

Building a Large Lexical Databank Which Provides Deep Semantics*

Charles J. Fillmore, Charles Wooters, and Collin F. Baker

International Computer Science Institute

1947 Center Street

Berkeley, CA 94704, USA

{fillmore,wooters,collinb}@icsi.berkeley.edu

Abstract

This paper reports on the design of a lexical database for English which is currently under construction ("FrameNet-2"), and describes the kinds of linguistic facts that the database is intended to make available, for both human and computer consumers. Building on a recently completed pilot study ("FrameNet-1"), it is centered on the nature of the relation between lexical meanings and the conceptual structures which underlie them (**semantic frames**). The database will show the semantic and syntactic combinatorial possibilities (based on frame membership) of the lexical items it includes, as these are documented through grammatical and semantic annotations of sentences extracted from a large corpus of contemporary written English. The notions of **profiling** within a frame, **frame inheritance** including **multiple inheritance**, **frame blending**, and **frame composition** will be explained and illustrated, and the manner of storing information about them in the database will be outlined. The building of the database, with its necessary labor-intensive manual component, will be explained.

1 Introduction

It is generally assumed that the goals of computational linguistics and those of ordinary linguistic analysis are unlikely to coincide in the area of lexicon building. The full range of the knowledge speakers have about the words in their language appears to be out of reach of practical efforts to acquire and organize such knowledge for efficient use in NLP applications.

FrameNet, as a computational linguistics project, aims at a compromise between two extremes. At the one end it is clearly possible to do careful and subtle conceptual analysis of the items in a small lexicon designed for restricted purposes in a narrow domain, where the goal of achieving reliable language understanding and accurate inference generation seems attainable. At the other end, one can aim at fairly superficial information retrieval or other NLP applications by using statistical methods for acquiring very large lexicons or by creating clever programs for making maximal use of the information available in machine-tractable dictionaries. The latter contain large amounts of data but are limited by what their compilers believed would be relevant to human users (Wilks et al., 1996).

The FrameNet team are convinced that at a certain level, the depth and subtlety of careful linguistic analysis can be attained using the wisely exploited judgments of linguistically trained

* The authors are Principal Investigator, System Designer, and Project Manager, respectively, of the National Science Foundation sponsored project, "FrameNet++: An online lexical semantic resource and its application to speech and language technology", funded under Grant No. ITR/HCI 0086132. The project is administered at the International Computer Science Institute in Berkeley, California. The other Principal Investigators of FrameNet++ are Dan Jurafsky (University of Colorado at Boulder), Srinu Narayanan (SRI International), and Mark Gawron (San Diego State University).

researchers who annotate syntactically sorted sentences¹ taken from a large natural language corpus. Some of the information made available by such research will be recorded directly by the annotators, and some will be derivable—partly manually, partly automatically—from the results of such annotations. The labor-intensive nature of such research will be eased in several ways: by computational facilitation of the annotation process itself; by computing maximal likelihood estimates for certain levels of analysis, so that researchers can simply accept or reject these judgments; and by allowing the annotation to proceed automatically in some cases, at least for patterns that appear to be clear. All such automatically created data will be vetted by human reviewers.

2 Frame Semantics Background

The basic assumption of Frame Semantics as it applies to the description of lexical meanings is that each word (in a given meaning) evokes a particular frame and possibly profiles some element or aspect of that frame.² An “evoked” frame is the structure of knowledge required for the understanding of a given lexical or phrasal item; a “profiled” entity is the component of a frame that integrates directly into the semantic structure of the surrounding text or sentence. The frames in question can be simple—small static scenes or states of affairs, simple patterns of contrast, relations between entities and the roles they serve, or possibly quite complex event types—which we can call **scenarios**—that provide the background for words that profile one or more of their phases or participants.

For example, the word *bartender* evokes a scene of service in a setting where alcoholic beverages are consumed, and profiles the person whose role is to prepare and serve these beverages. In a sentence like *The bartender asked for my ID*, it is the individual who occupies that role that we understand as making the request, and the request for identification is to be understood against the background of that frame.

The background frame can be a “history” and the profiled entity can be some person or object that figures at some point in that history. Thus the word *widow* evokes a complex historical frame in which a woman is seen as having a particular social status because the man to whom she was (most recently) married has died. The frame requires an understanding of marriage, family, death, and social status. A sentence like *She was a widow for three years* permits the interpretation that there was an interval of three years between the death of her husband and some other salient status change, such as a remarriage or perhaps her own death. The complementation syntax of the word *widow* allows mention of the deceased husband, either with a possessive determiner (*John’s widow*) or with a post-nominal *of*-phrase (*the young widow of a famous general*).

The kinds of semantic frames that have interested FrameNet the most are those that support the meanings of large numbers of words. One of these is what can be called the **REPLACEMENT** frame. An abstract description of this frame requires the positing of a Place (this can stand for a role, a function, a location, a job, a status, etc.), and two distinct entities (X_1 and X_2) that both have histories with this Place. We can then line up the separate pieces of the **REPLACEMENT Scenario** as follows:

¹It may be misleading to refer to this as simply “annotating sentences”: each of the sentences is seen as a concordance line illustrating a context for a single word, and the annotations are limited to recording the relevant features of the context of that word.

²The term *profile* (used here as a verb) is borrowed from (Langacker, 1987), esp. pp. 183ff.

Scenario 1

- Scene 1. X_1 is at Place.
Scene 2. X_2 is not at Place.
Transition 1. X_1 moves away from Place.
Transition 2. X_2 moves to Place.
Scene 3. X_1 is not at Place.
Scene 4. X_2 is at Place.

Ordering Constraints

- Constraint: Scene 1 precedes Transition 1.
Constraint: Scene 2 precedes Transition 2.
Constraint: Transition 1 precedes Transition 2.
Constraint: Transition 1 precedes Scene 3.
Constraint: Transition 2 precedes Scene 4.

This Scenario 1 is common to (and is “inherited by”) two other scenarios which add certain elaborations. The temporal precedence relation holding between Transition 1 and Transition 2 is general enough to allow anything from immediate replacement of X_1 by X_2 , as in the game “King-of-the-Hill” (where X_2 pushes X_1 off the hill) to a temporal gap of considerable length, as when the role occupied by a retired or deceased company officer remains unoccupied for many years.

In terms of their separate historical association with the Place in Scenario 1, we can refer to the entities X_1 and X_2 as the Old Theme and the New Theme respectively, or, simply, Old and New. We refer to such names as OLD, NEW, and PLACE as **frame elements** or **FEs**.³

Now we can look at some of the vocabulary that fits the description of this frame.

Limiting ourselves to the scenario just sketched out (which contains no suggestion of agency or causation), we can see that the noun *replacement* has two uses: in one of its senses it can profile the whole event, as with the complex subject NP in the sentence [*Sylvia's replacement by Adam on the committee*] was a surprise to everyone; or it can refer to (i.e., “profile”) only the NEW entity in the final scene, as identified in the subject of the sentence [*Sylvia's replacement*] turned out to be completely unprepared for the job. The verb *replace* temporally profiles Transition 2, as in *Adam replaced Sylvia on the committee*. This means that if there is a temporal qualification added to this clause, it identifies the time when Adam showed up in the PLACE, and that does not necessarily include the time when Sylvia left it.

So far we've seen only a scene with the NEW and the OLD. The somewhat unnatural first sentence for illustrating our frame is intended to have only the interpretation that Adam came to have the position on the committee which Sylvia used to have. But the verb *replace* actually has two more possible profilings, fitting two slightly different elaborations of our initial scenario, both involving an agency that brings about one or both of these transitions, i.e. two types of **causative**.

In the first (and more restricted) of these elaborations, there is an entity X_3 that brings about Transition 2 but not Transition 1. Thus:

³The FEs are in many ways analogous to the **case roles** or **thematic roles** of various frameworks. For a small number of frames, the FEs could be labeled by familiar terms from such traditions, like **Agent**, **Theme**, **Instrument**, etc., but we do not require ourselves to accommodate our FE labels to such lists in cases where there is no obvious fit. The FE names are understood in terms of roles in specific frames and do not need to be thought of as selected from a restricted universal set. These labels are *reusable*, to be sure—which is why we seek to include as many words as possible within individual frames—but our work can proceed without deciding on an initial complete inventory a priori.

Scenario 2

- Scene 1. X₁ is at Place.
- Scene 2. X₂ is not at Place.
- Transition 1. X₁ moves away from Place.
- Transition 2. X₂ moves to Place.
- Cause: X₃ causes Transition 2.
- Scene 3. X₁ is not at Place.
- Scene 4. X₂ is at Place.

In an interpretation supported by this second scenario, the entity X₃ (which we will refer to as AGENT) can participate in a profiling of the "Cause" clause, as in *Lydia replaced Sylvia (with Adam) on the committee*. In this use of *replace* it is possible that the AGENT, Lydia, had nothing to do with Sylvia's move in Transition 1. That may have happened years ago. Given this second scenario, even though a NP mentioning the OLD is grammatically required, Sylvia might not be a participant in the profiled event at all. A temporal qualifier would only identify the time of Transition 2.

In Scenario 3, the second elaboration on the simple REPLACEMENT frame, we have a situation in which the AGENT is responsible for both of the transitions. Thus:

Scenario 3

- Scene 1. X₁ is at Place.
- Scene 2. X₂ is not at Place.
- Transition 1. X₁ moves away from Place.
- Transition 2. X₂ moves to Place.
- Cause: X₃ causes Transition 1 & Transition 2.
- Scene 3. X₁ is not at Place.
- Scene 4. X₂ is at Place.

On the interpretation of causative *replace* in this third scenario, we see Lydia as performing the two-part act of (1) removing Sylvia from the position and (2) placing Adam in it instead. As noted, the interpretation of temporal modifiers goes along with profiling: in Scenario 3, the sentence *Lydia replaced Sylvia with Adam last week* places both transitions within the same period.

Whether the existence of two interpretations corresponding to the two causative scenarios requires the positing of two senses for *replace* may be a problem for theories of polysemy, but there is no doubt of the difference in the implications. If you are told that *you can be replaced*, you have been warned that you might lose your job: you could become an unwilling participant in Transition 1. But if you are told that *you can never be replaced*, or equivalently that you are *irreplaceable*, then you will be pleased: they will never find anyone who can fill your shoes.

There are many other words that evoke one or more of these scenarios. The verbs *succeed* and *supersede* are for the most part limited to Scenario 1. The verb *supplant* has more or less the same possibilities, syntactic and semantic, as *replace* in all three scenarios. The verb *substitute* fits all three scenarios, but has a different syntax. In Scenario 1, where the verb is used intransitively, the NEW *substitutes* for the OLD, differing from *replace* in requiring the oblique rendering of the NEW with the preposition *for*. In the causative scenarios, the two verbs exhibit an alternation between direct and oblique objects, and a difference in prepositions for the oblique: the AGENT *substitutes* the NEW *for* the OLD, but the AGENT *replaces* the OLD *with* the NEW.

The noun *successor* profiles the NEW in respect to the OLD, whereas the noun *predecessor* profiles the OLD in respect to the NEW, at a time period at the end of the transitions. The phrasal preposition *instead of* also fits this frame: *You can use raw sugar instead of refined sugar* is a paraphrase of *You can substitute raw sugar for refined sugar*. The adjective *existing*

marks something as destined for replacement: if you see a campus map with one of the buildings identified as *the existing stadium*, you know that this structure is destined to be replaced. The time adjective *previous* belongs to this frame. If you tell me about your *former wife* or your *ex-wife*, I will know that she is someone who was once married to you; but if you speak of your *previous wife*, I will also know that she has in the meantime been replaced.

Numerous nouns are understood with reference to a specialization of the replacement frame: *crown prince*, *heir apparent*, *standby*, *understudy*, and so on. Thus, by defining the REPLACEMENT frame, we are able to concisely represent the semantics of at least twenty lexical items, including nouns, verbs, adjectives and a preposition.

3 The FrameNet Task

The job of FrameNet is to document from attested instances of contemporary English the manner in which frame elements (for given words in given meanings) are grammatically instantiated in English sentences and to organize and exhibit the results of such findings in a systematic way. We have seen, for example, that in causative uses of the words, an expression about *replacing NP with NP* takes the direct object as the OLD and the oblique object as the NEW, whereas *substituting NP for NP* does it the other way around. A commitment to basing such generalizations on attestations from a large corpus, however, has revealed that in both UK and US English, the word *substitute* also participates in the valence pattern found with *replace*, i.e. we find examples of *substituting* the OLD *with* the NEW.

More specifically, the actual work of the FrameNet lexicographers is to record the variety of combinatorial patterns found in the corpus for each word in the FrameNet lexicon, to present the results as the **valences** of the words, to create software capable of deriving as much other information about the words as possible, from the annotations, and to add only that information manually which cannot—or cannot easily—be derived automatically from the corpus or from the set of annotated examples.

3.1 The Corpus

The corpus used for the work in FrameNet-1 is the British National Corpus, more than 100,000,000 running words of contemporary British English.⁴ In the current phase we have added the North American newswire corpora from the Linguistic Data Consortium (<http://www ldc.upenn.edu>), and eventually we hope to be able to add the full resources of the American National Corpus (<http://www.cs.vassar.edu/~ide/anc/>).

The reason for using a corpus should be obvious. Anyone working with large masses of linguistic data learns quickly that native-speaker intuition is unreliable and that there is no principled way of resolving differences between native speakers. The observations about the alternative valences of causative *substitute* were not known to us in advance, but became obvious from the corpus. It is for these reasons that the central data of FrameNet are the collection of attested sentences.

3.2 Lexicographic Relevance

The part of FrameNet descriptions concerned with the combined semantic and syntactic valence of lexical items requires consideration of both frame structure and syntactic structure. Our objective is to record exactly how the language relates frame elements to those syntactic constituents that are syntactically dependent on the lexical items that evoke the frame.

Clear instances of lexicographically relevant positions are those accepted in narrow conceptions of *subcategorization frames*. If the target word is the verb *demonstrate* and the sentence is [*The*

⁴Our use of the BNC is by courtesy of Oxford University Press, through Timothy Benbow. The version of the corpus we use was tokenized at Oxford, lemmatized and POS-tagged at the Institut für Maschinelle Sprachverarbeitung at the University of Stuttgart. Information about the BNC can be found at <http://info.ox.ac.uk/bnc>.

instructor] demonstrated [*the process*] [*to the students*], the direct object of the verb (*the process*) is clearly relevant. Since the act of demonstration requires an audience, most analysts would include the *to*-phrase (*to the students*). In other words, a lexical description of the verb must say that the subject and these two syntactic positions function in identifying frame elements that are necessary parts of the concept of the frame evoked by it. For the deverbal noun *demonstration*, clearly relevant positions for annotation would be a preceding possessive determiner or following prepositional complements, as in the NP [*the instructor's*] *demonstration* [*of the process*] [*to the class*].

An example of a context providing information that could be relevant to completing information about the frame evoked by a particular lexical item but which is out of reach of lexically determined contexts is to be seen in the reference to *the students* and *the process* in the following sentence, *The students complained that they couldn't really understand the process without a demonstration*. It is clear from the meaning of the sentence as a whole that the audience of the demonstration is supposed to be *the students*, and that *the process* is what needs to be demonstrated, but such information does not come from the combinatorial requirements of the noun *demonstration* itself. The process of understanding the sentence as a whole needs to call on some means of recognizing such information in the surrounding discourse. But here the task of lexicography is more restricted: to characterize the combinatorial requirements and possibilities that are specifically associated with individual words rather than with the possibilities provided by the whole grammar of the language. On lexicographic relevance, see (Fillmore and Atkins, 1998).

Briefly, we regard the following constituents as worthy of annotation for lexicographic purposes.

- For verbs, nouns, adjectives and prepositions: their post-head complements, i.e., constituents of the phrase headed by the target (within the VP, NP, AP or PP) which amplify our understanding of the frame evoked by the head. (Complements which are "extracted" are similarly tagged.)
- For verbs, constituents external to the VP which instantiate an FE of the verb, either directly (as the verb's subject) or indirectly (by being a direct argument of a predicate which governs the VP through any of various "control" relations).
- For nouns, we annotate frame-relevant possessive determiners (*[Roger's] decision to join the party*), relational adjectives (*[environmental] protection*) and modifier nouns in compounds (*[environment] protection*).
- For nouns which occur with support verbs, we annotate constituents capable of being seen either as complements of the noun or as complements of the verb. That is, unlike a parser, we don't have to decide the correct constituent structure of a sentence like *She registered a complaint to the committee about her treatment by your colleagues*. If *complaint* is the target word, and *register* is the support verb, it does not matter for our purposes whether *to the committee* and *about her treatment by your colleagues* belong to a nominal phrase headed by *complaint* or to a verbal phrase headed by *registered*.

4 Performance of FrameNet-1

The pilot on which the current project is founded was officially titled "Tools for lexicon building" and was funded under NSF grant No. IRI 9618838. Since that project, which we refer to as FrameNet-1, accomplished both more and less than it set out to accomplish, something needs to be said about its history.

FrameNet-1 built on proposals for a frame-based lexicon first formulated in joint work by Fillmore and lexicographer Sue Atkins, of which the first appearance in print was a series of studies of the English verb *risk* (Fillmore and Atkins, 1992; Fillmore and Atkins, 1994). The original task was defined as associating a word with a frame, choosing a list of frame elements for the frame, looking at sentences which showed a use of the word in the appropriate sense, and selecting and labeling those constituents in those sentences which (within the constraints of lexicographic relevance) instantiated concepts from the frame. The constituents identified as frame elements were then to be classified (automatically if possible) as to their phrase type (PP, etc.) and in respect to their grammatical function (Object, etc.). A lexical entry was to be constructed automatically from the set of annotations selected for illustrating its use in a given sense, involving only minimal manual manipulation, such as the provision of a standard dictionary definition. The product of FrameNet-1 was to be full descriptions of the frame-semantic and syntactic combinatorial properties of the words in the lexicon (Lowe et al., 1997; Baker et al., 1998).

4.1 Software

The FrameNet-1 project used a combination of publicly-available and in-house software. The British National Corpus (BNC) was exploited using the Corpus Workbench (CWB), a suite of tools developed at Stuttgart by Oliver Christ and Bruno Schulze and made available to us through Ulrich Heid of IMS. (Information about IMS can be found at <http://www.ims.un-stuttgart.de>.) The exploratory searching of the BNC was carried out using *xkwic*, an interactive keyword-in-context program, which allows easy searching and sorting on user-defined patterns. In-house software allowed users to set parameters (via a web form) for the production of subcorpora for annotation, this information was then passed to the batch-processing CWB tool, called *cqp* ("corpus query processor") which did the actual searching, using a set of complex pattern searches over part-of-speech strings.

The annotation was done with the Alembic Workbench software, a product of the Mitre Corporation; information is available at <http://www.mitre.org/technology/alembic-workbench>. In-house software allowed annotators to create Alembic tag-sets from frame definitions, find the correct files and corresponding tag-sets, and copy the annotated files back into the correct directory locations.

In-house programs (mostly Perl CGI scripts) were the "glue" holding these applications together and enabled researchers to perform much of their work from anywhere, using web forms. This included the maintenance of databases defining frames and frame elements, which were stored in "flat" ASCII files with tab-separated fields; the considerably more complex database structures used in FrameNet-2 will be discussed in Section 7.

4.2 Support Verbs

FrameNet-1 was mainly limited to verbs, adjectives and deverbal nouns. In the work with adjectives and nouns we had to find a justified way to include XP-external arguments, and that required us to recognize a class of **support verbs**, verbs whose main function seemed to be that of associating their subjects with the frame evoked by their dependent noun or adjective. We found reasons for tagging the support verbs themselves, since in doing that we were recording information about frames that could be useful in its own right. For example,

- different senses of a noun can be selected by different support verbs (e.g., *have an argument* - 'quarreling'; *make the argument* - 'reasoning');
- different participants in an eventuality can be profiled with different support verbs (e.g., *perform an operation* vs. *undergo an operation*);

- different phases of a complex event can be signaled with different support verbs (e.g., *make a promise* vs. *keep a promise*); and
- different elaborations of what might seem a single frame can be signaled by different support verbs (e.g., *make a complaint* vs. *register a complaint* (the latter expressing some sort of official act)).

What started out as a simple decision to identify those “semantically bleached” verbs whose syntactic subjects counted as elements of the frame structure of their syntactic complements is now evolving into something approaching the “lexical functions” of Igor’ Mel’čuk (Mel’čuk, 1996). However, in FrameNet-1 we did nothing but record the support verb itself and the frame element of the argument it supported.

4.3 Implicit FEs

In order to be able to have an automatic way of grouping lexical units in the same frame according to the constellations of frame elements which conceptually accompanied them, we needed to recognize FEs that were conceptually present but not expressed in the sentence. This was made difficult because the annotation tool we were using, the Alembic Workbench, only permitted an operation of wrapping SGML tags around pieces of actual text. We resolved this problem by adding to the sentences ready for annotation symbols that the annotators could label. Even if we wanted to, there was no way of adding “empty categories” in syntactically appropriate contexts (the Corpus is not parsed), so these symbols were added immediately after the target word. In the process of making this possible, we felt that it would be useful to distinguish the various kinds of reasons why arguments could be omitted. Speculating that when there were two or more omissions in the same sentence we would never find two instances of the **same** omissibility condition, we provided just one symbol for each type of zero element.

The three types were:

- **Constructional**, where the absence of a constituent representing a particular FE was authorized by the grammar of the language: the missing subject of an imperative sentence (*Say something!*), the missing subject of an infinitive phrase (*To eat that fish might be dangerous*), the missing agent of a passive sentence (*The document had already been submitted*), the agent in a middle construction (*This product wipes clean with a soft cloth*), and so on.
- **Existential**, where the missing element could be given a generic or existential interpretation, such as when the objects of certain common verbs are not mentioned: *sew*, *eat*, *bake*, etc.
- **Anaphoric**—the most interesting cases—in which the missing element has to be “understood” or “given” in the discourse context. This covers two situations, one grammatical, one lexical: (1) where the genre determines the omissibility, as in the case of “instructional imperatives” (*Shake before using*) as in recipes and assembly manuals (Atkins et al., 1988) and (2) where the determination is lexically licensed (Fillmore, 1986; Condoravdi and Gawron, 1996).

Examples of frame-related words exhibiting the difference between the existential and the anaphoric can be provided with our verbs *replace* and *substitute*. The missing element in *We had to replace Harry* is of the existential type: the context doesn’t have to contain information about who Harry was replaced with. The missing element in *We had to substitute honey* is of the anaphoric type: in this case the discourse context has to have mentioned something, perhaps sugar, that was otherwise called for in some recipe.

In the end we also added a fourth type, **Incorporated**, to cover cases in which the meaning of the word itself incorporated the FE represented by the label. This could be used, for example, for the unmentioned target for the verb *spank*.

4.4 Lexical Entries

In principle, FrameNet lexical entries are sets of lexical sub-entries, each of these being a record of what FrameNet has recorded for the lemma in one of its senses. However, since our work proceeded one frame at a time, rather than one lemma at a time, we did not conduct a systematic exploration of polysemy for its own sake. The only cases of polysemy as such that were treated in FrameNet-1 were those in which the same lemma appeared in more than one of the frames we examined.⁵

An individual lexical entry covers a lemma in a particular part of speech, e.g., as verb or as noun. A lexical sub-entry is intended to represent a single lexical unit, i.e., a lemma in a given part of speech in a single sense.

A lexical sub-entry comprises the following components:

1. **Headword:** The lexical unit to be defined
2. **Frame:** Identification of the individual background frame, e.g., "Communication/Argument" (COMMUNICATION domain, ARGUMENT frame)
3. A definition, if relevant, of the meaning associated with the lexical unit taken (with permission) from the Concise Oxford Dictionary, 10th Edition
4. **Table of Frame Element Realizations:** a list, for each FE, of the syntactic ways in which Frame Elements have been expressed in the annotated sentences in terms of grammatical function and phrase type
5. **Table of Valence Patterns:** a list of the sets of Frame Elements with their syntactic realizations as found in the annotated sentences
6. **Annotated sentences** (where each sentence is annotated in respect to a single target word and the semantic roles which neighboring phrases bear in relation to that word)

The Frame Element Realization table and the Valence Pattern table are derived automatically from the annotated sentences.

In short, a FrameNet entry provides information, for each sense, about frame membership and the syntactic means by which each Frame Element is realized in the word's surroundings, and summarizes, as Valence Patterns, the full range of combinatorial possibilities as attested in the Corpus.

5 The MySQL Viewer

A MySQL interface which permits users to browse the FrameNet-1 database was created by Hiroaki Sato of Senshu University. It allows efficient collection of information of many kinds, based not only on the annotations themselves but also on minimal syntactic parsing of the annotated phrases. Here are some examples of the kinds of information that can be retrieved from the data.

FrameSQL * Domain*: communication_ Frame*: gesture_ FE set*: Add*, Com*, Manr*, Mns*, Msg*, Other*	Lemma* beckon.v gesticulate.v gesture.v motion.v nod.v signal.v wave.v	FE* Add Com Manr Mns Msg Other	lemmaPOS* <input type="text"/>	headN of XP* <input type="text"/>	cType* <input type="text"/>
	Phrase Type* <input type="text"/>	sort&disp* PT <input type="text"/>	type* <input type="text"/>	GF* <input type="text"/>	verb infl.* <input type="text"/>
			headX of XP* <input type="text"/>	limit* 9999 <input type="text"/>	<input type="button" value="Reset"/>

Figure 1: Searching for *motion* in the GESTURE Frame

5.1 FE realization

Figure 1 shows a web form that appears on selecting the GESTURE frame. The highlighted entities in the first and second boxes are traces of the user's selection of the verb *motion* and the frame element MESSAGE(MSG). The sort and display menu ("sort&disp") shows the selection of Phrase Type (PT). Clicking the "Search" button in this state will find those sentences in the database in which the target is the verb *motion*, as analyzed in the GESTURE frame, and which contain an annotation for the MESSAGE (what gets communicated with the gesture), and will group these according to the syntactic type of the phrases instantiating this information.

No.	domains	frames	lemma	pos	goto	Definition from Concise Oxford Dictionary
1	communication*	gesture*	motion	v	entry	1a: direct (someone) with a gesture.

FE set: Addressee Communicator Manner Message

valenceFreq	patterns
01	AVP
03	DNI
01	NP
05	PP
01	Sfin
06	Sforto
12	VPto
29	

Figure 2: Phrase Types for MESSAGES for GESTURE.motion

Figure 2 shows the top of the window created in response to that search. From this table we learn that the annotation set contains 12 sentences in which the MESSAGE is expressed with an infinitive phrase (VPto).

Clicking on the number 12 to the left of VPto produces the window shown in Figure 3, which gives all of the annotated sentences in which the MESSAGE communicated by the gesture is expressed as an infinitival VP. (The FEs are differentiated by color.)

5.2 More Complex Searches

The next example shows a different kind of search. The web form shown in Figure 4 is for the frame we call STATEMENT. (One person speaks to another; in contrast with CONVERSATION, which is reciprocal, and with other monological but interactional speech frames such as those

⁵There are, however, cases in which annotators have noted multiple senses by marking the sentences exhibiting them with sentence-initial sense numbers. Unfortunately, no mechanism was developed in FrameNet-1 which would allow such individually marked senses to be associated with the correct frames, definitions and valence descriptions.

- [\[back to the main table\]](#)
1. [edic] **He**^(NP) turned and motioned **Carter**^(NP) **to take the lead**^(VPto) . [1333222] [Top]
 2. [edic] **She**^(NP) subsided gracefully on to a mat and motioned **Frika**^(NP) **to join her**^(VPto) . [1321469]
 3. [edic] **She**^(NP) motioned **him**^(NP) **to stay**^(VPto) . [1321229]
 4. [edic] **Mother Bombie**^(NP) motioned **him**^(NP) **to keep his voice down**^(VPto) but he ignored her . [1321224]
 5. [edic] **She**^(NP) seated herself by the fireplace , and motioned **the two men**^(NP) **to sit**^(VPto) . [1321224]
 6. [edic] **He**^(NP) motioned **Parker**^(NP) **to sit down**^(VPto) . [1321224]
 7. [edic] We followed the tweed skirt into a kitchen where **its owner**^(NP) motioned **us**^(NP) **to sit**^(VPto) , **in the manner of someone training a dog**^(PP) . [1321224]
 8. [edic] **With a practised hand**^(PP) **the senior partner**^(NP) motioned **a waiter**^(NP) **to bring a fresh pot of coffee and five more pipes of tobacco**^(VPto) . [1321224]
 9. [edic] **He**^(NP) motioned **the young officer**^(NP) **to sit down**^(VPto) . [1321224]
 10. [edic] **He**^(NP) motioned **to the barman**^(PP) **to refill their glasses**^(VPto) , a feeling of satisfaction running through him . [1321224]
 11. [edic] **His wife**^(NP) motioned **to him**^(PP) **to be quiet**^(VPto) as she moved around the counter . [1321224]
 12. [edic] **Nicholson**^(NP) motioned **to them**^(PP) **to take a seat**^(VPto) . [1321224]

Figure 3: Examples of MESSAGE with Phrase Type VPto

FrameSQL * Domain: communication* Frame: statement* FE set: Add*, Degr*, ICause*, Manr*, Medium*, Mns*, Msg*, Null*, Other*, Spkr*, Top*	Lemma* add.v address.n address.v admission.n admit.v advise.v affirm.v affirmation.n allegation.n allege.v	FE* Add Degr ICause Manr Medium Mns Msg Null Other Spkr Top	lemmaPOS* v Phrase Type* NP GF* Obj headX of XP* 1	headN of XP* 1 sort&disp* lemma verb infl.* limit* 5000	cType* type* Search! Reset
	<input type="text"/>				

Figure 4: Searching for verbs with Addressee-Direct Objects in the STATEMENT Frame

involving asking, answering, accusing, etc.) This time no particular word is chosen. The highlighted element of the FE column is ADDRESSEE (ADD); the search is limited to verbs (note the v in "lemmaPOS"). The phrase type NP and the grammatical function Obj were chosen to select the grammatical way in which the ADDRESSEE role is to be expressed. The selection "lemma" in the sort&disp menu means that the display should be ordered alphabetically by the verbs. The search will produce a list of verbs in the STATEMENT frame that can express the ADDRESSEE as a NP-Object. (Other contexts for STATEMENT verbs have ADDRESSEE expressed obliquely, as the object of the preposition *to*.) Figure 5 shows the results.

valenceFreq	domains	frames	lemmas
44	communication	statement	address
31	communication	statement	advise
17	communication	statement	assure
19	communication	statement	caution
39	communication	statement	inform
12	communication	statement	lecture
16	communication	statement	notify
14	communication	statement	promise
59	communication	statement	tell
02	communication	statement	write
253			

Figure 5: Results of the previous search

One of the verbs in this collection is *lecture*. This verb, it turns out, appears in both this list and in the list of verbs that identify the ADDRESSEE with a prepositional phrase, i.e., with *to*. We can examine these sentences by clicking on the word *lecture*, producing the display shown in Fig. 6, which reveals that the annotator noted a difference between these uses and the others by attaching the label "sense2" to these sentences. In these uses the verb *lecture* has a meaning closer to 'reproach' than to 'inform'; there is no separate entry for the word in this sense (see footnote 5).

- [edict] [sense2] Robert Saxton^(Sbj), in these pages, recently denied that there was any longer a problem with contemporary music, and instead lectured the audience^(Add) [redacted]^(Top) [redacted]^(Top)
- [edict] [sense2] Sitting by her bedroom window later that night, Laura^(Sbj) lectured herself^(Add) for having been so naive as to have been upset by the remarks he had originally made on the dance-floor^(Obj) [redacted]^(Top)
- [edict] [sense2] Now the Tory party^(Sbj) are forever lecturing me^(Add) [redacted]^(Top) are they not [redacted]^(Top)
- [edict] [sense2] Besides, [redacted]^(Sbj) 're in no position to lecture me^(Add) [redacted]^(Top)." [redacted]^(Top)
- [edict] [sense2] No matter how long the language master^(Sbj) lectured him^(Add) [redacted]^(Top), the words seemed hollow in the light of his first remarks [redacted]^(Top)
- [edict] [sense2] " Start lecturing [redacted]^(Sbj) me^(Add) [redacted]^(Top) [redacted]^(Top)
- [edict] [sense2] We are doing more for training than any previous Government and the Labour party^(Sbj) is in no position to lecture us^(Add) [redacted]^(Top) [redacted]^(Top)
- [edict] [sense2] Experience convinced her that the man who was drunk^(Sbj) would lecture him^(Add) [redacted]^(Top), a theme prompted by the first mention of hospices [redacted]^(Top)
- [edict] [sense2] No longer did he^(Sbj) lecture people^(Add) [redacted]^(Top) [redacted]^(Top)
- [edict] [sense2] Walsh^(Sbj), lecturing the others^(Add) [redacted]^(Top) [redacted]^(Top) had the most important job of the moment as fund-raising co-ordinator [redacted]^(Top)
- [edict] [sense2] In a bang-up-to-date textbook (see Bicknell 1988), she lectures future generations of health and social service professionals^(Add) [redacted]^(Top) [redacted]^(Top)
- [edict] [sense2] He^(Sbj) always used to lecture people^(Add), tell them about the honesty of the artist, the necessity not to speak like a bourgeois [redacted]^(Top)

Figure 6: *Lecture Sense 2*

6 Goals for FrameNet-2

In FrameNet-2, one of the main differences will be a much larger vocabulary, the increase in great part made possible by the addition of large numbers of nouns that serve as the noun heads of FE-instantiating NPs. There will be two main sorts of sub-entries: lexical units will be described both in their roles as **governors** and in their roles as (the heads of) **governed** constituents. In general, verbs will be described only as governors, i.e., as the carriers of frame structure, and their analyses will be expressed in terms of the frame elements generated by their frame and the manner of the syntactic instantiation of these. Nouns will in general be annotated as governors, by analyzing their frame elements (as post-head complements, possessive modifiers, or modifying nouns or adjectives in noun compounds); but many will also be described in respect to their use as dependent elements, by finding definitionally relevant governors and by marking the phrase within which they serve as slot fillers of the frames they evoke.

For example, in the case of nouns that are the names of tools and other artifacts, in sentences which illustrate canonical uses of these artifacts, we will record the governing word (usually a verb) which evokes the frame which welcomes the noun and identify this frame. Then we will find the constituent containing the target noun which marks the frame element in which the noun is serving, and identify that frame element. Thus, given the target *knife*, for a sentence like *I tried to slice the ham with a dull knife* we would mark *slice* as the governor, indicate CUTTING as the name of the frame, select the phrase *with a dull knife* as the phrase containing the target which identifies a frame element of *slice* in that sentence, and give its name as INSTRUMENT. In another example for the same noun, *He was stabbed three times with a knife*, we would tag *stab* as the evoker of an ATTACK frame and select *with a knife* and tag it as WEAPON. Thus, for certain classes of dependent nouns, the entries would provide information in such a form as

Dependent	Governor	Frame	Frame Element Name	Phrase
<i>knife</i>	<i>slice</i>	CUTTING	INSTRUMENT	<i>with a dull knife</i>
<i>knife</i>	<i>stab</i>	ATTACK	WEAPON	<i>with a knife</i>

A very large number of nouns will be treated in respect to their function as FE-slot fillers, including names of tools, dwellings, geographic locations, diseases, kinship relations, body parts, species and varieties of plants and animals, etc.

To improve the performance of software written to discover typical nouns filling FE slots (see Section 8), we have also committed ourselves to the tagging of selectionally transparent nouns, e.g., nouns designating types, parts, portions, aggregates, borders, classifiers, quantifiers, etc., which are “transparent” to the selectional preferences of their governors (e.g. *It’s hard to cut paper with that kind of knife*). We also hope to expand the typology of what we have been calling “support verbs” to the full set of lexical functions of the Mel’čukian kind (possibly by incorporating into the FrameNet database work such as (Fontenelle, 1997)). We also plan to provide information regarding the fit of particular lexical items in selected grammatical constructions, such as appearance in various expressions of time (dates and other calendric phrases), place (addresses), titles, institution names (governmental, corporate, religious, etc.).⁶

In addition to information that is directly shown in the annotations, there will be information of various sorts derivable from them. Thus, it will be possible to derive information about functions of prepositions (the FEs they mark in given frames), support verbs (the kinds of nouns they serve and the FEs their subjects instantiate within the frames of those nouns), etc. Further information about nouns will be provided through links with taxonomies provided by other services, such as WordNet.

⁶To facilitate such collections, we will use the FSAs of the SRI FASTUS system, through a subcontract with SRI International. More information can be found at <http://www.ai.sri.com/natural-language/projects/fastus.html>.

6.1 Multi-Word Units

FrameNet-1, because of some difficulties in the subcorporation and annotation software, was not capable of dealing with multi-word expressions. FrameNet-2 will cover compounds, phrasal verbs, and idioms. (Idioms will be included only as they are encountered, in connection with words we are studying, rather than as a result of searches dedicated to finding them, but at least we now have the means of representing them.) Among the multi-word expressions we need to recognize are phrasal verbs of the following types:

verb plus preposition	<i>think about</i>
verb plus particle	<i>take off</i>
verb plus particle plus preposition	<i>put up with</i>
verb plus NP plus preposition	<i>prevent NP from</i>
verb plus NP plus particle	<i>take NP away</i>

The FrameNet-2 annotation software allows us to recognize multiply represented constituents, including phrasal and compound target words, by annotating on multiple "layers". This allows us to correctly annotate the occasional sentences in which constituent structure and frame element boundaries do not coincide. (Thus we should be able to take advantage of some of the insights, regarding the lexicon/grammar interface, of autolexical syntax (Sadock, 1991).)

6.2 Profiling

Profiling, it has been noted, is the representation of the foregrounded part of a frame, the participant, prop, phase or moment which figures centrally in the semantic interpretation of the sentence within which the frame is evoked. Some examples were given in sections 2 and 4.2.

6.3 Inheritance

Inheritance is a relation between two frames such that one of them has all of the properties of the other, plus something else. Perhaps the simplest kind of inheritance-plus-elaboration can be seen between abstract frames and frames with specific kinds of profiling.

For example, there is a general MOTION frame involving little more than location changes whose components are a starting point (SOURCE), a trajectory (PATH), and a destination (GOAL). Verbs like *move* or *travel* are instances of this general frame. There are elaborations in which the SOURCE is profiled (and hence the moment of departure), as with *depart*, *exit*, *leave*, *vacate*, etc.; there are some in which the GOAL is profiled (and hence the moment of arrival), as with *arrive*, *enter*, *reach*, *attain*, etc.; and some in which the PATH is profiled, as with *pass*, *cross*, *traverse*, etc. (Many of these verbs have their own further elaborations, involving figural properties of the landmark with respect to which the motion is conceived, requiring further levels of inheritance; e.g. *enter* elaborates *arrive*, adding the requirement that the GOAL be a CONTAINER.) Caused-motion frames can be similarly organized, with *insert* being GOAL-profiled, *remove* being SOURCE-profiled, etc.

6.4 Multiple Inheritance and Blending

Multiple inheritance is used to show the simultaneous operation of two or more frames which do not have inheritance links between them, where mappings are specified between their FEs. For example, there is a large class of English verbs involving someone's passing judgment on the character or behavior of someone else, which we call the JUDGMENT frame. The evaluation can be positive (as in *admire*, *cherish*, *appreciate*) or negative (as in *blame*, *disapprove of*, *disesteem*, etc.). FEs for this frame will include the JUDGE, the EVALUEE, and the BEHAVIOR being judged, as in [*She*] scorned [*him*] [*for saying that*].

Some verbs add to this abstract characterization the idea that the person making a judgment is also communicating such judgment; i.e., we see a **blend** of the frames JUDGMENT and COMMUNICATION, in which the JUDGE in the one frame is identified with (or “bound to”) the SPEAKER in the other frame. This is the case for verbs like *praise* and *acclaim* for positive judgments, and *criticize* and *denounce* for negative judgments, e.g. *She criticized him for saying that*. But there is a third class of verbs, in which the JUDGE is equated with the SPEAKER and the EVALUEE is also equated with the ADDRESSEE of the COMMUNICATION. This is what we find in *flatter* and *compliment*, and in *scold* and *upbraid* (*She scolded him for saying that*).

Multiple inheritance is also found in verbs that incorporate both the idea of CONVERSATION and the idea of FIGHTING. Simple CONVERSATION verbs include *converse*, *chat*, etc., and simple FIGHTING verbs include *fight*, *battle*, *clash*, etc. Verbs that include both of the ideas (i.e., that designate fighting with words) include *quarrel* and *argue*. Syntactic traces of this blend include the fact that the preposition *over*, which marks the “issue” in a dispute, occurs with the talk-fight verbs but not with the neutral conversation verbs. (Thus, we find *We fought over the inheritance* or *We quarreled over the inheritance* but not **We conversed over the inheritance* or **I spoke with my brother over the inheritance*(Johnson, 1999).)

6.5 Frame Composition

By frame composition is meant the situation in which a complex frame is made up of parts which are also frames, as with the scenarios mentioned earlier in connection with the REPLACEMENT frame. This is true of a large family of EXCHANGE frames, in which ownership relations in one stage are reversed in a transition of transferring ownership. Verbs incorporating caused events also have frame embeddings and are thus further instances of frame composition.

7 The FrameNet-2 Database

In this section we show some of the ways in which information created through FrameNet research is to be stored in a MySQL database. We will do this by briefly examining some of the tables linked in the database, leading up to a diagram which shows the entire database.

Many of the tables have ID numbers, prefixed by “PK” (primary key) in a hexagon; these are the unique numbers by which information is cross-indexed in other tables. Such ID numbers are repeated in the cross-referring tables, prefixed by “FK” (foreign key). As can be seen in Fig. 7, every lexical unit is linked to an appropriate MultiWord (explanation below) and also to an appropriate Frame. Those tables representing information entered into the database by a FrameNet researcher have, for record-keeping purposes, information about the date of their creation and an identifier of the researcher.

7.1 Lexical Units

The concept of **lexical unit** is central to the FrameNet lexicon. It is the entity that needs a definition and whose combinatorial properties we wish to understand. In the simplest case, a lexical unit is defined as a triple consisting of a lemma, a part of speech, and a frame: a lemma in turn is a set of word form alternatives, and is named by one of its variants (the so-called “citation form”). Thus the lemma *ring* as a noun has word-forms *ring* and *rings* (singular and plural); the lemma *ring* as a verb has word-forms *ring*, *rings*, *ringing*, *rang*, *rung*; the noun lemma *ring* when paired with a SOCIAL COMMUNITY frame (as in *ring of smugglers*) is one lexical unit; when paired with a JEWELRY frame (as in *wedding ring*) is a different lexical unit.

Fig. 7 is to show how lexical units are connected in the FrameNet-2 database to their formal properties (lemma, word-forms, and part of speech) and to their frames. The information that we want to connect with each lexical unit includes: its definition (sense description), the lemma of which it is an instance, its part of speech, the word forms that are tokens of the lemma, and

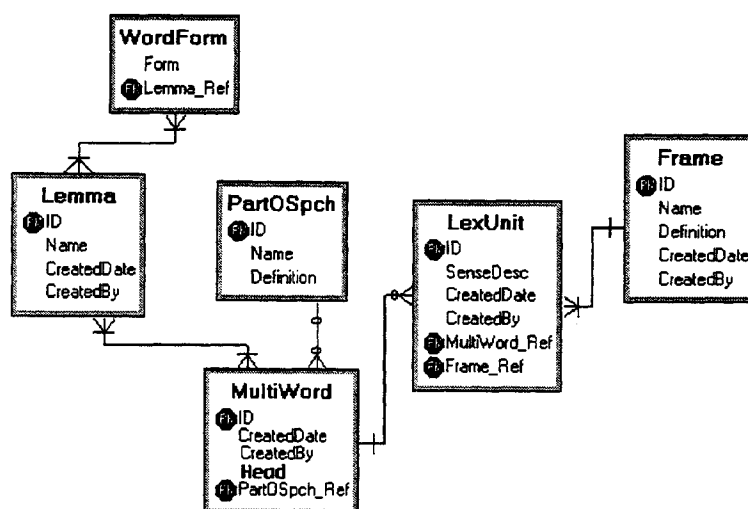


Figure 7: Tables associated with Lexical Units

the frame which it evokes. The figure shows an intervening table with the misleading name “MultiWord” which includes both simple lemmas and lemmas that are made up of two or more lemmas.

The definition (sense description) of a lexical unit is stored in the LexUnit table. This table also stores information about the frame to which the lexical unit is linked.

One of the features of the FrameNet-2 database is the ability to represent the relationships between multi-word lemmas such as *take off* and closely related single-word lemmas such as *take*. For example, the lemma *take* has word forms *take*, *took* and *takes* (etc.), and the multi-word lemma *take off* has the forms *take off*, *took off* and *takes off*. If we were to store multi-word lemmas as separate entries in the database, information about this relationship would be lost. So, to represent this relationship between multi-word lemmas and their single-word relatives, each lemma, whether single-word or multi-word, is linked to an entry in the MultiWord table. The relationship between the Lemma table and the MultiWord table allows an entry in the MultiWord table to be linked to one or more entries in the Lemma table, one of which will be labeled the head of the multi-word. It is the head which bears the part of speech of the whole and whose word-forms vary in the word-forms of the multi-word unit, e.g., the *take* of *take off*, the *house* of *bird house*, and the *attorney* of *attorney general*.⁷

7.2 Frames, Frame Elements, and Scenes

Fig. 8 shows how information about scenes is stored in the FrameNet-2 database. Scenes are subordinate states-of-affairs or events that make up a frame, or, put differently, are frames that are parts of other frames.

The structures shown here can be used to represent complex scenarios like those discussed in Section 2. Information about scenes includes: the name of the scene, a description, a flag to indicate whether or not a particular scene is profiled in respect to the frame in which it occurs and a reference to the frame to which the scene belongs. The ordering constraints on pairs of scenes within a frame are stored as entries in the SceneOrder table.

Another example of such a complex scenario is the COMMERCIAL TRANSACTION frame, where the transitions represent changes of ownership of GOODS and MONEY between a BUYER and a

⁷The relationship between multi-lemma words and their components has a number of complexities not touched on here.

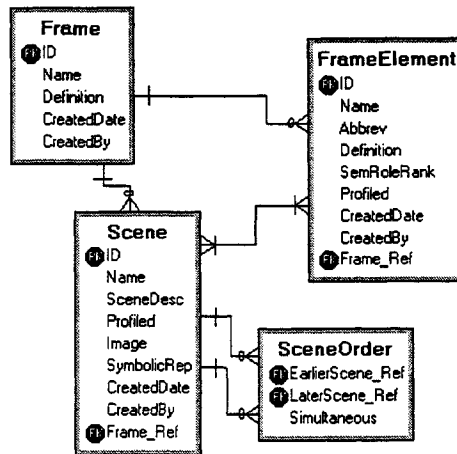


Figure 8: Tables linking Scenes, Frames, and Frame Elements

SELLER. There is an ordering relationship among scenes, but the transitions may occur simultaneously (especially if the BUYER and SELLER don't trust each other), or they may be separated by a large span of time.

The FrameElement table contains information for each frame element in the database. This includes: the name of the FE, an abbreviation for the FE, the definition of the FE, the rank of this FE with respect to other FEs in the same frame, and a field indicating whether or not this FE is profiled in this frame. The rank of an FE will determine the order in which the members of the valence sets will be displayed, but it will also be predictive of the linking to grammatical functions. (For verbs, for examples, rank #1 will be uniformly associated with the FE which, in active sentences, is realized as the sentence's subject.)

7.3 Frames and Inheritance

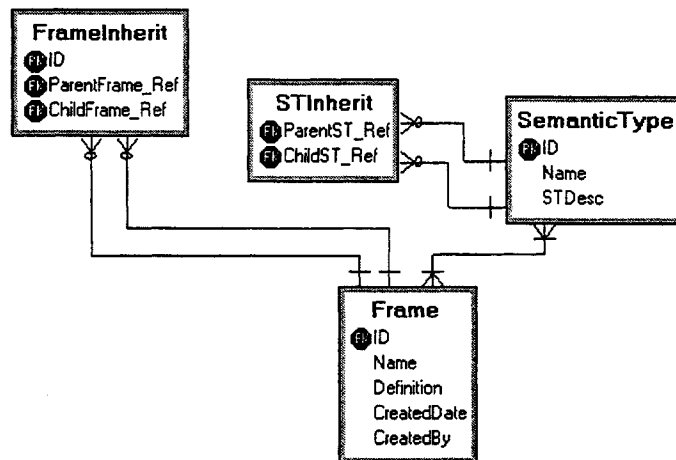


Figure 9: Tables representing Frame Inheritance

Fig. 9 shows how frame inheritance and semantic type information are stored in the FrameNet-2 database. The frame inheritance information is stored in the FrameInherit table, which contains one record for each frame inheritance relationship. Each entry in the Frame table may serve as either a "parent" or a "child" in an inheritance relationship. This table stores only immediate inheritance relationships, i.e., each entry in the Frame table only has direct access

to its immediate parent or child. However, a complete inheritance tree can be computed by iterative lookups using the FrameInherit table.

Relations between the meanings of lexical units of the sort where one incorporates the semantic properties of the other—but adds some detail—are introduced indirectly, through inheritance relations between their respective frames. This would apply to lexical relations between, say, *stare* and *gawk*, where the latter has an ‘open mouth’ component; between *say* and *whisper*, where the latter has a manner and loudness component; and so on.

7.4 Layers

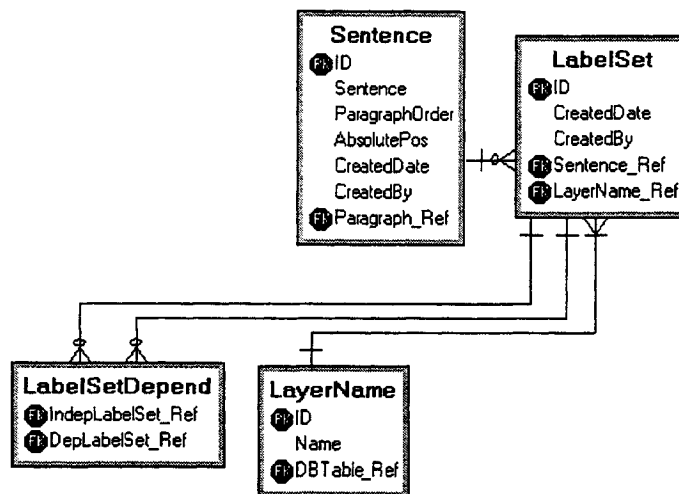


Figure 10: Representing Annotations with “Layers” of Labels

Fig. 10 shows that the annotation of a sentence consists of zero or more label sets, which can be thought of as “layers” of annotation. In the simplest case, we would have a label set for the FEs, another for the phrase types (PTs), and third for the grammatical functions (GFs), and the boundaries of the FEs would coincide with the boundaries of the syntactic constituents, each of which would be labeled with a PT and a GF. The entire sentence would also have a label in the Frame label set, giving the name of the frame, and the target itself would have a layer of its own. We would thus have a minimum of 5 label sets for annotated sentences. The FrameNet-2 software will enable annotators to view and edit any combination of these label sets.

We also need the ability to represent a dependency relationship between label sets. For example, in annotating a sentence in the COMMERCIAL TRANSACTION frame, the frame label set would contain only one label—“Commercial Transaction”; the FE label set will contain the labels SELLER, BUYER, etc. We want to be able to represent the fact that the FE label set is dependent on the frame label set, by storing the information in the LabelSetDepend table.

The number of label sets (layers of annotation) will be higher in cases of blended frames. E.g., the collection of verbs that evoke the COMMERCIAL TRANSACTION frame are all instances of other blended frames as well: *buy* is an instance of an ACQUISITION frame, and *sell* is an instance of a GIVING frame. Constituents which represent FEs in more than one frame would bear labels from correspondingly more label sets. (This sort of information is presented formally in terms of **multiple inheritance**, discussed in Section 6.4.) In such cases, the annotation software will allow users to view and annotate FEs in any frames that might be needed. It also will allow automatic presentation of the relations between the layers, such as the mapping of FEs between frames in cases of multiple inheritance, the mapping of FEs to phrase types or grammatical functions, and so on.

7.5 Labels

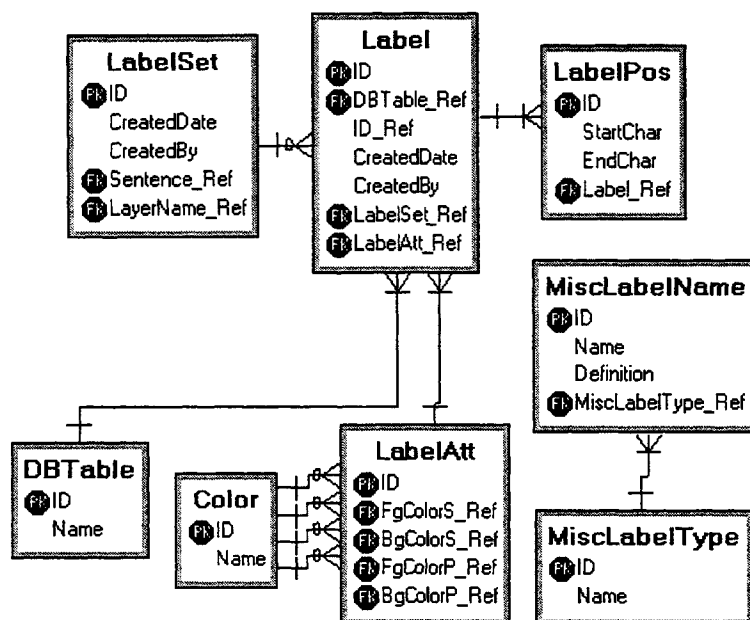


Figure 11: Labels used in Annotation

Fig. 11 shows how the database stores information about the labels used in FrameNet-2 annotation. We have chosen to store the annotation information inside the database, rather than as separate files. This allows us to link the names of frames, frame elements, and other items in the database directly to the annotation tags. Thus if we change the name of a particular FE, then all annotations that use that FE will automatically be updated to show the revised name.

The primary table used for annotation is the Label table. This table holds the information needed for a single label in an annotated sentence. The primary pieces of information that we need to store for a label are: the label name, the label position, the label attributes, and a pointer to the set of labels to which this label belongs. Here we will discuss only label position, since that is the feature which allows an improved treatment of compound lexical units and discontinuous FEs.

In order to place a label in the correct position in a sentence, we need to decide its starting and ending points, and store these in the LabelPos table. The units we will use for defining label positions will be characters rather than words. The one-to-many relationship between the Label table and the LabelPos table means that a label may have one or more label positions. This allows us to create a single label that is discontinuous. For example, in the sentence *They put me up for a month*, if the target is *put (NP) up*, we would want a single label "target" to refer to characters 6-8 and 13-14 of the sentence (covering the words *put* and *up* respectively); rather than annotating this with two separate labels, we can use a single label with multiple label positions. (The two words would, of course, have different labels on the part-of-speech layer.)

Because the FrameNet-2 database delimits labels in terms of characters, we can overcome some difficulties inherent in the "word-based" approach of the FrameNet-1 system. For example, in our treatment of compounds in FrameNet-1⁸ we were limited to units separated by spaces in the written text. As is well known, there are many inconsistencies in the ways in which compound

⁸In FrameNet-1, there were actually no entries that were themselves compounds. Rather we examined the

words are written in English, and character-based positioning allows us to deal equally well with such varieties as *teaspoon*, *tea-spoon* and *tea spoon*. Using character positions also has the potential to enable us to properly annotate text in languages such as German and Finnish, with extensive affixation and compounding, and in languages such as Chinese and Japanese, written without spaces between words.

7.6 Corpus

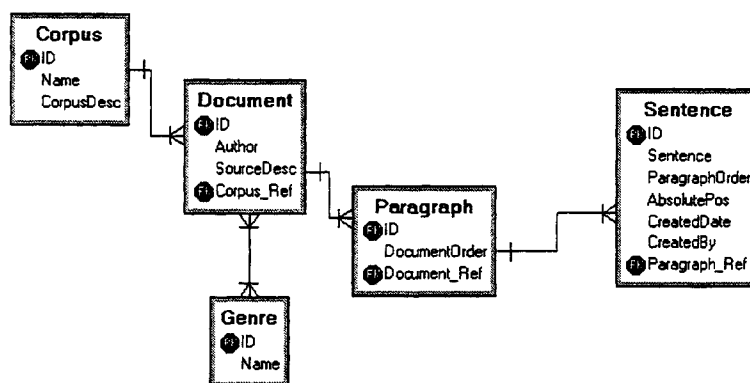


Figure 12: Situating Example Sentences within the Corpus

The tables in Fig. 12 show the hierarchical relationship between a sentence and a corpus, with intervening structural units (such as paragraph and document) and identification, where available, of the **genre**. This information will permit certain kinds of statistical studies of word usages over document types.

7.7 Full Display of FN-2 Database

Fig. 13 shows the structure of the entire FrameNet-2 database, including all the relations among tables.

8 Advances and Applications

As evaluation of the work of FrameNet-2, several NLP applications, or more accurately, feasibility studies for such applications, will be conducted, especially through subcontracts with San Diego State University, the University of Colorado, and SRI International. The database of annotated sentences from FrameNet-1 is currently providing a training corpus for **automatic FE labeling**; a Computer Science dissertation on this topic is in progress (Gildea and Jurafsky, 2000).

Information about the separate syntactic combinatorics of different senses of words will figure in the development of techniques for **word sense disambiguation**. (E.g., the 'quarrel' sense of *argue* or *argument* can quite reliably be selected in the neighborhood of *with-* phrases followed by personal nouns, or *over-* phrases, while the 'reasoning' sense is selected in the presence of *that*-clauses and prepositional phrases with *for* and *against*.) Current annotations do not reflect the frequency of items in the Corpus, but to the extent that automatic tagging and sense selection become successful, meaningful frequency studies of senses, valences, and FE-realization preferences will become possible and be carried out under the direction of co-PI Dan Jurafsky.

Two developments of FrameNet-2 work should lead to progress in **machine translation** and **information extraction**. The first is a decision to expand the vocabulary coverage in a number of restricted areas (one of them being newspaper accounts of crimes), and the second elements of compounds from the point of view of their heads—and as part of the entries for those heads—by identifying the modifying element in a compound in terms of the FE it satisfied in the frame evoked by the head.

is to develop meaning-representations from which inferences can be generated (using a type of **simulation semantics** developed at ICSI, under the direction of co-PI Srinu Narayanan, and its symbolic representations in a “flat semantics” modeled after Minimal Recursion Semantics (Copestake et al., ms.)). The translation work will be directed by co-PI Mark Gawron, building on a pilot Japanese FrameNet based on the language of newspaper accounts of crimes.

9 Conclusion

In claiming that the FrameNet-2 database will offer a lexicon with “deep semantics” we have in mind

- the core of annotated sentences and what it offers directly, i.e.: lexical items and their combinatorial properties, both from the point of view of frame semantics and grammar; selectionally transparent nouns; support verbs and lexical functions;
- the information that can be derived automatically from that core, e.g., by parsing the labeled phrases, revealing commonalities in the relation between syntactic marking (complementation types, preposition marking, etc.) and FEs, the semantic properties of governing words, etc.;
- the lexicographer-added information about profiling, frame composition (scene structure), inheritance, and blending; and
- the formalization of frame structure that we expect will allow the operation of compositional semantic principles (for integrating the meanings of words into phrases and phrases into sentences and texts).

To all of this we should add

- the value that linguistic analysis can offer when the data are examined for generalizations nobody connected with FrameNet had in mind.⁹

Acknowledgments

The authors are indebted to the contributions of John Lowe (Technical Director of FN-1), Sue Atkins (lexicography consultant), Christopher Johnson and Josef Ruppenhofer (senior linguists in FN-1 and FN-2 respectively), Susanne Gahl (designer of local corpus searching tools), and Hiroaki Sato (creator of the FrameSQL search interface).

⁹This seems like a place where corpus-based linguistics and introspection-based linguistics can meet. At present there is speculation among us that certain resultative expressions with *into* require or impose an iterative and/or cumulative interpretation on the associated verb. Examples that support this—including grammaticality judgements—are the contrast between *beat me into submission*, *pummeled me into submission*, etc., and the unacceptable **hit me into submission*. The former two imply repeated impact, but *hit* does not. Other examples: *nagged me into taking them to the circus*, *badgered me into taking them to the circus*, as opposed to **begged me into taking them to the circus*. One instance of begging does not constitute nagging; one annoying appeal does not constitute badgering.

FrameNet II Entity Relationship Diagram

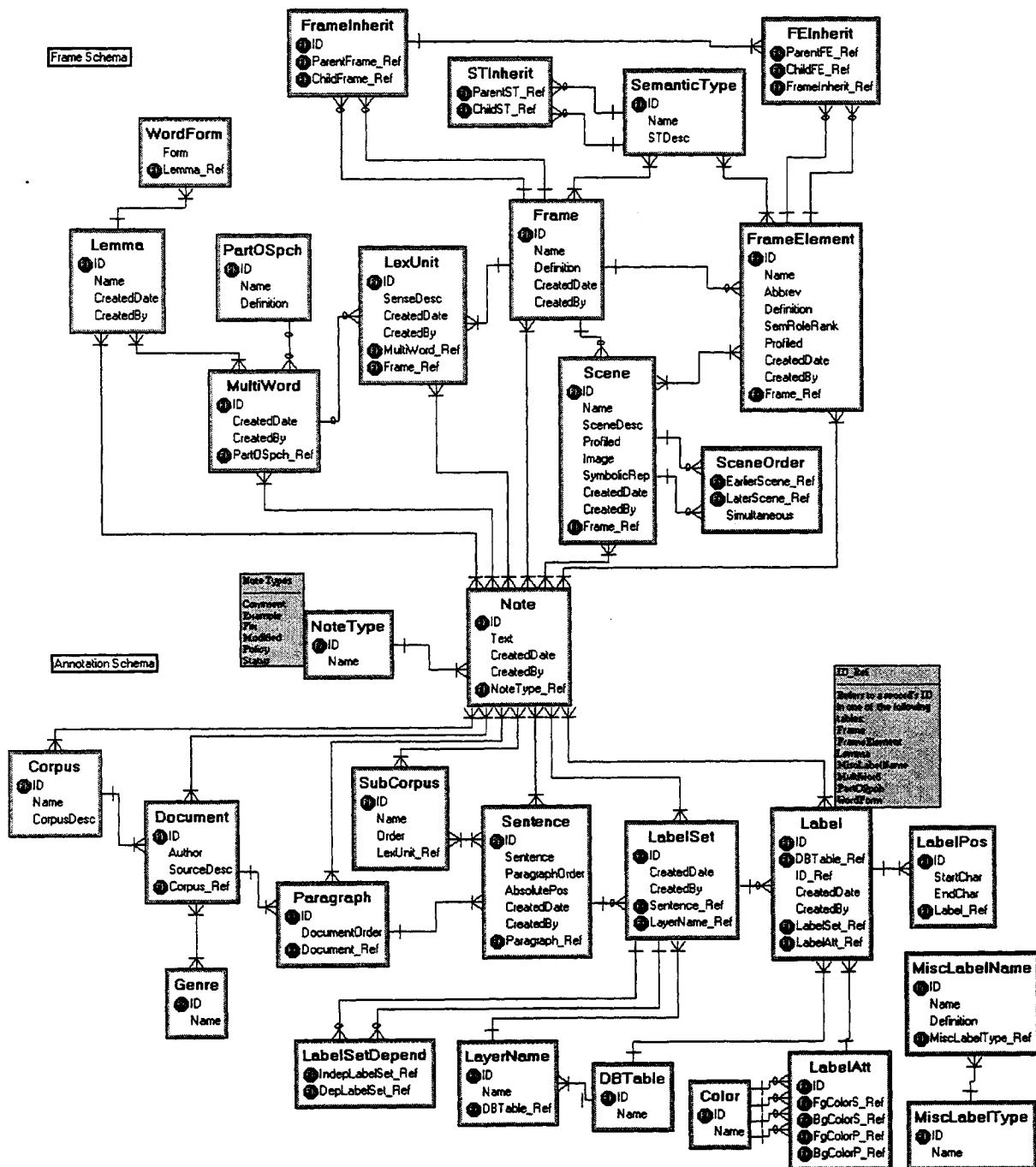


Figure 13: Overall database, showing all Tables and Relations

References

- Atkins, B.T.S., J. Kegl, and Beth Levin. 1988. Anatomy of a verb entry: from linguistic theory to lexicographic practice. *International Journal of Lexicography*, 1(2):84-126.
- Baker, Collin F., Charles J. Fillmore, and John B. Lowe. 1998. The berkeley framenet project. In *COLING-ACL '98: Proceedings of the Conference, held at the University of Montréal*, pages 86-90. Association for Computational Linguistics.
- Condoravdi, Cleo and J. Mark Gawron. 1996. The context dependency of implicit arguments. In Makoto Kanazawa, Christopher Piñon, and Henriette de Swart, editors, *Quantifiers, Deduction and Context*. CSLI Publications, Stanford, pages 1-32.
- Copestake, Ann, Dan Flickinger, Ivan Sag, and Carl Pollard. ms. Minimum recursion semantics: An introduction. <http://www-csli.stanford.edu/aac/papers.html>.
- Fillmore, Charles J. 1986. Pragmatically controlled zero anaphora. In *Proceedings of the 12th Annual Meeting of the Berkeley Linguistics Society*, pages 95-107.
- Fillmore, Charles J. and B.T.S. Atkins. 1992. Towards a frame-based lexicon: The semantics of RISK and its neighbors. In Adrienne Lehrer and Eva Feder Kittay, editors, *Frames, Fields and Contrasts*. Lawrence Erlbaum Associates.
- Fillmore, Charles J. and B.T.S. Atkins. 1994. Starting where the dictionaries stop: The challenge for computational lexicography. In B.T.S. Atkins and A. Zampolli, editors, *Computational Approaches to the Lexicon*. Clarendon Press.
- Fillmore, J. Charles and B.T.S. Atkins. 1998. Framenet and lexicographic relevance. In *Proceedings of the First International Conference on Language Resources and Evaluation, Granada, Spain*.
- Fontenelle, Thierry. 1997. *Turning a Bilingual Dictionary into a Lexical-Semantic Database*. Number 79 in Lexicographica Series Maior. Max Niemeyer Verlag, Tübingen.
- Gildea, Daniel and Daniel Jurafsky. 2000. Automatic labeling of semantic roles. In *ACL 2000: Proceedings of ACL 2000, Hong Kong*. Association for Computational Linguistics.
- Johnson, Christopher. 1999. Multiple frame inheritance in lexical descriptions. Annual Meeting of the Linguistic Society of America, held in Los Angeles, January.
- Langacker, Ronald W. 1987. *Foundations of Cognitive Grammar, Volume I: Theoretical Prerequisites*. Stanford University Press, Stanford, California.
- Lowe, John B., Collin F. Baker, and Charles J. Fillmore. 1997. A frame-semantic approach to semantic annotation. In Marc Light, editor, *Tagging Text with Lexical Semantics: Why, What and How?* Special Interest Group on the Lexicon, Association for Computational Linguistics.
- Mel'čuk, Igor'. 1996. Lexical functions: A tool for the description of lexical relations in the lexicon. In L. Wanner, editor, *Lexical Functions in Lexicography and Natural Language Processing*. Benjamins, Philadelphia, pages 37-102.
- Sadock, Jerrold M. 1991. *Autolexical syntax: A Theory of Parallel Grammatical Representations*. Studies in contemporary linguistics. University of Chicago Press, Chicago.
- Wilks, Yorick, Brian M. Slator, and Louise M. Guthrie. 1996. *Electric Words: Dictionaries, Computers, and Meanings*. ACL-MIT Press series in natural-language processing. MIT Press, Cambridge, Mass.

