FrameNet and the Linking between Semantic and Syntactic Relations

(The author apologizes for submitting a padded outline instead of a full-blown paper. The presentation itself will include data samples and software demos, or simulations thereof.)

1. Introduction.

My motivations in offering this talk are two-fold. First, I would like to acquaint this community with the goals, procedures, and the ultimate products of a research program with which I have been affiliated during the past five years. My reason for this is that I believe the kind of lexical resource we are trying to create can serve linguistic and NLP researchers in a number of ways, just as I hope that it will be possible to improve our resource by adapting and mobilizing data from other existing sources. Secondly, I want to define a task for which the data provided through this project can serve as input and as a tool. That task is to derive from the data developed in the main project a secondary resource consisting of structured clusters of lexical items (called *kernel dependency graphs*) in which each such cluster contains a governor and the lexical heads of all of its dependents, each of these marked for the nature of the semantic role it bears to the governor. These kernel dependency graphs when based on already annotated sentences can provide information about collocational and selectional requirements of lexical items, and when automatically discovered in text will make possible such operations as word sense disambiguation and topic detection.

The project in question is called FrameNet.[1] It is funded by the U.S. National Science Foundation and is administered at the International Computer Science Institute in Berkeley. The project employs about a dozen people - slightly more in the summer months - and over the years has also benefited from contributions of labor and wisdom from numerous consultants, volunteers and associates. It is basically a lexicon-building effort; as such it requires an amount of human input of the kind avoided on many computational linguistic projects, but there are various NLP applications which it is intended to serve, among them word sense disambiguation, information extraction, question answering, topic detection, and machine translation. The data, descriptions of the methodology and software tools will soon be made available to researchers and should be adaptable to lexicon-building efforts in specialist domains. Co-PIs in charge of the affiliated applications are Dan Jurafsky, currently working on question-answering, Srini Narayanan, dealing with information extraction, and Mark Gawron, using FrameNet data as part of a machine translation project. Word sense disambiguation is of course a necessary underpinning for all of these operations.

[1] The project is supported under NSF grant ITR/HCI 0086132, "FrameNet++: An On-Line Lexical Resource and Its Application to Speech and Language Technology." PI is Charles Fillmore. The public website is http://framenet.icsi.Berkeley.edu/~framenet/.

Situating FrameNet. FrameNet differs in several ways from a number of other existing lexical resources.

- Machine-readable dictionaries. The FrameNet work involves careful examination of corpus evidence and generally leads to several sorts of information, in terms of usage and combinatorial properties, that are generally not recorded in any standard dictionaries. Though many kinds of information can be derived from currently available machine-readable dictionaries (Wilks et al. 1995), they are nevertheless essentially limited to the same information found in the print dictionaries on which they are based.
- WordNet. The WordNet database[2] is much larger than FrameNet, but it has very limited combinatorial information, and what it has is limited to verbs, It fails to connect semantically related words in different parts of speech, and it does not connect semantic and syntactic information. WordNet links words by a large variety of traditional lexical semantic relations (starting with synonymy, since the basic units are synonym sets - synsets), FrameNet lacks any direct way of showing such paradigmatic relationships, but does show word-to-word relations by way of showing membership in particular semantic frames, and indirectly by showing relations between frames.
- Levin Verb Classes. The verb class inventory constructed by Beth Levin (Levin 1993), a rich source of information on variation in combinatorial patterns for English verbs, is based on linguists' intuition rather than corpus attestations, is limited to verbs, and in fact only verbs that exhibit valence alternations. Many of the verb groupings correspond to FrameNet frames, but many do not.
- PropBank. The Penn "PropBank" database[3] describes argument structures found in sentences in the gold standard subsection of the Penn TreeBank, but only for verbs, and at present only for the highest verb in each sentence; it has only minimal specification of the connections between the argument types and semantic roles. FrameNet, by contrast, is as interested in examining the argument structure (frame structure) of frame-bearing nouns and adjectives as much as in verbs. (The possibility of blending PropBank and FrameNet data is under discussion.)
- COMLEX. The COMLEX database[4] has syntactic subcategorization frames only, but these are not separated according to the word's meanings, and the components of the sub-cat frames are not connected to semantic roles. Valence descriptions produced by FrameNet are designed to display the structured ways in which semantic roles, for a given sense of a word, are paired with their syntactic realization in terms of grammatical functions and phrase types.
- Dependency Database of Dekang Lin. The invaluable dependency database maintained by Dekang Lin[5], built on a dependency grammar parse of newspaper text, does not sort dependency links according to the senses of the words at either end of the dependency. Though limited to sentences that happen to have been

[2] http://www.cogsci.princeton.edu/~wn/

[3] http://www.cis.upenn.edu/~ace

[4] http://www.cs.nyu.edu/cs/faculty/grishman/comlex.html

[5] http://www.cs.ualberta.ca/~lindek/demos/dep.htm

annotated through the FrameNet process, FrameNet database makes it possible to browse for examples exhibiting a wide variety of conditions, e.g., nouns that head the Topic constituents in verbs of conversation, and the like.

- Word Sketch. Adam Kilgarriff's Word Sketches[6] offer information highly relevant to lexicographers, in displaying statistically relevant collocations of each word. Kilgarriff's work is developed statistically from the BNC, whereas FrameNet is based on manual tagging of sentences exhibiting selected senses of each word.

Many of these resources serve as checks and progress measures for the FrameNet efforts, all of them are consulted by members of the FrameNet team, and several of them invite the possibility of combining results, either directly into a single resource or through hyperlink paths.

The central activity of FrameNet can be summarized as follows:

- It groups *lexical units* (in the sense of Cruse 1986, pairings of words and senses) from the general vocabulary of contemporary English into sets according to whether they permit parallel semantic descriptions. (Thus, in the appropriate senses, *tell*, *inform* and *notify* would be assembled into a single set - along with many others.)
- Targeting these words one at a time, it examines sentences containing them extracted from a very large corpus. (Until now this has been mainly the British National Corpus[7], but we are currently adding the newswire texts from the Linguistic Data Consortium plus some separately acquired texts for special purposes.)
- It identifies the structure and the components of the semantic frame underlying each such group of words, characterizes the senses of the words in terms of the semantic frames that the word evokes.
- Through the semantic roles or frame elements determined for each relevant frame, it documents the ways in which phrases that accompany each word fit the semantic and syntactic combinatorial requirements of each lexical unit by manual annotation of sentences exemplifying its uses. (Thus, among the larger set of words inviting the expression of Speaker, Addressee, Message and Topic, the verbs *tell*, *inform* and *notify* agree in their ability to express the Addressee as the verb's direct object.)
- Automatic processes assemble information from XML record of the annotated sentences and produce a variety of reports displaying the attested combinations of semantic roles and grammatical realization features, the central ones of these being valence descriptions, combining semantic and syntactic combinatorial requirements and privileges.

The general division of labor for the lexicon-building portion of FrameNet work is something like this

- Computer scientists on the project have created software for implementing or assisting in all the operations of (1) corpus manipulation, (2) annotation, and (3)

[6] http://www.itri.bton.ac.uk/~Adam.Kilgarriff/WORDSKETCHES/
[7] http://www.hcu.ox.ac.uk/BNC/

editing, and for (4) storing the information resulting from those operations in a MySQL database.

- Linguists with training in syntax and semantics choose vocabulary sets for investigation, explore the properties of the semantic frames that underlie the meanings of the words in each such set, and, after scanning KWIC samples of a given lexical unit, prepare initial descriptions of the meanings and forms needed for describing the combinatorial properties of the words.
- Student annotators equipped with such initial descriptions examine sentences that can serve for illustrating the meanings and discovering their valence patterns, select sentences representing typical uses of the target word, seeking to exemplify each basic pattern found in the corpus, and annotate these sentences by tagging the constituents that express frame-relevant semantic roles and associate such constituents with the names of the semantic role (*frame element*). They also check and edit the automatically assigned grammatical function and phrase type on each of the phrases they tagged.
- Computer scientists have designed the means of displaying and summarizing these results, and are providing representations that permit the intended applications, and viewing and searching the resulting database.

A suite of pre-defined reports has been created in-house; our Japanese associate, Hiroaki Sato, has created a MySQL viewer and browser that enables a wide variety of queries on the FrameNet data. (The Sato viewer is accessible through the FrameNet website.)

The kinds of information made available for each valence-bearing lexical unit include

- its membership in a particular semantic frame. (Thus the pairs of lexical units with the word-forms *argue* or *argument* are separated according to their participation in a frame that has to do with Quarreling or one that has to do with Reasoning. That is, each of these frames contains each of these words, as different lexical units. The noun, but not the verb, participates in a separate technical frame, not currently covered in the project, as when it is used to refer to the argument of a variable.)
- the semantic roles, or so-called frame elements, that figure in the description of each frame. (In the case of Quarreling that would include reference to the Interlocutors, the issue or Topic of their disagreement, and so on.)
- the manner in which the frame elements are syntactically realized in sentences found in the corpus. (The Interlocutors in the Quarreling frame are found expressed either jointly, in a plural NP - where the phrase would receive the label Interlocutors - or disjointly, where one of the parties represented obliquely with the preposition *with* and the two elements are labeled Interlocutor-1 and Interlocutor-2. (*Jones and Smith argued* vs. *Jones argued with Smith*.) The Topic, in the case of Quarreling but not in the case of other frames of Conversation, can be expressed with either the preposition *over* or the preposition *about*: *Jones and Smith argued over the inheritance*.)
- the binding of the frame elements of a given frame with the higher-level or more abstract frames whose properties it inherits. (Thus, words in the Quarreling frame have some of their properties accounted for by the fact that Quarreling is a

subtype of Conversation, and others by the fact that it is also a subtype of Disagreeing.)

Much of this information is made available because the annotation process includes both the manual work done by human annotators in identifying the semantic roles for each frame, and the assignment of grammatical functions and phrase types which is done automatically but subject to manual editing. Some of the information, however, requires a certain amount of *metalexicographic* work on the part of analysts seeking to discover and represent generalizations about frame-to-frame relationships, cross-frame similarities, etc.

Valence: the semantic role component. The semantic part of the valence descriptions goes beyond the traditional thematic roles or case roles. The usual set of general-purpose semantic roles includes a limited list of concepts like Agent, Patient, Theme, Experiencer, Instrument, Source, Goal, Path, Location, and so on. (Fillmore 1977, 65; for a differently structured list, Somers 1987, 206) It has been our experience that the "standard" lists do not cover all of the semantic roles needed for the description of our frames, and distorting their interpretation for the sake of staying with the limited list does not seem helpful. For example, in a sentence like *you risked death*, it is hard to imagine connecting either the subject or the object of that sentence to any of the standard roles; in a sentence like *after Harry died, I replaced him on the committee* the direct object of *replaced* is clearly not somebody affected by the replacing act; and in *Harry resembles my cousin* and countless other examples, we see no reason to force the semantic roles we find into a ready-made inventory. Hence, in FrameNet we depend on frame-specific role names, and insist that in principle the frame element name used in one frame needs to be defined specifically for that frame without requiring us to show its commonalities with the role that received the same name in another frame. (Recall the terms Speaker, Addressee, etc., proposed for the speaking words.)

We have no interest, of course, in neglecting the numerous well-known generalizations across the semantic roles of predicates, generalizations having particularly to do with predictions of syntactic realizations dealt with in the literature on *linking.* FrameNet efforts to reflect such generalizations are not made on the run, while the annotation work is going on: rather, we hope to be able to capture such phenomena at a time when we feel we have the whole picture in view, perhaps in terms of frame inheritance, frame-structure homologies, or the like. Straightforward reduction to the abstractions proposed by various authors - e.g., Van Valin, Dowty - have not been of help.

Valence: the syntactic component. The syntactic part of the valence descriptions in the FrameNet efforts goes beyond what is usually thought of as argument structure or subcategorization-plus-subject, since the constituents that are frame-relevant are not limited to those that occur only in major syntactic positions or that are obligatory. In general, we try to characterize the central conceptual structure of the frame and then we look for the ways in which information that completes or elaborates that structure gets expressed in sentences headed by the words that evoke the frame. In doing this, we distinguish *core* and *peripheral* elements, generally assuming that certain kinds of modifying structures are appropriate to large classes of predicates and do not need specification in individual low-level frames. Thus, Place and Time adverbials tend to be compatible with any kind of event predicate; Purpose and Attitudinal adverbials are compatible with any agentive frame; and so on. The valence descriptions reports can be asked to display only the core elements. Certain concepts which are peripheral with some predicates, of course, are core with others. Thus a Place modifier is has peripheral status with *execute*, core status with *banish*, peripheral with *buy*, core with *reside*; a Manner modifier is peripheral with *speak*, core with *behave* or *phrase*.

In the case of noun valences, we include the noun's complements, of course, but also a possessive determiner if it introduces information about a participant in the frame evoked by the noun (*my brother's decision*); arguments of support verbs; and since we are interested in analyzing compound nouns as well as noun-headed phrases, modifiers of nouns in compounds: in particular, modifying nouns, as in *child abuse*, or "pertinative" ("relational") adjectives, as in *educational policy*. The inclusion of the arguments of support verbs make it unnecessary to concern ourselves with questions about the constituent structure of certain phrases with support verbs and event nouns: thus we will find reason to tag the PP *about the President's policy* in the sentence *he made a statement about the President's policy*, as an element of the frame evoked by *statement*, without worrying about attaching the PP to the noun or the verb.

2. The new task.

The particular task for which we are beginning to use the FrameNet data concerns representing the mapping between a predicating word and the dependent semantic elements of a predication headed by that word and the lexico-syntactic realization of those elements. Here the goal is not one of using insights based on linguistic intuition for contriving and testing "linking rules" for predicting the connection between semantic function and syntactic realization, but of documenting the semantic/syntactic connections that we find attested in corpora. More specifically, I want to do this in terms of lexical dependencies, both collocational, in terms of word-to-word dependencies, and selectional, in terms of the slot-filling functions of particular classes of words or constituents.

Here I define predication as a semantic structure expressed as a governing word and its dependents, where the dependent elements are identified both semantically and in terms of grammatical function and form. In further particular, I would like to produce, both from annotated sentences selected from the corpus for illustrating lexicographically relevant information, and from further corpus material to which automatic annotation processes have been applied, an inventory of structures, conceived in a roughly dependency-grammar manner, creating a repertory of clusters of lexical sets to be called *kernel dependency graphs* (KDGs), in which each one contains a governing word and lexical dependents as discovered in representative uses of that governing word. (In the simplest cases lexical-unit to lexical-unit is sufficient, but for constructions in which lexical headedness is problematic, as in the case of minigrammars for such named entities as dates, addresses, person names and the like. (The nature of such elaborations will come up later on.)

To carry out this task we have needed to expand the basic activities of the project beyond what originally seemed to be the fairly simple and straightforward activity of identifying predicators, and locating and tagging their dependents as we find them in corpus sentences illustrating their typical use. Among the additional procedures we needed to develop are the following:

- For verbs, we have tagged the arguments of matrix structures which syntactically control grammatical positions (typically subject) in the target words. Thus, if we are annotating sentences with a verb like *interpret* in the cognitive sense, we could tag as the Cognizer (the one who makes the interpretation) the subjects of subject control Equi or Raising verbs or adjectives, the objects of object control Equi or Raising verbs, the oblique experiencers of certain kinds of adjectives and nouns (*fun, difficult*), the possessors of controlling nouns (*the mayor's attempt to interpret the statute*), or chainings of these. If we were doing this work with a parsed corpus - it would have to be a very accurately parsed corpus, of course - we could count on the structure of the sentences providing this information; but for the sake of being able to collect information about collocates, we have annotated these constituents as well.
- This decision made it necessary for us to devise a pseudo-grammatical-function "Ext" (external to the phrase headed by the target), since using the gf "Subject" would be misleading: the controlling NP of a nonfinite VP is often not a subject in its direct context.
- In the case of frame-bearing nouns, we identify any and all *support verbs*, and find the arguments of those verbs which are, of linguistic necessity, construed as participants in the event or state of affairs designated by the noun. (Consider the subjects of the verbs in *pay a compliment, make an announcement, say a prayer, wage war*, etc.) In the course of carrying this work out, we needed to expand the concept of support verb beyond the standard light verbs in the direction of the lexical functions of Igor' Mel'cuk. (Related work is being carried on in the NOMLEX project at NYU under Catherine Macleod; discovering such information has also been the mission of Thierry Fontenelle's dictionary-based derivations of Mel'cukian lexical functions for French and English.)
- Further developments required us to posit *support prepositions*. In the way that the combination of a support verb and a noun creates a verb-like entity, there are cases in which the combination of a preposition with a noun creates an adjective-like entity. Examples are *at risk, in danger, on fire,* etc.
- Also in the case of nouns, for the sake of acquiring meaningful collocational information, we have added various kinds of typical modifiers (such as size, color, style, etc., for garment nouns) as well as information that reveals the qualia structure of nouns, in the sense of Pustejovsky.
- For nouns in general, both frame-bearing and dependent, we have needed to identify the kinds of nouns that occur in N+of+N constructions in which it is the second noun, not the first, that has selectional or collocational relations to the context of the whole phrase. These nouns, which are called transparent to indicate that a collocation-detector sees through the syntactic head to find the semantic head, express such meanings as Types, Aggregates, Parts and Portions, Quantities, and Classifiers. (The syntactic pattern does not always determine

transparency interpretations: the verb-noun collocation in *eat a number of apples* involves transparency, that in *estimate the number of apples* doesn't.)

- We have added a separate kind of annotations for dependent nouns as targets. where we identify for each example sentence their governors and information about the boundaries of the phrases that mark the nature of their dependency, hoping to provide more detailed frame information in a later pass through the data.
- Also for frame-bearing nouns, we have needed to introduce an awareness of relational or *pertinative* attributive adjectives (the adjectives whose definitions contain the phrases "of or pertaining to") since such adjectives typically provide information about frame structure or qualia structure for the nouns that they modify.
- In addition to straightforwardly finding phrases that count as semantic dependents of frame-bearing words, we have expanded the question to a more general one, of just how information about the participants in a frame is encoded or understood in the annotated sentences. This requires us to say something about constituents which are understood but absent. Thus we recognize the non-instantiation of frame elements in particular constructions (constructional null instantiation, such as the missing elements of passives or imperatives, though this is in general is not lexicographically significant), but more importantly lexically-determined omission possibilities of two sorts, *existential null elements* (as with intransitive uses of verbs like *eat, sew, bake*, etc.) and *anaphoric null elements* (as for the missing complements of *she found out*, *we won*, *they've already arrived.*
- The same motivation has made it necessary for us to recognize cases where a frame element common to a particular frame is incorporated directly into a verb (*shelved the books, bottled the wine,* etc.) as well as cases in which two frame elements are signaled either in single or multiple constituents (*punched his nose, punched him in the nose*) or combined lexically in a single word (*cured the leper* vs. *cured* NP *of leprosy, appointed the chairman* vs. *appointed* NP *as chairman*)
- Having recently acquired the means of recognizing certain classes of named entities, such as addresses, dates, time-telling formulas, personal names, institutional names, etc., we hope eventually to be able to use this information for generalizing over the KDGs that we construct, replacing particular dependent elements with labels like Person, Institution, Place, Time, etc.

Appendix

Here are some invented examples showing KDGs and the sentences on which they can be derived. KDGs could be represented as dependency graphs in tree form, with lexical forms as the nodes and the branches labeled with frame element names; here we use a mock RDF format.

The idea is that frame-annotated sentences can provide input for a process that finds the target governor and displays the semantically relevant lexical heads of the dependents and associates with them the manner of their "marking" in the sentence and the semantic role, within a given frame, that they bear within the predication.

Simple cases:

1. *The boy caught a spider.*
(The frame elements are subject and object of a transitive verb.)

```
<KDG rdf:ID="1137864">
  <governor>catch</governor>
  <frame rdf:resource="Capture"
   <agent>boy</agent>
   <victim>spider</victim>
  </frame>
</KDG>
```

2. *The man was caught stealing a fish.*
(One of the frame elements is itself the governor of a second frame; the unexpressed agent of the passive is rendered as *SOMEONE.*)

```
<KDG rdf:ID="46823">
  <governor>catch</governor>
  <frame rdf:resource="Spotting2"
   <observer>SOMEONE</observer>
   <observed>man</observed>
   <act>stealing</act>
  </frame>
</KDG>
```

3. *The teacher talked to the students about ambition.*
(Two frame elements are represented obliquely: the prepositional "marker" is shown.)

```
<KDG rdf:ID=“21718644”>
 <governor>use</governor>
 <frame rdf:resource=“Talk”>
   <speaker>teacher</speaker>
   <addressee>to: students</addressee>
   <topic>about: ambition</topic>
 </frame>
</KDG>
```

Cases recognizing control and/or support verbs.

4. *The freshmen have to take a chemistry test.*

(The semantic governor is *test*, support verb is *take*; subject matter is expressed modifier noun in a compound; the "control" context *have to* is ignored.)

```
<KDG rdf:ID=“47623”>
 <governor>test</governor>
 <support>take</support>
 <frame rdf:resource=“Examination”>
   <examiner>SOMEONE</examiner>
   <examinee>freshmen</examinee>
   <subject>chemistry</subject>
 </frame>
</KDG>
```

5. *The senator paid me a compliment on my work.*

(The support verb is *pay*.)

```
<KDG rdf:ID=“9637615”>
 <support>pay</support>
 <governor>compliment</governor>
 <frame rdf:resource=“Compliment”>
   <speaker>senator</speaker>
   <addressee>me</addressee>
   <reason>on: work</reason>
 </frame>
</KDG>
```

Case recognizing transparent nouns.

6. *The majority of tobacco producers use a variety of asbestos in this kind of filter.*

(The syntactic heads are *majority*, *variety* and *kind*; the collocationally relevant dependents are different.)

```
<KDG rdf:ID=“256”>
 <governor>use</governor>
 <frame rdf:resource=“Use3”>
   <agent>tobacco producers</agent>
   <ingredient>asbestos</ingredient>
   <product>in: filter</product>
 </frame>
</KDG>
```

Predication within a NP

7. *our religious discussion*

(Interlocutors expressed as possessive determiner, topic as pertinative adjective; all relevant frame elements expressed within the NP.)

```
<KDG rdf:ID=“143301”>
<governor>discussion</governor>
 <frame rdf:resource=“Dialogue”>
   <interlocutors>genitive: we</interlocutors>
   <topic>pertinative: religion</topic>
 </frame>
</KDG>
```

References

Bouillon, Pierette, Vincent Claveau, Cécile Fabre, Pascal Sébillon (2002), "Acquisition of qualia elements from corpora - evaluation of a symbolic learning method", *LREC 2002: Third International Conference on Language Resources and Evaluation,* Vol. 1, pp. 208-215

Cruse, D. A. (1986), Lexical Semantics. Cambridge: Cambridge University Press.

Fillmore, Charles J. (1977), "The case for case reopened," in: Peter Cole and Jerry Sadock, eds., *Syntax and Semantics, Vol. 8: Grammatical Relations*. New York, 59-??

Levin, Beth (1993), *English Verb Classes and Alternations: A Preliminary Investigation*. Chicago: University of Chicago.

Mel'cuk, Igor' A., et al. (1984), *Dictionnaire Explicativ et Combinatoire du Français Contemporain. Recherches Lexicosémantiques* 1. Montréal.

Pustejovsky, James (1995), *The Generative Lexicon*, Cambridge: The MIT Press.

Somers, H. L. (1987), *Valency and Case in Computational Linguistics.* Edinburgh.

Wilks, Yorick, Brian Slator and Louise Guthrie (1995), *Electric Words*, Cambridge, MA: MIT Press.