we were left with 7,804 frames across the lexicon as a whole (down from 9,180).

Of the 7,804 subcategorisation frames left in the lexicon, 1,253 have potentially duplicate arguments, i.e. where two or more arguments have semantic roles from the same coreness set. Thus, we next eliminated all duplicate arguments from individual subcategorisation frames, resulting in a decrease in the total number of arguments across all extant subcategorisation frames, from 16,795 to 16,406. Finally, after again eliminating duplicate subcategorisation frames from within each verb entry, the lexicon contained 7,672 frames across the 2,770 verb entries. This constitutes an average of 2.8 subcategorisation frames per entry and a reduction of 16% on the original number of 9,180.

4.3 Evaluation

We wanted to evaluate whether the use of coreness sets to consolidate pairs of subcategorisation frames corresponds with linguistic intuitions about which subcategorisations frames in a verb entry are really 'equivalent' and hence 'collapsible'. To this end, we selected 100 random cases where our procedure had used coreness sets to make a judgment that two distinct subcategorisation frames were essentially the same. We ensured that our sample contained only one instance from each semantic type, so as to counteract the bias in the FrameNet corpus whereby certain types include more verbs than others and certain verbs have been more fully annotated. Where necessary, we referred to the equivalent verb entries in VerbNet and the TRIPS lexicon.

Of the 100 entries chosen, 17 involved variations of the 'trajectory' coreness set discussed above, associated with an assortment of motion, orientation and spatial extension predicates. It is important to note, first of all, that this coreness set is independently motivated, for example in the ontology of paths outlined in Jackendoff (1983), where source, goal, and other unbounded path expressions are treated as equivalent in the sense that they are alternative realisations of one and the same thematic role in conceptual structure. We verified that in all 17 cases, the coreness set *did* in fact correlate with this linguistic intuition, and hence that combining the two subcategorisation frames was valid. Take for example, the following subcategorisation frames of the verb *buzz* from the `Motion_noise` semantic type:

(5) `Sbj:NP:Theme Dep:PP:Goal`
`Sbj:NP:Theme Dep:PP:Path`

The first of these includes a `Goal` argument (e.g. *buzz into the room*) and the second a `Path` (e.g. *buzz across the room*). Since the FrameNet ontology lists these in a coreness set for `Motion_noise`, the two subcategorisation frames are combined into the following unified representation:

(6) `Sbj:NP:Theme Dep:PP:Goal/Path`

This decision corresponds with our linguistic intuitions about the argument structure of the verb *buzz*, which subcategorises for an unbounded number of trajectory expressions (e.g. *The fly buzzed from the doorway across the room to the window*). We used similar reasoning with the other 16 instances involving the 'trajectory' coreness set in our sample.

Of the remaining cases in our sample, a substantial number (around 40) involve what can loosely be termed 'part-whole' alternations in the relevant argument. For example, the verb *claw* from the `Manipulation` type subcategorises for subjects with two distinct semantic roles, `Agent` and `Bodypart_of_agent`, related through a coreness set. These two usages are exemplified in the following two sentences:

(7) Jane clawed at his back
Fingers clawed at his back

Other examples are somewhat more abstract. For example, the verb *eclipse* from the `Surpassing` type subcategorises for two kinds of subject in the FrameNet lexicon, `Profiled_item` and `Profiled_attribute`, again related through a coreness set, and where the latter can be approximated as a 'part' (or possibly 'feature') of the former:

(8) John eclipsed Mary
John's talent eclipsed Mary's

Again, the consolidation of these arguments was judged to be linguistically valid, in part because VerbNet treats them as encoding the same thematic role (i.e. `Theme1`).

Other coreness sets which occurred repeatedly throughout our sample involved agent-cause alternations (e.g. *John/The blackout disabled the alarm system*) and speaker-medium alternation (e.g. *The critics/survey labelled her a has-been*). Again the intuitiveness of these coreness sets is supported by VerbNet thematic roles.

However, there were at least ten cases where the coreness sets lead to an invalid consolidation of arguments, in general caused by the fact that FrameNet syntactic information, and hence our lexical entry extraction procedure, does not distinguish between preposition phrases headed by different prepositions. For example, consider the following two example sentences involving the verb *jab* from the `Cause_impact` type:

(9) Mary jabbed John with a bayonet
Mary jabbed a bayonet at John

In both these sentences, *John* would be annotated as an `Impactee` and *a bayonet* as an `Impactor`. Since these two roles are part of the same coreness set, the subcategorisation frames underlying both sentences are consolidated into the following unified representation:

(10) `Sbj:NP:Agent`
`Obj:NP:Impactor/Impactee`
`Dep:PP:Impactor/Impactor`

This is clearly undesirable, since it leads to an unnecessary level of ambiguity for a parser, a conclusion reinforced by the fact that VerbNet treats the impactee and impactor arguments with distinct thematic roles (i.e. `Destination` and `Instrument` respectively).

It is worth dwelling a little on the possible reasons for FrameNet annotators formulating such an obviously unintuitive coreness set. In previous work (McConville and Dzikovska, 2007), we have noted the tendency to incorporate *all* uses of a particular verb into the same frame, even when syntax disagrees. For example, take the two uses of the verb *rip* in the following sentences:

(11) John ripped his trousers below the knee
John ripped the top off his packet of cigarettes

In both sentences, annotators have judged that the target verb *rip* evokes the `Damaging` frame, which has two important 'core' roles — `Agent` (i.e. the 'ripper') and `Patient` (the object that gets ripped). In this respect, annotation of the first sentence is simple — *John* is the `Agent`, *his trousers* is the `Patient`, and the prepositional phrase *below the knee* is assigned to a 'non-core', locative role called `Subregion`.

Assuming that the use of the target verb *rip* in the second sentence also involves the `Damaging` frame causes problems however — *the top*, is assigned to the non-core `Subregion` role, despite being realised as a (syntactically obligatory) direct object. Thus, in this case the syntactic generalisation that subjects and onjects realise core roles is overuled in favour of keeping all uses of the target verb within the same frame. A more appropriate analysis would have been to assign the use of the target verb in the second sentence to the `Removing` frame.

Considering again the examples involving the target verb *jab* in (9), we see that similar forces are at work. The hypothesised reason for grouping roles into coreness sets is where a number of distinct roles are realised by the same syntactic role — in this case, the direct object can realise either the `Impactee` (i.e. *John*) or the `Impactor` (i.e. *a bayonet*), so the formulation of a coreness set `Impactor/Impactee` makes sense. Note however that this is purely an artifact of the decision to treat both uses of the target verb *jab* as evoking the same frame. If the second sentence were treated as involving the `Cause_motion` frame, the undesirable coreness set would not have been formulated.

Therefore, we can conclude that, although the FrameNet coreness sets correspond in the vast majority of cases with valid underlying thematic roles, there are a number of problematic cases, at least some of which involve target verbs being assigned to suboptimal frames by annotators.

Note that information about the particular kind of preposition which can head a given PP argument is often considered to be a part of a subcategorisation frame, especially for deep parsers (c.f. the commonly used PFORM feature). If such information were available in FrameNet annotation, this would have the side effect of avoiding some of the problems caused by this kind of unintuitive coreness set, since the argument structures derived from the two sentences in (9) would not be identical — the first would have a `PPwith` dependent, whereas the second would have a `PPat`. However, it would also make it more difficult to merge arguments from some of the intuitive coreness sets such as that involving trajectory arguments, since these can be introduced by a large variety of prepositions.

In the future, we are planning to improve our lexicon extraction algorithm so that prepositions are taken into account in extracting and differentiating subcategorisation frames. This would require a more detailed investigation which arguments can be merged despite using different prepositions, and in which cases they should be kept separate. One possible solution is suggested by the approach taken in the VerbNet. The arguments in the VerbNet subcategorisation frame can either be associated with a single preposition (such as *with*), or with a class of prepositions (such as *P:loc* corresponding to a set of locative prepositions). This encodes the intuition that in some cases the preposition is fixed by the verb, and therefore 'meaningless', while in other cases the preposition is 'meaningful' in that it corresponds to a specific predicate (*on*, *in*, *under*) and can be drawn from a large set of possibilities. We therefore are considering using the FrameNet corpus data to see if a preposition associated with a given role appears to be fixed, or can be drawn from a larger set, and using this as a basis for making the distinction between meaningless and meaningful prepositions associated with coreness sets.

5 Discussion

In this paper, we argued that for purposes of parsing and semantic interpretation, a less specific set of semantic roles would ease lexicon construction and disambiguation. Consider an analogy with word sense distinctions: Palmer et al. (2004) argue that different levels of granularity are

needed for different applications. For example, information retrieval may require coarser distinctions, at the level of PropBank sense groupings, while machine translation may require much more fine-grained distinctions, such as those found in WordNet (Miller, 1995). Similar reasoning can be applied to semantic roles: coarser distinctions, such as the argument labelling assumed in PropBank (i.e. ARG0, ARG1, etc.), may be the easiest to disambiguate and annotate; thematic roles as used in VerbNet (i.e. AGENT, THEME, etc.) may provide an appropriate level of generalisation when linking syntactic and semantic structure; and the fine distinctions encoded in FrameNet (i.e. COOK, FOOD, etc.) may be useful for reasoning. Ideally, these different levels could be mapped to each other, similarly to the way WordNet senses are linked to VerbNet and PropBank entries. Our study is a first step in evaluating to what extent the different levels of generalisation could be linked in FrameNet through the use of features defined in its ontology, and in attempting to automatically derive a set of semantic roles and lexical entries at lower granularity.

While our research is primarily centered on the needs of a deep parser and lexicon, the algorithms we developed could also contribute to ongoing research on linking various lexical resources and annotated corpora, for both manual and automatic linking approaches. In case of manual linking, the SemLink project[5] aims to develop correspondences between the semantic types and roles underlying PropBank, VerbNet and FrameNet. In the future, we plan to compare results of our automatic procedure with the correspondences made by human coders. Assuming that there is sufficient agreement, this automatic approach could be adapted in the future to reduce the need for manual linking. For automatic linking, Kwon and Hovy (2006) propose an automatic algorithm for aligning role names between semantic lexicons, which achieves around 78% accuracy in aligning FrameNet and PropBank roles based on corpus evidence. It may be interesting to consider whether using either inheritance or coreness set information could improve the accuracy of the alignment algorithm.

Finally, statistical parsers and semantic role labellers (Gildea and Jurafsky, 2002) could benefit from having a smaller set of semantic roles, because this would reduce the data sparsity problem. Using the hierarchy to reduce the role set could be useful under the circumstances, without loss of data. It is admittedly less clear how the coreness set information could be used, but this too may be worth exploring if it could be utilised as a way of backing off to more general role names in a statistical model.

6 Conclusion

The aim of the project reported in this paper was to take a verb lexicon harvested fairly directly from the FrameNet semantically annotated corpus, and to apply some of the mechanisms within the FrameNet ontology to make this lexicon more effective for use with a deep parser. We argued that the lexicon would be improved with a more concise and generic role set, because it will simplify making links between syntax and semantics in the lexical entries. We examined: (a) the inheritance relation on semantic roles, and the corresponding links between semantic roles of increasing granularity, as a means of reducing the size of the vocabulary of roles across the lexicon as a whole; and (b) the coreness sets of related semantic roles specified within the FrameNet ontology, with the aim of consolidating subcategorisation frames within individual verb entries. In both cases, we concluded that the annotation scheme provide useful, though not perfect, mechanisms for our purposes. This is in part due to the fact that the relevant aspects of the scheme are not always applied in systematic manner across the FrameNet ontology. Making this part of the FrameNet annotation more consistent could benefit not only our application, but also applications that support linking between different resources, and potentially semantic role labelling applications.

[5] http://verbs.colorado.edu/semlink

Acknowledgements

The work reported here was supported by grants N000140510043 and N000140510048 from the Office of Naval Research.

7 References

James Allen, Myroslava Dzikovska, Mehdi Manshadi, and Mary Swift. 2007. Deep linguistic processing for spoken dialogue systems. In *Proceedings of the ACL'07 Workshop on Deep Linguistic processing*, pages 49–56.

C. F. Baker, C. Fillmore, and J. B. Lowe. 1998. The Berkeley FrameNet Project. In *Proceedings of COLING-ACL'98, Montreal*, pages 86–90.

Ann Copestake and Daniel Flickinger. 2000. An open-source grammar development environment and broad-coverage English grammar using HPSG. In *Proceedings of LREC'00, Athens, Greece*, pages 591–600.

Benoit Crabbé, Myroslava O. Dzikovska, William de Beaumont, and Mary D. Swift. 2006. Increasing coverage of a domain independent dialogue lexicon with VerbNet. In *Proceedings of the Third International Workshop on Scalable Natural Language Understanding (ScaNaLU 2006)*.

Myroslava O. Dzikovska, Mary D. Swift, and James F. Allen. 2004. Building a computational lexicon and ontology with FrameNet. In *Proceedings of the LREC'04 Workshop on Building Lexical Resources from Semantically Annotated Corpora*.

Myroslava O. Dzikovska, Mary D. Swift, and James F. Allen. 2007. Linking semantic and knowledge representations in a multi-domain dialogue system. *Journal of Logic and Computation, Special Issue on Natural Language and Knowledge Representation*.

Daniel Gildea and Daniel Jurafsky. 2002. Automatic labeling of semantic roles. *Computational Linguistics*, 28(3):245–288.

Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. Ontonotes: The 90% solution. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pages 57–60.

Ray Jackendoff. 1983. *Semantics and Cognition*. MIT Press.

Michael Kaisser and Bonnie Webber. 2007. Question answering based on semantic roles. In *Proceedings of the ACL'07 Workshop on Deep Linguistic processing*.

Karin Kipper, Hoa Trang Dang, and Martha Palmer. 2000. Class-based construction of a verb lexicon. In *Proceedings of AAAI'00*.

Namhee Kwon and Eduard Hovy. 2006. Integrating semantic frames from multiple sources. In *Proceedings of CICLing'06*.

Beth Levin. 1993. *English Verb Classes and Alternations*. The University of Chicago Press.

Mark McConville and Myroslava O. Dzikovska. 2007. Extracting a verb lexicon for deep parsing from FrameNet. In *Proceedings of the ACL'07 Workshop on Deep Linguistic processing*, pages 112–119.

Mark McConville and Myroslava O. Dzikovska. 2008. Evaluating complement-modifier distinctions in a semantically annotated corpus. In *Proceedings of LREC'08*.

Mark McConville. 2006. Inheritance and the CCG lexicon. In *Proceedings of EACL'06*, pages 1–8.

G. Miller. 1995. WordNet: A lexical database for English. *Communications of the ACM*, 38(5).

Martha Palmer, Olga Babko-Malaya, and Hoa Trang Dang. 2004. Different sense granularities for different applications. In *HLT-NAACL 2004 Workshop: 2nd Workshop on Scalable Natural Language Understanding*, pages 49–56, Boston, Massachusetts, USA, May 2 - May 7.

Martha Palmer, Paul Kingsbury, and Daniel Gildea. 2005. The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106.

Mihai Surdeanu, Sanda M. Harabagiu, John Williams, and Paul Aarseth. 2003. Using predicate-argument structures for information extraction. In *Proceedings of ACL-03*, pages 8–15.