ESK: An Event and State Knowledge Base

Greg Whittemore
John Hicks
Language Technology Group
Human Communications Research Centre
University of Edinburgh
Edinburgh, Scotland EH8 9LW
greg@cogsci.ed.ac.uk
jwjhix@cogsci.ed.ac.uk

Abstract

Within is a description of both the contents and employment of a knowledge base used for storing event and state lexical semantic information. Of particular focus is an explanation of how the knowledge base is currently used to define thematic roles and to enrich and disambiguate logical forms.

1 Introduction

Within is a synopsis of ESK (event and state knowledge base). ESK is being created simultaneously with the development of the MENELAS English natural language understander (henceforth MENLU), a system designed as part of the multi-lingual MENELAS natural language understanding / information retrieval project in the medical domain[1]. In particular, ESK is a lexical knowledge base constructed to enhance the descriptive capacity of MENLU, facilitating the interpretation of events and states (henceforth simply events) as found in free text.

The input for MENLU comes in the form of patient discharge summaries (PDSs) dealing predominantly with cardio-vascular case histories. The information targets established for 'understanding' via MENLU include past patient histories, lifestyles, diagnostic tests, hypothesized diagnoses, and current and future treatments. For the purposes of MENELAS, terms referring to these targets and sufficient contextual information must be identified and represented in such a way that subsequent queries can exact appropriate responses from the system. Interpretations in MENLU are limited. We are not seeking to put in every last detail known about an event into actual output representations. We simply want to make an 'adequate' or 'sufficient' interpretation that can subsequently be used synergistically alongside a knowledge base and inference component in order to formulate appropriate responses under 'retrieval mode'.

Limiting the detail in our interpretations is mostly an engineering response. Computational linguists and knowledge base developers alike have long been plagued with the difficulty of establishing the 'proper' or 'sufficient' level of representational detail in output forms and knowledge bases. At times, it even seems that there can be no 'right' answer. A lot can depend on the needs of a particular application or even on the particular perspective held by a specific user. Nirenburg and Defrise (1992) nicely discuss such difficulties confronted by knowledge base engineers in their description of DIONYSUS; their response being that one must 'provide concrete answers to such questions' (p. 294). We agree in practice with their concrete suggestion, but in theory we would think that we are building a flexible platform which can be customized for a given application (and we do not mean to imply that their platform is not a flexible one).

The lack of a 'right' answer regarding the level of detail required in a knowledge base was especially evident when we were deciding on how to best represent the relationship between events and event participants. Assignment of singular thematic roles (Fillmore (1968) and Grimshaw (1990)), has long been a traditional method for some, but other camps have never given them serious consideration (such as with strict model theoretic Montagovians). Many current notions also permit multiple semantic role assignments (Jackendoff (1990), Wilkins and Culicover (1986), and Bresnan and Kanerva (1988)). There is also a strong movement for a version based on 'cluster' definitions for which holding the largest number of basic proto-thematic-role properties determines thematic role type, a notion which was first started by Dowty (1988) and more currently discussed in Sanfillipo (1990) and Manandhar (1993) (Sanfillipo and Mandanhar also suggest possible implementations of feature set thematic role accounts).

Our particular solution for denoting the link between event/states and arguments depends on a reader's perspective. We have attempted, in a modest way, a flexible solution that has at its core a knowledge base rich in details regarding events and participants, but in our interpretation form such detail is kept at a minimum. If more detail is required for a particular application or subsequent linguistic process, the knowledge base is available as a resource.

A primary task of ESK is to provide sufficient linguistic wherewithal to identify, disambiguate, and validate predicate/argument relationships. The nature and organization of this linguistic support and how it is used ser-

[1]MENELAS is a multi-national project funded by the European Communities initiative for Advanced Informatics in Medicine, Project # A2023.

ves as the primary focus for the paper. Also of concern, as mentioned above, is the matter of building a bifurcated response system that combines information from a resulting semantic interpretation and support knowledge from a knowledge base. Some discussion on this is also included.

An outline of the remaining sections is as follows:

- a description of the kind of text to which ESK is currently applied
- an overview of MENLU
- ESK's role in processing
- a description of ESK's architecture
- a description of ESK's content

2 The Targets

The overall task of the MENELAS project is to develop a message understanding/information retrieval system that can operate over PDSs. For prototyping purposes, texts were initially limited to those regarding patients who have suffered myocardial infarctions. Through the advice of medical professionals, approximately 200 information targets were identified as 'relevant' for medical research and administrative purposes in these texts. These targets served as guide for establishing understanding goals for the three language systems used in MENELAS (Dutch, English, and French). For the English system, MENLU, the text contained in 475 unedited English PDSs (approximately 98,000 words of text) has been used to establish an initial grounding for the MENLU lexicon, syntactic and semantic components. Of relevance to this paper are the approximately 400 headword verbs the corpus contains. Of note too is the form of PDSs the different language groups worked with. The Dutch and French PDSs primarily are couched in a 'telegraphic' style almost entirely based on noun phrase structure. The English PDSs on the other hand were primarily written as sets of fully structured sentences.

Each language understander in MENELAS has its own processing components that are to generate fairly rich semantic representations that are consumable by a language independent pragmatic analyzer/database generator. The design and development for the individual language processing components were left for each language group to decide, with the requirement that output is adequately rich in semantic detail and compatible with the shared pragmatic analyzer. Also shared by the three languages is a heart-domain type hierarchy, initially embodying approximately 950 types. Most of the entries contained in the hierarchy pertained directly to the 'myocardial infarction' domain (e.g. entries for anatomical structures and diagnostic procedures), but a large number of items were of a more general linguistic nature (e.g. 'measure' types such as 'gram' and time units). An example type-hierarchy segment looks as follows:

```
therapeutic_substance
                > anticoagulant.
      anticoagulant
                > injectable_anticoagulant.
      injectable_anticoagulant
                > heparin.
      anticoagulant
                > oral_anticoagulant.
      anticoagulant
                > aspirin.
          aspirin > aspegic.
therapeutic_substance
                > anti_inflammation_drug.
      anti_inflammation_drug
                > corticosteroid.
```

where ' x > y ' means x is a super-type of y

3 MENLU — The English Natural Language Understander

When the project began, a design choice had to be made as to which morpho-syntactic processor should be used for the English understander. One available option was the Alvey broad-coverage parser/grammar (Briscoe, et al. 1993). The Alvey system has a fairly traditional syntactic/semantic architecture which incorporates an extensive GPSG grammar rule set with each grammar rule accompanied by one or more semantic mappings. Semantic representation in the publically available Alvey is couched in unscoped first-order predicate logic and most akin to what is often known as initial logical form (henceforth ILF). It has an 'event-based' semantics of a neo-Davidsonian (1967) or Parsonian (1990) flavor. Alvey event representations employ an 'event variable', a representation of tense, and an ordered set of event arguments aligning with 'deep-structure' notions for subject and direct/indirect objects. Alvey also associates arguments with predicates in traditional control, passive, and 'causative verb phrase' (such as in 'make her change her therapy') contexts. It does not, however, make use of thematic roles and lacks semantic functional assignments for prepositional phrases. It also does not attempt to map predicates or arguments contained in an ILF to individual semantic senses nor does it have a means for validating concepts such as could partially be achieved through verification of presupposed type constraints.

In contrast were the available smaller-scale graph-unification, lexically-based grammars of HPSG, UCG, and CCG.[2] With such systems it is fairly easy to incorporate rich semantic representations and incorporate substantial type-checking/ validation procedures into semantic processing. Though these systems are theoretically pleasing we did not have a version of any one of

[2]HPSG — Head-Driven Phrase Structure Grammar (Pollard and Sag 1994),
UCG — Unification Categorial Grammar (various treatments combining graph-unification with Montague-style categorial grammar (Sanfillipo's 1990 dissertation employs a UCG), and CCG — Combinatory Categorial Grammar, first introduced by Ades and Steedman (1981).
The PLEUK system (Calder 1992) is a publically available shell that has prototype versions of these three grammar formalisms.

them that had been proven robust enough and that had substantial grammar coverage. We chose the Alvey system largely because of its coverage, but also because it is computationally sound and returns results in reasonable parse times. By adding to the Alvey core-processing unit a preprocessing and preparser component and additional semantic interpretation facilities, we have ended up with a fairly substantial natural language understanding system.

4 Enriching Event Semantics in MENLU — The Representation

The expected input to the pragmatic analyzer shared on the MENELAS project requires much of the information contained in a basic Alvey ILF along with semantic functional assignments and 'type' information that it does not routinely contain (as described immediately above). In order to build sufficiently rich representations an 'embellisher' component was created that takes initial logical forms, verifies concepts they contain, identifies and incorporates appropriate semantic senses and relations, and scores and ranks resulting FLFs (final logical forms). An example transformation between an ILF and FLF is given immediately below. The sentence represented in these forms is:

Dr. John admitted Mrs. Smith for angioplasty.

```
(decl
  (and
    (admit.vp/np
     (uqe (some (e1)
       (attr e1 past//syn_tense)))
      (name ( the (x1)
        (and (attr x1  sg//syn_number)
                (named x1 Dr. John))))
      (name ( the (x2)
        (and (attr x2 sg//syn_number)
                (named x2 Mrs. Smith))))
  (for.p e1
   (uq
   (angioplasty//vascular_surgery x3))))))
```

ILF

```
                -.-.-.-.-.-.-.-.-.-.-.-.
(DECL
     (uqe  (some (e1)
                (and (admit/vp_np//ag_th/vsoci e1
      (attr e1 past//syn_tense)
      (AGENT e1
        (name ( the (x1)
            (and (--> x1 human_being)
                      (attr x1 sg//syn_number)
                      (named x1 Dr.John )))))
      (THEME e1
        (name ( the (x2)
            (and (--> x2 human_being)
                      (attr x2 sg//syn_number)
                      (named x2 Mrs. Smith)))))
  (PURPOSE e1
     (uq (some (x3)
        (and
         (angioplasty//-
                        vascular_surgery x3)
          (--> x3 treatment)))))
  (LOC.GOAL e1 (uq (some (i2)
       (--> i2 medical_center)))))))
```

FLF

In the example, embellishment to the final form is manifested in five primary ways:

- the verb *admit* has been linked with a specific semantic sense, namely the sense of admit that has as a concept type 'vsoci' (meaning a verb of social interaction) and underlying semantic arguments of 'ag(ent)', 'th(eme)',[3] 'purpose', and 'locative' and a surface form corresponding to that of a transitive verb.[4]
- the individual arguments to the verb admit are placed within two-place thematic-role functions with the first argument being a variable bound to the *admit* event and the second to a noun phrase object.
- an implicit 'argument' of *(loc.goal e1 i2)* is inserted into the 'admit' event and functions to represent a 'salient', unfilled role.
- presupposed types or constraints have been posted for each argument by way of the ' --> ' predicate.
- the ambiguous *for.p* preposition has been assigned a semantic function of 'purpose'.

4.1 Enriching via Thematic Roles

The characterization of the thematic roles in the final form is related to that proposed by Carlson (1984, p. 268) who states "*....thematic roles can be looked upon as functions which map individuals to sets of events — the events in which that individual participates in that particular way.*" In the example above, the meaning of the function *(theme e1 (Mrs.Smith x2))* is the set of *Admit-to-the-hospital-for-angioplasty'* events in which *Mrs.Smith* functions as a *theme*. Carlson adds that each thematic role 'assignment' is dependent on verb meaning and verbs that "*assign different thematic roles should be considered as meaning different things.*" Thus the verb *open* in *The door opened* and *The door was opened* assigns a *theme* in both instances and an additional *agent* in the latter but not the former. The lack of an *agent*, in *The door opened* helps to explain how the former *open* is different than the latter one.

[3] Unfortunately, the thematic role 'theme' means different things to different researchers. We will have an underlying prototypical meaning of 'something that moves or is located.'

[4] This particular sense is labeled in ESK with the index admit/vp_np//ag_th/vsoci. We could have chosen a numerical-based indexing scheme — this version made the labelling more perspicuous.

What precisely a *Theme'* or likewise an *Agent'* means (and similarly with other thematic roles) has been at the heart of much controversy in recent linguistic theory. Dowty (1991) and Ladusaw and Dowty (1988) have brought to the forefront the problems of thematic role labels. They point out the futility in attempting to derive a set of labels that can accurately cover all cases.

In an attempt to infuse more precision into identifying a subset of roles — agents and patients — Dowty has proposed that thematic roles can be 'entailed' by way of counting the number of 'agent-like' or 'patient-like' properties an argument has. An argument that has the most 'agent-like' properties (including *volition*, *sentience*, and *causes an event*, and *movement*) for a given event is the 'agent' and the one with the most 'patient-like' properties (including *change of state*, *incremental theme*, and *causally affected by event*) is the patient. The properties and their respective values do not 'define' what agents or themes mean, but they do help to characterize them and, perhaps more importantly, characterize them with respect to the verbs with which the roles are associated. Different verbs will have differing numbers of positive values for their agent properties (e.g. some will have (volition +) and (cause +) others will not) and their patient properties when they have them.

Jackendoff (1990) offers another perspective on thematic roles. One of the frequent complaints about thematic roles is that it is often hard to choose the 'right' one — often two (or more?) roles seem applicable. Instead of trying to derive a single, all encompassing set of roles that could cover the full range of roles that arguments play, he has developed a multi-level thematic tier system in which single arguments can be assigned multiple thematic role functions.[5] Employing a decompositional representation, each sentential argument can serve as an argument in one or more predicates which in turn can be assigned to different thematic tiers. For instance, the subject in an example sentence of Jackendoff's (p. 127) *the car hit a tree* is the *actor* argument that '**affects**' in the **action tier** and the *theme* that moves on the **theme tier**. His representation for this event is:

$$[\text{EVENT}] \longrightarrow \begin{bmatrix} INCHOATIVE[BE([CAR],[AT[TREE]])] \\ AFF([CAR],[TREE]) \end{bmatrix}$$

Similarly a *baseball* in a sentence like *Hank hit the baseball* could be both a *patient* and a *theme* in that it is both affected, a characteristic of *patients*, and it moves, a characteristic of *themes*.

Further flexibility in his system is exemplified by the way variations on predicates can be made through featural modifications. For instance, to distinguish the difference in various kinds of force-dynamic verbs[6], such as in

Harry forced Sam to go away

Harry pressured Sam into going away

Jackendoff makes featural variations on his CAUSE predicate in his **theme tier** (p. 131). For forced type verbs he assigns a '+' to designate successful cause — the first argument (the antagonist) asserts an influence and the 'effect takes place', which he represents as

$$\begin{bmatrix} CAUSE^{+}([HARRY], \begin{bmatrix} GO([SAM],[AWAY]) \\ AFF([SAM],) \end{bmatrix}) \\ AFF([HARRY],[SAM]) \end{bmatrix}$$

For verbs like *pressure* he uses the feature '^{u}' to assert that the end result is *unknown*.

$$\begin{bmatrix} CAUSE^{u}([HARRY], \begin{bmatrix} GO([SAM],[AWAY]) \\ AFF([SAM],) \end{bmatrix}) \\ AFF([HARRY],[SAM]) \end{bmatrix}$$

Returning to the question of what a *THEME* and *AGENT* (and similarly other thematic roles) mean, it would appear that Dowty is right in his assessment of assignments such as *(AGENT e1 x1)*, in so far as there is a lot of variation in what a function like *Agent'* could denote. In order to get a more complete characterization of what *Agent'* means, a more intimate examination of how an argument interacts with particular verbs or verb types is required. Dowty's method of partly characterizing thematic roles is via a simple feature/attribute scheme in which he depicts whether or not an agent has positive or negative values for properties such as 'volition' and 'causer' for individual events. His (1991) scheme, however, was not designed as a means for 'defining' thematic roles, but instead as a means to help identify thematic roles (in particular agents and themes).[7]

His 'clusters' do not have much of a 'cognitive' or conceptual orientation to them (nor were they intended to) — especially when compared with Jackendoff's representations. Jackendoff's representations go into fine detail about how a semantic argument relates to other event participants; he makes use of various predicates (in which semantic arguments are contained) and these predicates can in turn participate in multiple thematic tiers and be modified through additional featural modifications.

If 'thematic' roles mean, in Jackendoff terms, those notions couched in argument slots within multiple predicates available in multiple thematic tiers, then the thematic role names included in our FLFs would seem to have little meaning. Not only would they seem to fail to reflect subtle differences in how a thematic role can vary from verb to verb, but using a single role name when in fact multiple thematic role assignments might be appropriate would seem wrong.

To remedy the seemingly apparent shortcoming of the labels used in FLFs we could choose to dramatically alter

[5] Other researchers who have permitted arguments to have multiple thematic role assignments are Bresnan and Kanerva (1988) and Wilkins and Culicover (1986).

[6] Jackendoff attributes the notion of force dynamic verbs to Talmy (1985).

[7] In fact values for individual property-clusters are not 'assigned' a priori by verbs. A positive value for 'sentient', for instance, is not 'assigned' by the verb, but instead comes about through identification of the 'sentientness' of an argument. When a possible proto-agent is found to be sentient, the 'sentient' quality adds to its agenthood qualities.

our current representations and reconfigure them as conceptual structures of a Jackendoff variety. Doing away with the current overall FLF structures and replacing them with Jackendoff conceptual structures would certainly provide detailed accounts of individual thematic roles, but it would also mean a dramatic change in our current stuctures, a change we have reservations about because it is unclear how beneficial such complex structures would be for every sentence in our message understanding application. We are not necessarily prepared to build Jackendoff structures right now, but they do contain what seems to be useful information. As an alternative, we would like to capture a lot of the information he contains in his structures, and then use such information when we know it is necessary.

Our attempt at doing exactly this borrows from Dowty's approach for partially characterizing proto-roles. We, where more detail is required, use multiple featural assignments in order to characterize individual verb arguments. Many notions evident in Jackendoff's conceptual structures (and we couch other features as well) can be recouched as feature/value pairs: (CAUSE U) for 'unknown success for a causer', (CAUSE +) for 'a successful causer', etc. (as in the 'force-dynamic' examples given earlier), and where appropropriate we can use multiple typical thematic role labels such as saying an argument is both (THEME +) and (PATIENT +). These feature/values in our current system are provided on a per verb basis — each verb can potentially make a number of such specifications to each of its arguments.

Adding aspects of Jackendoff concepts via individual features not only allows us to make more 'intimate' correspondences between arguments and events, it also gives us the potential for creating Jackendoff like structures on top of our current representations (through an added inferencing component) if the need arises. This *on-demand* approach is computationally more attractive and could be a good platform for designing natural language systems for different applications (e.g. a spatially intensive graphics+natural language package might benefit more from Jackendoff 'thematic tier' representations than say a general message understanding application.)

One final detail that needs to be accounted for is "where are the features?" If one looks back at the FLF above, one will note that it contains only thematic role labels such as *agent* and *theme*. There are no features of the type just described. The answer is the features are kept within a separate knowledge base, ESK. The thematic role labels that play a part in describing the semantics of individual events have explicit extensions within ESK. The extensions are couched as feature clusters (sometimes just single features are needed) and are assigned on a per verb basis. Thus an event like the *operate* event in *Dr. John operated on the patient* incorporates , in FLF representations, only predicates headed with role-names like *Agent* and *Patient* in order to specify thematic roles. The role names in turn have grounding in ESK in the form of featural descriptions for the individual roles of the verb operate. In ESK the *Agent* for *operate* has as part of its value [[volitional +][cause +]] and likewise its *Patient* is a slot and has as part of its value [[affected +][altered +]]. An account of how they are represented and what is represented is given in the 'ESK Contents' section later in the paper.

4.2 Enriching via Implicit Roles

The inclusion of implicit roles, such as the *(loc.goal e1 (uq (some (i2) (-> i2 medical_center))))*, in FLFs but not in ILFs also originates from Carlson (1984). He defends a separation of syntactic subcategorization requirements and underlying semantic specification. Carlson notes the problematic accounts of others concerning 'detransitivizing' between pairs such as *John ate a sandwich* and *John ate* which *"has the effect of deleting the direct object from the verb's subcategorization frame"* and employing the additional operation *of existentially quantifying over the object position in the intransitive version* (p. 263). The approach taken by the MENLU logical form 'embellisher' in order to handle similar 'detransitivized' constructs is to meld an ILF with a verb semantics. In Alvey, an ILF for an intransitive verb is simply the result of employing a syntactic intransitive verb rule in order to obtain a syntactic representation of *[sent [np JOHN][vp ATE]]* and an accompanying ILF of *(DECL (Eat (uqe (some (e1) (attr e1 past) (John x1)))))*. The verb semantics for *eat* specifies the thematic role relations of *agent* and *theme* — *((EAT e) ∩ (Agent e (some x)) ∩ (Theme (some y)))*. The embellisher, as illustrated below, applies a mapping function that aligns the arguments available in the ILF with that of the lexical semantic representation for *eat*, resulting in the interpretation $((EAT\,e) \cap (Agent\,e\,John') \cap (Theme(some\,y)))$.[8]

(arg-map
$(DECL\,(Eat\,(uqe\,(some\,(e1)\,(attr\,e1\,past)(John\,x1)))))$
$(((EAT\,e1) \cap (Agent\,e1\,(some\,x)) \cap (Theme\,(some\,y)))$
$\Longrightarrow$
$((EATe1) \cap (Agente1John) \cap (Theme(somey)))$

Possibly somewhat controversial for the MENLU approach is that it includes not only implicit traditional thematic role arguments, but also arguments for relations sometimes thought to be derived from adjuncts (cf. the *loc.goal* role in the illustrated FLF above) and even temporal roles depending on the telicity of a verb. Many adjunct-like notions are included in MENLU FLFs for several reasons. By far the most important one is that they seem to be as much a part of event semantics as subjects or directly stated direct objects. Including implicit roles for them in an event interpretation is no less relevant than noting the missing object in the sentence *John ate*. The same mapping function described above that aligns syntactic arguments with thematic roles, while also making explicit implicit roles, is used for the adjunct situa-

[8] Additional support for having an underlying semantics that includes thematic roles can be found in the psycholinguistic study of Carlson and Tanenhaus (1988). Sanfillipo (1990) provides a very good review of Carlson's work and related research, and in general gives a good defense for Carlson's approach. Lastly, in another paper Carlson and colleagues (Whittemore, Macpherson, and Carlson, (1991)) propose a method for enriching discourse understanding through relating event participants with 'open' roles in inter-clausally.

tions. More details about this mapping function and the part that ESK plays in it are addressed later, along with points about constraint posting and verb sense selection.

4.3 Identification of Semantic Sense

Clearly, the primary goal for MENLU is to generate stable and consistent end-form results, given often disparate and ambiguous input. The resolution of this input into a normalized form rests on determining the appropriate sense of a head word when confronted with alternatives. If we have the sentence containing a verb like *admit*, then the preference for the appropriate semantic sense is syntactically based; the *admit* that subcategorizes for a clausal complement is a verb of communication – *He admits that he smokes* –, while the one that takes an NP complement will be the social interaction verb we have already seen.[9] However, other common head words that have more than one semantic sense, say *give*, may not be distinguishable syntactically: *He gives money every week*; *He gave no symptoms of this yesterday.* Along with the featural specifications for variations on thematic roles, we also include in ESK semantic type restrictions on arguments. Satisfaction of such constraints helps to 'guide' the selection of the appropriate semantic sense in the case where the categories of syntax leave us none the wiser. For instance, the argument filling the theme of *give* in the second sentence being a MEDICAL_SIGN lets us know that we are talking about the "exhibit" sense and not the "transfer" sense of the verb since the "exhibit" sense of *give* posts a semantic restriction on possible themes that limits them to things that are typically 'exhibited'. Sense selection is also aided by a method that checks for the observed 'integration successfulness' of a sense in terms of the number of its roles that have been filled. The 'integration success' score for the verb *admit* that has filled roles for its primary roles of *agent* and *theme* and adjunctive roles of *purpose* and *loc.goal* will be higher than one with just filled *agent* and *theme* roles (see McRoy (1992) for a description of similar contributors to sense selection).

4.4 Remaining Interpretation

'Embellishment' of the type discussed above for verb-based events, is also desired for event interpretations syntactically expressed as nominal. The target representation for nominalized forms such as *admission into hospital* , *pulse-rate change*, *change in ecg*, *cigarette-smoking*, is the same as that of their verb counterparts. Lastly, for inter-clausal situations — such as in 'light verb' constructions — identifying participants in one clause that play a part in an event in another clause is desired. For instance in a sentence like *John underwent an operation*, we want there to be a link made between John, the subject of the matrix verb, and the patient role underlyingly available in the nominalized *operate* event.

[9] Actually, the -ing NP *He admits smoking* theoretically embodies this latter NP subcategorization, but as Alvey distinguishes syntactically between gerund and non-gerundial nouns, this does not pose a problem. Admit in other contexts also takes simple NPs as in *He admitted several crimes.*

In this section, we have described the semantic content of our event representation. We now move on to a description of the ESK knowledge base used to help derive these representations.

5 Architecture

The object-oriented system CLOS (CommonLisp Object System) was chosen for representing ESK. CLOS naturally incorporates type hierarchies and employs slots/messages inheritance. With object-oriented systems generalities are simpler to maintain — general information can be infused from higher-order types to sub-types and instances they dominate, and with enhanced versions, like CLOS, inheritance can be manipulated to open up methods for non-monotonic knowledge base techniques, permitting users to specify values for slots when default inherited values are to be overridden. With CLOS there is also the option of multiple inheritance. With multiple inheritance a single class can receive specifications across several 'higher-order' classes. Multiple-inheritance-based designs not only permit very rich representations to be cast, but the organizational properties are also aesthetically and functionally pleasing. For instance, ESK is organized by way of five main groups: Semantic_Concepts, Proposition_Types, Semantic_Relations, Word_Entries, and Syntax. The subclasses and instances of the Word_Entries class are used as a focal point to which information from all the other classes is inherited. Members of Word_Entries essentially are verb-frames which draw their contents from the other major branches: syntactic specifications from the Syntax super-class, labels for thematic roles from the Semantics_Relations, etc. If global modifications are required, they can be made at any of the major super-classes and all changes will be inherited down to the individual verb frames. (see Russell, et al., 1992, for a description of a CLOS-based lexicon and Kiczales, et al., 1992, for a description of and methods for developing a CLOS-based system)

The authors also suggest Manandhar (1993) and Sanfilipo (1990) for perspectives on rigorously typed feature-based approaches to lexicons.

6 Contents of ESK

ESK is organized as a default- inheritance network and contains information relevant to characterizing events and states with a bias towards semantics. Currently, the information it contains is as follows:

- thematic roles
- semantic types for arguments
- expected adjuncts for each verb
- syntactic information
- linking specifications for 'light verbs'
- event (or propositional types)
- correspondences between syntactic and semantic representations
- WordNet verb concept types.

- WordNet sense indices (WordNet 1.4 Miller, et al., 1993)

In this section we discuss briefly the notions above, along with some of the motivation behind its current implementation of ESK.

6.1 Thematic Roles

As discussed above, thematic roles are defined within ESK as feature/value pairs and are tuned according to the particular role an entity plays for a given verb. We do **not** intend for the set of feature/value pairs assigned a role to be construed as fully defining it. The features are more intended as a minimal specification from which other implicatures can be formulated in order to provide fuller (but again not complete) accounts of the roles played.

The default inheritance available lends itself to easy manipulation of the defining features. For each thematic role we have a default set of features available and individual verbs can use the default mechanism to inherit individual role features. Also briefly discussed above was the manifestation of thematic roles in FLFs. We employ typical role names in FLFs, but retain in underlying ESK verb representations the list of feature/value pairs used to minimally specify the role.

In ESK we use 'typical' role names (like the ones in FLFs) as slots in verb/state frames and add further refinements by way of feature clusters in their slot values. Currently, within ESK the value of a thematic role slot is somewhat complicated in that we also use the slot-value to retain other information as well. Besides the features that refine thematic-role notions we also use the slot-value to specify semantic-type restrictions on entities that can fill the slots (e.g. [human +] and [inanimate +]) and where appropriate we allow a third information-type to express *factivity* and *degree of implied factivity.* This information helps to further specify the semantics regarding objects of attitudinal verbs such as those in the sentences:

We suspected myocardial infarction

I believe a further myocardial infarction

The slot-value of a thematic role is maintained as a tri-part value, vis-a-vis an ordered list of three members, with the first reserved for semantic-restrictions, the second for thematic-role features, and the third for 'factivity' notions (we borrow from Givon (1984) for our factivity assignments).

We have mentioned that we do not include additional thematic-role featural specifications in FLF (and the same is true of 'factivity' statements, but we do post semantic restrictions in FLF) mainly because we wish to not encumber FLFs with too much information especially since this information can be retrieved from ESK when needed. We have another reason for retaining simple labels in FLF — so that we can more easily take advantage of much of the research regarding thematic roles. From an engineering perspective having the simpler labels to latch onto offers a straightforward mechansism for dealing with event/state arguments A case in point is Nishigauchi's (1984) work on purpose clauses in which he provides an account of how interpretation of the controller of purpose clauses, for many (most?) cases can be aided by finding the last available 'goal' as in:

Bill bought for Susan$_i$ a flashy car to [PRO$_i$ to drive]. John$_i$ received from Susan a book [PRO$_i$ to read]. Though his account has been proven to not accurately handle all such cases (i.e. Ladusaw and Dowty (1988) because deeper pragmatic knowledge is often needed, if an NLU is not capable of reaching indepth understanding of such situations, a simple inference employing a method like Nishigauchi's to calculate controllers has, at the least, 'promise'.

Nishigauchi incorporates a thematic role hierarchy to determine which entities are likely to be 'goals' and such hierarchies have enjoyed a wealth of popularity in the related literature. Besides Nishigauchi, Grimshaw (1990) gives a very promising account of limiting passive constructions for both verbs and nominalized verb forms through a combination of a thematic-role and aspectual-oriented hierarchies. Hierarchies used for limiting surface forms are typically ordered in terms of externality. Roles that more characterize arguments found in subject position are 'external' and those with less subject-like qualities are 'internal'. Grimshaw's hierarchy, for example, is ordered in this fashion.

(Agent(Experiencer(Goal/Source/Location (Theme))))

Currently, the list of predominant thematic role labels employed in ESK are:

Agent, Theme, Patient, Instrument, Goal, Source, Experiencer, Stimulus, Actor, Measure, Identical_Object, Referent_Object.

Their default definitions more or less correspond with Jackendoff's descriptions.

We are also currently working on the set of variations on the roles. Some examples of variations on 'patient' are:

```
+effected --- patients of creation
+destroyed --- patients of eradication.
+altered -- altered patients
```

6.2 Argument Semantic Types

Each thematic role contains information regarding the expected (or presupposed type) for its filler. When producing FLFs MENLU employs a type-checking scheme for validating fillers for argument roles. The posted types are included within the slot-value for individual event/state thematic roles as described above. The actual types employed in ESK are incorporated within a simple domain-specific type hierarchy that resides outside of ESK (see Section 2).

6.3 Expected Adjuncts/Prepositional Phrases

One of the aims for MENLU is to incorporate knowledge about the relationships between different verb forms and adjuncts. Expectancy about adjuncts is couched in terms of functional/semantic roles. Verbs of movement, for instance, characteristically co-occur with adjuncts denoting a *goal location.* These verbs have an available *loc.goal* as an 'expected role'. Prepositions in turn are part of ESK and have specifications for the functions they serve, the semantic type of structure to which they attach, and the semantic type that can serve as an object to the preposition. MENLU has a prepositional phrase interpretation scheme that scores and ranks interpretations for available prepositions taking into account type constraints and 'expectancies'.

6.4 Syntactic Contributions

ESK also provides a network for aligning results from parses with proper verb senses and thematic role combinations. ESK maintains a separation of syntactic and semantic information, providing the potential to exchange the syntactic segment from one theory to another. MENLU incorporates the Alvey/GPSG-based software [12], and ESK has in its syntax-related hierarchy information reflecting Alvey's GPSG-grammar. The information contained in this section of the hierarchy pertains to predicate/argument alignments (via culling information from available subcategorization features) and is used to align syntactic arguments (e.g. subjects and direct objects) with thematic arguments (e.g. agents and patients). Each meld of syntactic and semantic argument structures results in an ESK class entity that can then be inherited into subsequent verb frames.

An example of how syntactic and semantic information comes together into a single verb frame is as here:

```
AGENT        ((HUMAN)...)
PATIENT    ((HUMAN)...)
ROLE_ALIGNMENT
  (((|agent| |human|) ?X0)
   ((|patient| |human|) ?X1))
ARG0                    ?X0
ARG1                    ?X1
SYN_RULE                VP/NP
```

From the syntactic knowledge stems information regarding the syntactic rule (which reflects subcategorization requirements) and the syntactic arguments encapsulated by the conventionalized use of ARG0, ARG1, etc. to represent subject, direct object, etc., respectively.

When ESK was first designed and implemented, explicit syntactico-semantic links were incorporated in the fashion just described. However, as we learned more about other research regarding such linking and as we gained confidence in believing that we could actually incorporate the kind of information we desired, it has become apparent that we might be able to forego the explicit syntactico-semantic linking component. Grimshaw (1990) promotes such a notion and suggests that a combination of a thematic role hierarchy and aspectual knowledge would be enough for creating syntactic arguments to underlying thematic roles and ESK contains both aspectual and thematic role knowledge for individual verb types. We have not yet made this jump, but in time such an exploration will be made.

6.5 Light Verb Links

ESK contains linking specifications that hold between matrix verbs and embedded deverbal event arguments (such as with 'light verb' constructions, e.g. she(agenti) made a [(agenti) promise]). ESK provides a simple language in which such links can be declared. Currently, this feature of ESK is very experimental.

6.6 Proposition Types

ESK includes event (or propositional) type labels that organize events into one of four types: process, achievement, accomplishment, and state. The general proposition type distinctions will provide support for dealing with verbs' aspectual character (see Somers' survey (1987:80)).

6.7 Single Syntactic Entries Mapped to Multiple Semantic Senses

Part of the inheritance pathway within ESK has been designed for establishing correspondences between single syntactic surface representations and multiple underlying conceptual forms. This mapping involves an indexing scheme with the semantic index of Alvey lexical entries co-indexed with a class held in ESK. The class in ESK can potentially be linked to sub-classes from which multiple semantic-senses arise. These correspondences are currently being used for cases involving verbs such as *give* in sentences like *The patient gave a history of cardiovascular disease* where there is not only the potential for ambiguity in semantic senses (vis-a-vis the "transfer" sense and the "exhibit" sense), but also the potential ambiguity regarding whether or not underlying implicit roles are involved; the former ("transfer") sense always has an implied beneficiary which is usually explicit in the corpus, while the latter ("exhibit") does not demand an "experiencer/goal" role[10]

6.8 Verb Concept Types

For verb-based semantic (or conceptual) types, ESK draws from WordNet. WordNet organizes the verb portion of its semantic network into 14 categories: Competition, Change, Consumption, Contact, Motion, Emotion and Psych, Stative, Social, Interaction, Weather, Perception, Possession, Cognition, Creation.

Though individual ESK senses are linked to individual WordNet senses (and thus information regarding the top-most super-classes could be obtained) we made the decision to also maintain these primary conceptual types directly in ESK. The reason for directly including these in ESK is that when MENLU attempts analysis we do not anticipate having an active link into WordNet.

[10] It is arguable that *He gives a history of heart disease* does always have an implied experiencer.

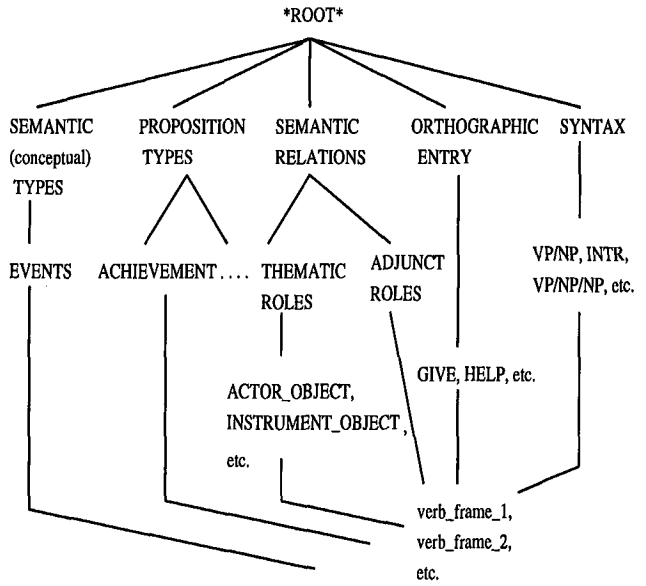

Figure 1: An overview of the ESK hierarchy

```
<GIVE.VP/NP//AG_TH.VCOMM>
 WNS                    (5)
 AGENT    (HUMAN)
 ROLE_ALIGNMENT
       ((((|agent| |human|) ?X0)
         ((|theme| |evidence|) ?X1))
 ARG0                   ?X0
 SUBCAT            "(SUBCAT NP)"
 EXAMPLE        "he gives a history"
 SYN_RULE               VP/NP
 NUMB_SYNCAT            (1)
 ARG1                  ?X1
 ENGLISH-DESCRIPTION
         ''The communication sense for give''
 VERB                   GIVE
 THEME  ((evidence) nil ((implicative +))*
```

** the nil value in the theme slot is used to say that nothing further is added to the default value for theme (namely (movable +) and the (implicative +) is used to say that the theme is positively implied.*

Figure 2: An example verb frame for the communication sense of the verb 'give'.

6.9 Synonyms, Hypernyms, Hyponyms, and Antonyms

Each sense in WordNet is placed within a hierarchy of senses. Super-types of individual entries correspond to hyponyms, sisters to synonyms, and daughters to hypernyms (see Fellbaum, 1992) . An example of this is provided below. ESK taps into WordNet by way of establishing for each of its own senses a link to the best corresponding sense in WordNet, thus making available the benefits of WordNet to ESK. This step also saves the developers of ESK from developing its own conceptual hierarchy.

```
Sense 6 of the verb OPERATE
operate on, operate, perform surgery on
 => treat, doctor -- (provide treatment for)
```

6.10 Overview of the Hierarchy and Example Frame

An overview of the ESK hierarchy is given above in Figure (1). As noted, there are several higher-order superclasses from which information is inherited down to individual verb frames, an example of which is given in Figure (2). Generalizations are made at higher levels of the organization and can be overridden at lower levels. The net result is a set of individual verb frames which contain the bulk of necessary information that MENLU requires in order to embellish and disambiguate information regarding events and states.

7 Future Exploration

With MENELAS being a message understanding task, we are interested in measuring the benefit in precision and recall that is gained per 'unit' of additional linguistic support. Of particular interest to our group is the degree to which rich verb information increases data-extraction performance and what types of event knowledge benefit the extraction process the most. Likewise, we want to know the amount of 'extra' processing capability beyond our current level we need. Having a full ESK representation in place will help us to make these kinds of exploration possible.

8 Conclusion

ESK provides a detailed representation of the verbs and states found in the MENELAS English PDS corpus. Its structure and contents not only make it possible for MENLU to achieve rich interpretation of states and events, but it also provides a means for inferring additional implicatures regarding events and event participants.

References

[1] Ades, A., Steedman, M., 1982. On the Order of Words *Linguistics and Philosophy*, 4, pp. 517-558.

[2] R. Beckwith, Fellbaum, C., Gross, D., and Miller, G., 1990. WordNet: A lexical database organized on psycholinguistic principles. *Cognitive Science Laboratory Report 42*. Princeton University.

[3] Bresnan, J., Kanevra, J., 1988. Locative inversion in Chichewa: A case study of factorization in grammar. CSLI TechReport, August, 1988.

[4] Calder, J., 1991. The Pleuk grammar development system. University of Edinburgh.

[5] Carlson, G., 1984. Thematic roles and their role in semantic interpretation *Linguistics*, 22, pp. 259-279.

[6] Carlson, G. and Tanenhaus, 1988. Thematic roles and language comprehension. In Wilkins, W. (ed.) *Syntax and Semantics, Vol 21*, Academic Press: NY.

[7] Dowty, D., 1991. Thematic Proto Roles. Subject Selection and Lexical Semantic Defaults. *Language*, 67:3, pp. 547-619.

[8] Fellbaum, C., 1990. English Verbs as a Semantic Net *Cognitive Science Laboratory.* Princeton University.

[9] Fillmore, C., 1968. The case for case. *Universals in Linguistic Theory*, Bach and Harms (eds.), pp 1-88.

[10] Givon, T., 1984. *Syntax A Functional-Typological Introduction*, vol. 1. John Benjamins: Amsterdam/Philadelphia.

[11] Grimshaw, J., 1990. *Argument Structure.* Cambridge, MA:MIT Press.

[12] Grover, C., Carroll, J., and Briscoe, T., 1993. The Alvey Natural Language Tools Grammar (4th Release). Technical Report no.284. University of Cambridge Computer Laboratory.

[13] Jackendoff, R., 1990. *Semantic Structures.* Cambridge, MA: MIT Press.

[14] Kiczales, G., des Rivieres, J., and Bobrow, D.G., 1992. *The Art of the Metaobject Protocol.* Cambridge, MA: MIT Press.

[15] Ladusaw, W., Dowty, D., 1988. Toward a nongrammatical account of thematic roles. in *Syntax and Semantics, Voume 21*, Wilkins (ed.) Academic Press, San Diego, CA.

[16] Manandhar, S., 1993. Relational Extensions to Feature Logic: Applications to Constraint Based Grammars PhD. Thesis, Dept. of AI, University of Edinburgh

[17] McRoy, S., 1992. Using Multiple Knowledge Sources for Word Sense Discrimination *Computational Linguistics*, 18(1), pp. 1-30.

[18] Nirenburg, S., Defrise, C., 1993. Lexical and Conceptual structure for knowledge-based machine translation. *Semantics and the Lexicon*, J. Pustejovsky (ed.), pp. 291-323

[19] Nishigauchi, T., 1984. Control and the Thematic Domain. *Language*, 60(2), pp. 215–250.

[20] Pollard, C., Sag, I., 1994. *Head-Driven Phrase Structure Grammar* University of Chicago Pressand CSLI Publications Chicago, IL. and Standford, CA.

[21] Russell, G., Ballim, A., Carroll, J., and Warwick-Armstrong, S., 1992. A practical Approach to Multiple Default Inheritance for Unification-Based Lexicons *Computational Linguistics*, 18(3), pp. 311–337.

[22] Sanfillipo, A., 1990. Grammatical Relations, Thematic Roles and Verb Semantics PhD. Thesis, Centre for Cognitive Science, University of Edinburgh

[23] Talmy, L., 1985. Force Dynamics in Language and Thought. (cited in Jackendoff, 1990.) In *Papers from the Twenty-first Regional Meeting of the Chicago Linguistic Society.* Department of Linguistics, University of Chicago, Chicago. IL.

[24] Somers, H.L., 1987. *Valency and Case in Computational Linguistics.* Edinburgh Press: Edinburgh, Scotland.

[25] Whittemore, G., Brunner, H., and Ferrara, K., 1990. Empirical Study of Predictive Powers of Simple Attachment Schemes for Post-modifier Prepositional Phrases. In the *Proceedings of the 27th Annual Meeting of the Association of Computational Linguists.* Pittsburgh, PA.

[26] Whittemore, G., Macpherson, M., and Carlson G., 1991. Event-building through Role-filling and Anaphora Resolution. In the *Proceedings of 28th Annual Meeting of the Association of Computational Linguists.* Berkeley, CA.

[27] Wilkins, W. Culicover, 1986. Control, Pro and the Projection Principle *Language* 62(1), pp.