OBLS HOBBLE COMPUTATIONS

Annie Zaenen and Dick Crouch
PARC MS-Powerset

Proceedings of the LFG09 Conference

Miriam Butt and Tracy Holloway King (Editors)

2009

CSLI Publications

http://csli-publications.stanford.edu/

Abstract

We discuss the problems created by the distinction between oblique arguments and adjuncts in general and in the XLE-based ParGram English grammar implementation. We argue that it is better to do away with the distinction for semantically marked obliques.

1 Introduction

One of the wonders of Natural Language is the way it manages to repurpose limited resources. Unfortunately for linguistic computations the way it does this is by relying heavily on non-linguistic context. This means that given the incomplete information that we have when we try to analyze text relying only on syntactic structure and limited lexical resources, our analyses are most of the time ambiguous in multiple ways. This problem can be compounded by the incompleteness of our understanding of syntax and of the structure of the lexicon. In this paper we discuss a case where ambiguities are created by incomplete syntactic and contextual knowledge and where attempts to remedy the explosion of ambiguities that they allow in fact make the problem worse. The problem we focus on is how the interaction between some obliques and adjuncts in the XLE-based ParGram English grammar (henceforth ParGramEnglish) and the interaction between its idea of subcategorization and VerbNet (http://verbs.colorado.edu/ mpalmer/projects/verbnet.html) information together create computational bottlenecks. We propose to solve part of these problems by assimilating one class of obliques to adjuncts.

2 Problem 1: The theoretical notion of oblique in LFG

Like most modern linguistic theories, LFG makes a distinction between syntactic arguments and adjuncts, or, in theory internal parlance, governable and non-governable grammatical functions (GFs). Within the governable GFs a further distinction is made between semantically *unrestricted* ones, subject (SUBJ), object (OBJ), and *restricted* ones, object theta (OBJ-TH) and obliques (OBL-TH) and within the OBL-THS we further distinguish between *idiosyncratically marked* ones and *semantically marked* ones. The topic of this paper is the status of semantically marked OBL-THS in computational embodiments of LFG. We argue that it would be better not to encode these as such but to assimilate them to ADJS.

Semantically marked obliques are marked by prepositions in English. The preposition is meaningful and indicates what the semantic role of the oblique is. A standard example is the *to*-phrase in sentences such as

(1) Mary gave the book to Bill.

[†] Thanks to Danny Bobrow, Lucas Champollion, Cleo Condoravdi, Martin Forst, Lauri Karttunen, Tracy Holloway King and John Maxwell for discussion .

As any textbook about prepositions for non-native speakers will tell you, *to* is an indicator of goals. The marking is not specific to the verb *give* and, depending on how general one thinks the notion of goal is, it can be argued it is not even specific to verbs of transfer of possession. This contrasts with the use of *on* in a sentence such as (2).

(2) John relied on Mary.

Here the use of the preposition is completely determined by the verb and there is no need to give it any independent meaning of its own. These obliques are classified as *idiosyncratic marked* (Bresnan, 1982a) or *quirky case-marked* (Butt and King, 2005).

2.1 Arguments and adjuncts

The main criterion that LFG uses to distinguish arguments from adjuncts is *uniqueness* as discussed in (Bresnan, 1982b). In a sentence arguments are unique, whereas adjuncts can be multiply specified. The example given in (Bresnan, 1982b) is

(3) *Fred* **deftly** [Manner] handed *a toy to the baby* **by reaching behind his back** [Manner] **over lunch** [Temp] **at noon** [Temp] **in a restaurant** [Loc] **last Sunday** [Temp] **in Back Bay** [Loc] **without interrupting the discussion** [Manner].

In this example we have italicized the arguments and given the adjuncts in bold.

This criterion at first seems reasonably straightforward but it requires careful syntactic analysis and an a priori agreement on what counts as the same or a different adjunct or argument. For instance, what is the difference in analysis between

(4) I count on you, on your kindness.

and

(5) He lives in France, in a small village.

Most people would take *you* in (4) to be an argument. This forces one to assume that *on your kindness* is a kind of parenthetical whereas *in France* and *in a small village* in (5) can be analyzed as separate locative adjuncts. There is no agreed upon list of either (oblique) arguments or adjuncts, and the same prepositions can introduce either. This means that in the absence of careful analysis it is often impossible to determine whether a prepositional phrase is an adjunct or an argument and careful analysis is lacking for most cases. For example, in

(6) He drove from Paris to Venice via Milan.

we could analyze *from Paris to Venice via Milan* as three adjuncts of the type directional or as three arguments of three different types or as a mixture with one or two arguments and one or two directional adjuncts. Maybe linguistic theory will eventually clarify these issues but at this point an implementation of LFG cannot make the required distinctions for all predicates, with the result that sentences like the one above will typically get several analyses.

As we will discuss in section 4, the distinctions that one wants to make in natural language processing are of a semantic nature. A uniqueness criterion is not particularly relevant in that respect. Other theories use more semantic criteria to make the distinction between adjunct and argument. For instance, (Dowty, 1989) proposes to use *semantic entailment*: a semantic argument is entailed by the meaning of its verb. For instance in *John walks*, John is an obligatory participant because there can be no walking without there being a walker. As formulated by Dowty, the criterion does not correspond to the pre-theoretic distinction between arguments and adjuncts because it would make arguments out of the elements that are most often classified as adjuncts, viz, locative and temporal elements. In

(7) John worked in the kitchen.

or

(8) Mary worked at noon.

in the kitchen and *at noon* are in general not seen as arguments although all *working* takes place somewhere and at some time. (Koenig et al., 2003) improve upon the criterion by stipulating that semantic entailment is a necessary but not a sufficient condition. They add to it a *specificity* condition: arguments are required only by a restricted set of verbs. This excludes immediately the locative and temporal elements mentioned above. (Koenig et al., 2003) are interested in psycholinguistic evidence for the argument adjunct distinction and argue that their criteria pick out classes that correlate with processing differences. This seem plausible but specificity in this sense is rather difficult to pin down as a crisp criterion for each verb. It seems then that in the current state of affairs no linguistic theory is developed enough to give criteria that allow us to straightforwardly distinguish arguments from adjuncts in many cases. So, even in the cases where we can hope one day to make the distinction based on syntactic and lexical criteria we are not able to do it now.

3 Problem 2: The implementation of obliques in the ParGramEnglish syntax

Subcategorized grammatical functions in LFG are unique. The theory assumes that different types of OBLs will be distinguished through different names. But the ParGramEnglish implementation chooses to allow only one oblique, OBL[1], which is treated like all other non-ADJ functions as being unique. This was done because the theoretical situation sketched above and the lack of contextual information would

[1] In fact, some further specialized obliques are used: OBL-AG, OBL-COMPAR, OBL-PART but they do not concern the type of obliques we are discussing here

allow for too many ambiguities: for all semantically marked obliques of verbs that also have simple transitive or intransitive subcategorization frames, we would also get an analysis that would treats these obliques as ADJs. For instance a verb like *drive*, has a subcategorization frame where it takes only a subject as in (9) as well as one where it takes a *from*-PP and a *to*-PP. Given this, a sentence such as (10) would at first blush get four analyses: (OBL,OBL), (OBL,ADJ), (ADJ,OBL), (ADJ,ADJ).

(9) John drove.

(10) John drove from the house to the school.

The restriction to one OBL eliminates one of these readings, the double OBL one.

Another analysis, the (ADJ,ADJ) one is eliminated by a feature of the ParGramEnglish implementation, called OT marks for Optimality Theory Marks (because it is in spirit related to Optimality Theory). The OT subsystem is described and motivated in (Frank et al., 1998). The XLE system allows the grammar writer to attach preference and dispreference marks to rules. These preferences and dispreferences can be further ordered in the configuration files, which are grammar specific. In the grammar under consideration OT marks are used to regulate OBLs, ADJs and PP attachment preferences. One OT mark says that, when the same c-structure span can cover either an OBL or an ADJ, the OBL is preferred. This excludes the (ADJ,ADJ) reading for the sentence above. But, in fact, the situation is more complex: without further information, the *to*-PP in the sentence above can also be attached as an NP adjunct (NADJ) to *house* and indeed that reading is not excluded on the syntactic level. These possibilities start to multiply when we consider a sentence such as:

(11) John drove the car from the house to the school.

Here the NP attachments can be be to *car* and to *house*.

If we look at the two PPs we see we can analyze both of them in three ways: as an OBL, as an ADJ and as a NADJ (nominal adjunct). A local OT mark tells us to prefer the OBL to the ADJ in each case, whereas there are no constraints on the NADJ combinations. The result is that we end up with the following possibilities: (OBL, NADJ), (NADJ,OBL), (NADJ,NADJ). This last possibility leads to two parses: one where the two NADJs are attached to the same noun and one where the second in embedded under the first one, but what is important for us is that the possibilities that we would want (OBL,OBL) or (ADJ,ADJ) or (OBL,ADJ) have all disappeared for various reasons. We could go back and reconsider how OT marks are assigned but assigning OT marks is a delicate balancing act and in this case it is not clear that we can improve the system as a whole.

This unfortunate result leads us to look more closely at whether there is an overriding reason to have the OBL-ADJ ambiguity for optional semantically marked obliques. As the discussion in section 2.1 shows, the status of these OBLs is very theory dependent and the LFG classification is very sui generis, which in itself leads us to think that we should not be too attached to it when it gets in our way.

This impression is reinforced when we consider how this class fares in further processing.

4 Problem 3: Constraining interpretations and combining lexical resources

A large coverage NLP system needs extensive lexical information. The systems that we are developing ((Bobrow et al., 2007)aim at a rather deep *normalization* of natural language text. We are developing a level, called AKR (Abstract Knowledge Representation) on which texts that mean the same thing are represented in the same way regardless of the variation in the surface string and texts that have different meanings are represented differently regardless of the similarity in the surface string.[2] For this deeper analysis, we definitely need information about how the meaning of one item in a sentence can constrain the meaning of other items. A subcase of this is the way the meaning of verbs constrain the meaning of their dependents. This information is typically encoded in the lexicon.

4.1 Is subcategorization information what computational approaches need?

An important subset of these constraints on verbal dependents are often talked about as semantic or thematic roles: the subject of a verb like *work* is the *agent* or the *worker* depending on the level of generalization/abstraction one wants to use for this type of information. This is useful information because it helps with paraphrases or entailments (for instance, the difference between transitive and intransitive *sink*) and, more generally, with the very necessary narrowing down of lexical choices: for instance, if we know that something has to be an *agent*, we know it has to have independent force.[3] If we are given this information we know, for example, that the meaning of *pilot* will not be the *pilot light* one in

(12) The pilot smiled at the passenger.

Computational lexicons, such as VerbNet, that encode these restrictions often give information about *alternations*: what is listed with a verb is the dependents that can be expressed in more than one way. This clearly doesn't represent the theoretical distinction between arguments and adjuncts. For instance, in the following alternation from (Levin, 1993), one can argue that, in the first variant, we have three arguments but in the second nobody will analyze *(the) horse's* as an argument of *touch*.

[2] It is clear that this cannot be achieved absolutely, or rather that there is no advantage in achieving it absolutely, as at the limit, in every pair of non literally matching texts, the two texts mean something different. We mainly try to achieve sameness of propositional meaning.

[3] At least, we would know this if these semantic/thematic categories to which grammatical functions map to were well-defined. This is far from being the case but we will ignore that problem here.

(13) Selina touched the horse on the back.

(14) Selina touched the horse's back.

The alternation information is the information that computational lexicons need and try to cope with. Because it often looks like the information that pre-theoretically can be thought of as subcategorization information, one has the tendency to assimilate it completely to this. But as the example above shows, this is not warranted. In fact the conflation of lexical constraints with subcategorization information has led computational linguists to neglect important lexical information that can constrain the interpretation of adjuncts.

(15) He left for three days. ⇒ The period of three days is after the leaving.

(16) He worked for three days. ⇒ The period of three days is the period of the working.

Here the choice of the verb determines the interpretation of the temporal phrase. This is not seen as a subcategorization restriction and hence there is much less knowledge about the verb classes involved in this and similar phenomena than there is about phenomena that are considered to be part of subcategorization.

4.2 Lexical resources

For lexical information, all systems are dependent on resources that have been created outside of the system because no one enterprise can do it all. Specifically our implementation based on the ParGramEnglish syntax also relies on VerbNet. VerbNet is based on (Levin, 1993) verb classes. It intends to describe the alternations for a large subclass of verbs. The alternation information about the verbal dependents is expressed in syntactic categories such as NP and PP that can be found in the immediate environment of a verb. These syntactic categories are mapped onto thematic roles such as *agent*, *patient*, and the like. These in turn are associated with a semantic frame that spells out the event structure of a verb argument combination in terms of semantic predicates such as *cause*, *manner*, *directed motion*, etc.

Our implementation combines the information from its own lexicon with information from VerbNet to create the Unified Lexicon (UL). [4] The VerbNet information is combined with the ParGramEnglish subcategorization information because that is the only information that the system has about the dependents of its verbs. If one interprets the information contained in VerbNet also as subcategorized information one has to figure out what type of notion of subcategorization VerbNet uses to ascertain whether this mapping is warranted. VerbNet does not tell us this. As far as one can derive from looking at what is in VerbNet, the notion used is a mix of the entailment and the alternation approach.[5] There is no reference whatsoever to a uniqueness criterion.

[4] See (Crouch and King, 2005) for details.

[5] It cannot be the entailment approach alone because for certain classes when (Levin 1993) clearly states that there is no entailment relation, VerbNet proposes a frame, e.g. class 11.5 (*drive* verbs).

As we discussed in section 4.1, seeing the information in VerbNet as subcategorization information might be the wrong idea but whether one assumes that the information in VerbNet is subcategorization information or not, it is clear that ParGramEnglish and VerbNet have a different view on which dependents should be listed in the lexicon with each verb. Given this, the discussion in section 2 and the OBL restriction noted in section 3, it comes as no surprise that the syntactic frames of VerbNet do not correspond well to either LFG or ParGramEnglish subcategorization frames. At the very least, the system has to provide a way to handle the multiple PP complements that VerbNet allows. This is currently done by allowing ParGramEnglish ADJs to function as VerbNet arguments. But, apart from the restrictions to one OBL, LFG and ParGramEnglish associate c-structure components with grammatical functions and VerbNet associates c-structure components with thematic roles. There is no one-to-one mapping between these two. This has as a consequence that the mapping rules often give incorrect results. Most of the errors are in the mapping of OBLs because that is where the two lexicons differ the most.[6]

Our UL then contains unreliable information about PP arguments. Moreover, it is unwieldy: it follows VerbNet in treating each possible subcategorization frame for each predicate as a separate lexical entry. Thus, all the possible combinations of PPs that can be associated with each argument taking predicate need to be spelled out. The spelling out can be done by rule but the result is still a list of lexical items. Given that there is no agreement on what belongs to a subcategorization frame, there is no end to the number of PPs that can be proposed as parts of a subcategorization frame. More importantly, given that, for normalization, our interest is in fact in alternations or in meaning restrictions, the notion of subcategorization frame is not the most relevant and possibly more combinations are relevant than anybody would put in a subcategorization frame.

5 Towards a solution

5.1 Proposal for the elimination of semantically marked obliques: A (partial) solution to the ambiguity problem

We have seen that treating *semantically marked* OBLs as complements leads to a proliferation of ambiguity in parsing. This is especially so in the case of multiple obliques, but is also true for single obliques, e.g.

(17) John sent flowers to Mary.

(18) John sent flowers.

Given the existence of subcategorization frames for *send* both with and without the *to* oblique, when a *to*-PP is present, the grammar has the option of treating it either

[6] The mapping from post-verbal NPs to OBJs is in most cases rather straightforward. There are also problems with the specific thematic role assignments that VN proposes, but that is not the topic of this paper.

as an oblique or as an adjunct. Deft placement of OT marks is required to eliminate the ambiguity.

We have also seen that the system has a restriction to single obliques, in an effort to reduce syntactic ambiguity. This makes the integration of VerbNet information more complex when in the VerbNet frame there is more than one PP argument. For a parse assigning a single syntactic subcategorization frame but with multiple PP adjuncts, one needs to look through a variety of VerbNet frames to see if any of the PP adjuncts can be treated as a semantically marked oblique.

From a parsing perspective, the adjunct/complement distinction for semantically restricted obliques is, as we have seen, hard to draw determinately; and for multiple OBLs, all but one are in any case treated as adjuncts. So why not treat *all* semantically marked obliques as PP adjuncts?

The effects of this on parsing would be threefold. First, there would be a reduction in the degree of ADJ/OBL ambiguity, with concomitant gains in parsing speed. Second, OT marks controlling this ambiguity could be simplified. Third, the number of verb/subcategorization-frame pairs in the lexicon would be reduced, simplifying the task of lexical maintenance.

From a semantic point of view, there is no loss of information if syntax treats semantically marked OBLs as ADJs. This is because mechanisms already exist for finding *some* of the VerbNet specified semantic obliques amongst the adjunct set; one merely has to apply this mechanism to find *all* such obliques amongst the adjunct set. But the more interesting question is whether lexical semantic processing could, like syntactic processing, be made simpler.

The processing complexity at issue arises from the fact that an (OBLless) syntactic subcategorization frame like V-SUBJ-OBJ for a verb such as *send* may have to be compared to a variety of semantic frames like V-SUBJ-OBJ-OBL(TO), V-SUBJ-OBJ-OBL(FROM), or V-SUBJ-OBJ-OBL-OBL(FROM,TO). Comparison would be facilitated if instead there was just one semantic frame, V-SUBJ-OBJ, with an additional specification that *from*- and *to*-PPs were candidates for mapping to semantically marked roles like source and destination. Another way of looking at this conceptually is that *from* and *to* do not introduce obliques at any level, merely PP adjuncts whose range of interpretation is constrained by the verb to which they apply.

Computationally, therefore, eliminating semantically marked obliques appears to lead to gains all round.

5.2 Efficient encoding of adjunct role restrictions: a (partial) solution to the mapping problem

By eliminating semantically marked OBLs, there is a reduction in the number of syntactic and semantic verb frames that need to be encoded. But a naive implementation would still record, for each frame with obliques eliminated, the role restrictions on the prepositional adjuncts. This is still more verbose than it needs to be. Most of the role restrictions are not dependent on the particular verb, at least not for the level of granularity at which VerbNet assigns roles.[7]

Indeed, most of the role assignments in VerbNet are based around alternation classes of verbs. For example, in verbs undergoing the instrumental alternation (*John broke the window with a hammer* vs. *A hammer broke the window*) *with*-PPs have a restricted Instrument interpretation open to them. Rather than recording this separately for each verb frame tagged as undergoing the instrumental alternation, the information can be encoded just once as a property of all instrumental verbs.[8]. This can be done in a similar way for the other verb classes that VerbNet encodes. This approach also allows us to specify the mapping of prepositional ADJs to thematic roles per verb class and in that way alleviate the problem of wrong mappings mentioned in section 4.2

6 Conclusion and outstanding problems

One of the main reasons for using lexical semantic information from VerbNet is to capture the kinds of paraphrase that are opaque if one looks at grammatical roles alone, e.g.

(19) John$_{\mathrm{SUBJ}}$ broke the window$_{\mathrm{OBJ}}$. ⇒ The window$_{\mathrm{SUBJ}}$ broke.

(where *window* maps onto the *theme* role in both cases). The elimination of semantically marked obliques does not interfere with the representation and recognition of alternation paraphrases. Both complements and semantically restricted adjuncts continue to receive their semantic role assignments, albeit through slightly different means of specification.

The proposal sketched above alleviates the ambiguity problem and the mapping problem that *semantically marked* obliques pose for the system. It does not eliminate all unwanted ambiguities: many are due to the lack of semantic information about the verbal dependents themselves. The proposal also does not deal with *idiosyncratically marked obliques*. Idiosyncratically marked obliques are ones that either have to be syntactically present (e.g. *rely on*), or are syntactically optional but their presence substantially alters the meaning of the verb (e.g. *answer* vs. *answer for*). Both kinds of idiosyncratically marked oblique need to be explicitly marked in verb frames.

The syntactically obligatory obliques are easy to identify, and clearly have to be encoded in syntactic subcategorization frames (or else their obligatory nature will not be reflected in parsing). It is less clear that optional obliques need to be recorded in syntactic subcategorization frames (since they are optional, they don't constrain parsing), and their identification relies on judgments of sense differences between verbs with and without the oblique. One can use WordNet to produce an

[7] VerbNet's *agent*, *theme*, *patient*, etc. roles are not inherently verb-specific in the way that finer-grained roles like *worker*, *employer*, *employee*, etc. would be.

[8] However, a lower level mechanism for recording role restrictions on individual verb-frame pairs may be needed to either override generalizations in specific cases, or to include additional role restrictions specific to the verb.

initial list of verb-preposition pairs that have different senses than verbs alone, and hence where the prepositions are candidates for idiosyncratic obliques; but the list is liable to be both incomplete and error-prone. However, since the oblique controls the sense of the verb, and not just to the role assignment to the prepositional argument, the oblique does at least need to be explicitly recorded in the semantic verb frame. For the moment we opt for leaving them as subcategorized.

References

Bobrow, Daniel G., Bob Cheslow, Cleo Condoravdi, Lauri Karttunen, Tracy Holloway King, Rowan Nairn, Valeria de Paiva, Charlotte Price, and Annie Zaenen. 2007. PARC's Bridge and Question Answering System. In T. King and E. Bender, eds., *The Proceedings of the GEAF07 Workshop*. CSLI.

Bresnan, Joan. 1982a. Control and Complementation. In J. Bresnan, ed., *The Mental Representation of Grammatical Relations*, pages 282–390. MIT Press.

Bresnan, Joan. 1982b. Polyadicity. In J. Bresnan, ed., *The Mental Representation of Grammatical Relations*, page 164. MIT Press.

Butt, Miriam and Tracy Holloway King. 2005. The Status of Case. In V. Dayal and A. Mahajan, eds., *Clause Structure in South Asian Languages*, pages 153–198. Kluwer.

Crouch, Dick and Tracy King. 2005. Unifying lexical resources. In *Proceedings of the Interdisciplinary Workshop on the Identification and Representation of Verb Features and Verb Classes*. http://www-csli.stanford.edu/ thking/verb05crouchking.pdf.gz.

Dowty, David. 1989. Grammatical Relations and Montague Grammar. In P. Jacobson and G. Pullum, eds., *The Nature of Syntactic Representation*, pages 69–129.

Frank, Anette, Tracy Holloway King, Jonas Kuhn, and John Maxwell. 1998. Optimality Theory Style Constraint Ranking in Large-scale LFG Grammars. *Proceedings of the LFG98 Conference* .

Koenig, Jean-Pierre, Gail Mauner, and Breton Bienvenue. 2003. Arguments for Adjuncts. *Cognition* 89:67–103.

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