The Berkeley FrameNet Project

Collin F. Baker, Charles J. Fillmore, and **John B. Lowe**
{collinb, fillmore, jblowe}@icsi.berkeley.edu
International Computer Science Institute
1947 Center St. Suite 600
Berkeley, Calif., 94704

Abstract

FrameNet is a three-year NSF-supported project in corpus-based computational lexicography, now in its second year (NSF IRI-9618838, "Tools for Lexicon Building"). The project's key features are (a) a commitment to corpus evidence for semantic and syntactic generalizations, and (b) the representation of the valences of its target words (mostly nouns, adjectives, and verbs) in which the semantic portion makes use of frame semantics. The resulting database will contain (a) descriptions of the semantic frames underlying the meanings of the words described, and (b) the valence representation (semantic and syntactic) of several thousand words and phrases, each accompanied by (c) a representative collection of annotated corpus attestations, which jointly exemplify the observed linkings between "frame elements" and their syntactic realizations (e.g. grammatical function, phrase type, and other syntactic traits). This report will present the project's goals and workflow, and information about the computational tools that have been adapted or created in-house for this work.

Introduction

The Berkeley FrameNet project[1] is producing frame-semantic descriptions of several thousand English lexical items and backing up these descriptions with semantically annotated attestations from contemporary English corpora[2].

These descriptions are based on hand-tagged semantic annotations of example sentences extracted from large text corpora and systematic analysis of the semantic patterns they exemplify by lexicographers and linguists. The primary emphasis of the project therefore is the encoding, by humans, of semantic knowledge in machine-readable form. The intuition of the lexicographers is guided by and constrained by the results of corpus-based research using high-performance software tools.

The semantic domains to be covered are: HEALTH CARE, CHANCE, PERCEPTION, COMMUNICATION, TRANSACTION, TIME, SPACE, BODY (parts and functions of the body), MOTION, LIFE STAGES, SOCIAL CONTEXT, EMOTION and COGNITION.

Scope of the Project

The results of the project are (a) a lexical resource, called the FrameNet database[3], and (b) associated software tools. The database has three major components (described in more detail below:

- **Lexicon** containing entries which are composed of: (a) some conventional dictionary-type data, mainly for the sake of human readers; (b) FORMULAS which capture the morphosyntactic ways in which elements of the semantic frame can be realized within

[1]The project is based at the International Computer Science Institute (1947 Center Street, Berkeley, CA). A fuller bibliography may be found in (Lowe et al., 1997)

[2]Our main corpus is the British National Corpus. We have access to it through the courtesy of Oxford University Press; the POS-tagged and lemmatized version we use was prepared by the Institut für Maschinelle Sprachverarbeitung of the University of Stuttgart). The European collaborators whose participation has made this possible are Sue Atkins, Oxford University Press, and Ulrich Heid, IMS-Stuttgart.

[3]The database will ultimately contain at least 5,000 lexical entries together with a parallel annotated corpus, these in formats suitable for integration into applications which use other lexical resources such as WordNet and COMLEX. The final design of the database will be selected in consultation with colleagues at Princeton (WordNet), ICSI, and IMS, and with other members of the NLP community.

the phrases or sentences built up around the word; (c) links to semantically ANNOTATED EXAMPLE SENTENCES which illustrate each of the potential realization patterns identified in the formula;[4] and (d) links to the FRAME DATABASE and to other machine-readable resources such as WordNet and COMLEX.

- **Frame Database** containing descriptions of each frame's basic conceptual structure and giving names and descriptions for the elements which participate in such structures. Several related entries in this database are schematized in Fig. 1.
- **Annotated Example Sentences** which are marked up to exemplify the semantic and morphosyntactic properties of the lexical items. (Several of these are schematized in Fig. 2). These sentences provide empirical support for the lexicographic analysis provided in the frame database and lexicon entries.

These three components form a highly relational and tightly integrated whole: elements in each may point to elements in the other two. The database will also contain estimates of the relative frequency of senses and complementation patterns calculated by matching the senses and patterns in the hand-tagged examples against the entire BNC corpus.

Conceptual Model

The FrameNet work is in some ways similar to efforts to describe the argument structures of lexical items in terms of case-roles or theta-roles,[5] but in FrameNet, the role names (called **frame elements** or **FEs**) are local to particular conceptual structures (**frames**); some of these are quite general, while others are specific to a small family of lexical items.

For example, the TRANSPORTATION frame, within the domain of MOTION, provides MOVERS, MEANS of transportation, and PATHS;[6] subframes associated with individual words inherit all of these while possibly adding some of their own. Fig. 1 shows some of the subframes, as discussed below.

frame(TRANSPORTATION) frame_elements(MOVER(S), MEANS, PATH) scene(MOVER(S) move along PATH by MEANS)
frame(DRIVING) inherit(TRANSPORTATION) frame_elements(DRIVER (=MOVER), VEHICLE (=MEANS), RIDER(S) (=MOVER(S)), CARGO (=MOVER(S))) scenes(DRIVER starts VEHICLE, DRIVER controls VEHICLE, DRIVER stops VEHICLE)
frame(RIDING_1) inherit(TRANSPORTATION) frame_elements(RIDER(S) (=MOVER(S)), VEHICLE (=MEANS)) scenes(RIDER enters VEHICLE, VEHICLE carries RIDER along PATH, RIDER leaves VEHICLE)

Figure 1: A subframe can inherit elements and semantics from its parent

The DRIVING frame, for example, specifies a DRIVER (a principal MOVER), a VEHICLE (a particularization of the MEANS element), and potentially CARGO or RIDER as secondary movers. In this frame, the DRIVER initiates and controls the movement of the VEHICLE. For most verbs in this frame, DRIVER or VEHICLE can be realized as subjects; VEHICLE, RIDER, or CARGO can appear as direct objects; and PATH and VEHICLE can appear as oblique complements.

Some combinations of frame elements, or **Frame Element Groups (FEGs)**, for some real corpus sentences in the DRIVING frame are shown in Fig. 2.

A RIDING_1 frame has the primary mover role as RIDER, and allows as VEHICLE those driven

[4] In cases of accidental gaps, clearly marked invented examples may be added.

[5] The semantic frames for individual lexical units are typically "blends" of more than one basic frame; from our point of view, the so-called "linking" patterns proposed in LFG, HPSG, and Construction Grammar, operate on higher-level frames of action (giving agent, patient, instrument), motion and location (giving theme, location, source, goal, path), and experience (giving experiencer, stimulus, content), etc. In some but not all cases, the assignment of syntactic correlates to frame elements could be mediated by mapping them to the roles of one of the more abstract frames.

[6] A detailed study of motion predicates would require a finer-grained analysis of the Path element, separating out Source and Goal, and perhaps Direction and Area, but for a basic study of the transportation predicates such refined analysis is not necessary. In any case, our work includes the separate analysis of the frame semantics of directional and locational expressions.

FEG	Annotated Example from BNC
D	[$_D$ Kate] **drove** [$_P$ home] in a stupor.
V, D	A pregnant woman lost her baby after she fainted as she waited for a bus and fell into the path of [$_V$ a lorry] **driven** [$_D$ by her uncle].
D, P	And that was why [$_D$ I] **drove** [$_P$ eastwards along Lake Geneva].
D, R, P	Now [$_D$ Van Cheele] was **driving** [$_R$ his guest] [$_P$ back to the station].
D, V, P	[$_D$ Cumming] had a fascination with most forms of transport, **driving** [$_V$ his Rolls] at high speed [$_P$ around the streets of London].
D+R, P	[$_D$ We] **drive** [$_P$ home along miles of empty freeway].
V, P	Over the next 4 days, [$_V$ the Rolls Royces] will **drive** [$_P$ down to Plymouth], following the route of the railway.

Figure 2: Examples of Frame Element Groups and Annotated Sentences

by others.[7] In grammatical realizations of this frame, the RIDER can be the subject; the VEHICLE can appear as a direct object or an oblique complement; and the PATH is generally realized as an oblique.

The FrameNet entry for each of these verbs will include a concise formula for all semantic and syntactic combinatorial possibilities, together with a collection of annotated corpus sentences in which each possibility is exemplified. The syntactic positions considered relevant for lexicographic description include those that are internal to the maximal projection of the target word (the whole VP, AP, or NP for target V, A or N), and those that are external to the maximal projection under precise structural conditions; the subject, in the case of VP, and the subject of support verbs in the case of AP and NP. [8]

Used in NLP, the FrameNet database should make it possible for a system which finds a valence-bearing lexical item in a text to know (for each of its senses) where its individual arguments are likely to be found. For example, once a parser has found the verb *drive* and its direct object NP, the link to the DRIVING frame will suggest some semantics for that NP, e.g. that a person as direct object probably represents the RIDER, while a non-human proper noun is probably the VEHICLE.

For practical lexicography, the contribution of the FrameNet database will be its presentation of the full range of use possibilities for individual words, documented with corpus data, the model examples for each use, and the statistical information on relative frequency.

Organization and Workflow

Overview

The computational side of the FrameNet project is directed at efficiently capturing human insights into semantic structure. The majority of the work involved is marking text with semantic tags, specifying (again by hand) the structure of the frames to be treated, and writing dictionary-style entries based the results of annotation and *a priori* descriptions. With the exception of the example sentence extraction component, all the software modules are highly interactive and have substantial user interface requirements. Most of this functionality is provided by WWW-based programs written in PERL.

Four processing steps are required produce the FrameNet database of frame semantic representations: (a) generating initial descriptions of semantic and syntactic patterns for use in corpus queries and annotation ("Preparation"), (b) extracting good example sentences ("Subcorpus Extraction"), (c) marking (by hand) the constituents of interest ("Annotation"), and (d) building a database of lexical semantic representations based on the annotations and other data ("Entry Writing"). These are discussed briefly below and shown Fig. .

Workflow and Personnel

As work on the project has progressed, we have defined several explicit roles which project participants play in the various steps. these roles are referred to as **Vanguard** (1.1 in Fig.), **Annotators** (3.1) and **Rearguard** (4.1). These are purely functional designations: the same person may play different roles at different

[7] A separate frame RIDING_2 that applies to the English verb *ride* selects means of transportation that can be straddled, such as bicycles, motorcycles, and horses.

[8] For causatives, the object of the support verb is included; for details, see Fillmore and Atkins (forthcoming).

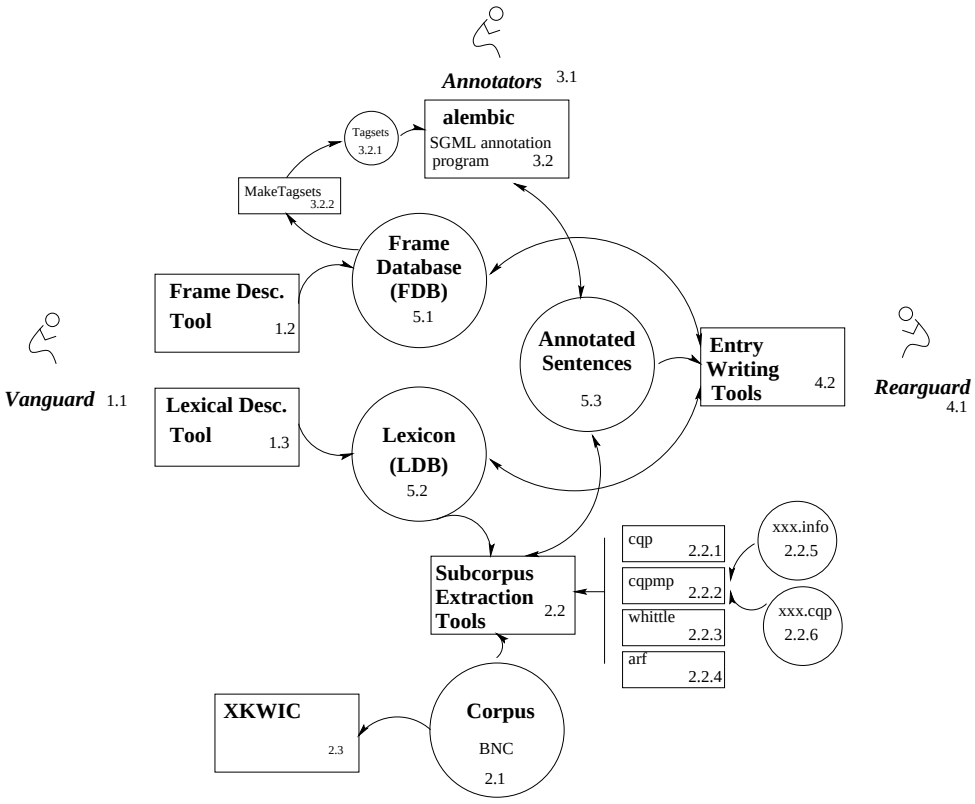

Figure 3: Workflow, Roles, Data Structures and Software

times.[9]

1. **Preparation**. The Vanguard (1.1) prepares the initial descriptions of frames, including lists of frames and frame elements, and adds these to the Frame Database (5.1) using the Frame Description tool (1.2). The Vanguard also selects the major vocabulary items for the frame (the *target words*) and the syntactic patterns that need to be checked for each word, which are entered in the Lexical Database (5.2) by means of the Lexical Database Tool (1.3).

2. **Subcorpus Extraction**. Based on the Vanguard's work, the subcorpus extraction tools (2.2) produce a representative collection of sentences containing these words.

This selection of examples is achieved through a hybrid process partially controlled by the preliminary lexical description of each lemma. Sentences containing the lemma are extracted from from a corpus and classified into subcorpora by syntactic pattern (2.2.1) using a CASCADE FILTER (2.2.2, 2.2.5, 2.2.6) representing a partial regular-expression grammar of English over part-of-speech tags (cf. Gahl (forthcoming)), formatted for annotation (2.2.4) , and automatically sampled (2.2.3) down to an appropriate number.

(If these heuristics fail to find appropriate examples by means of syntactic patterns, sentences are selected using INTERACTIVE SELECTION TOOLS (2.3)).

3. **Annotation**. Using the annotation software (3.2) and the tagsets (3.2.1) derived from the Frame Database, the Annotators (3.1) mark selected constituents in the extracted subcorpora according to the frame elements which they realize, and identify canonical examples, novel patterns, and problem sentences.[10]

4. **Entry Writing**. The Rearguard (4.1)

[9] Of course there are other staff members who write code and maintain the databases. This behind-the-scenes work is not shown in Fig. .

[10] We are building a "constituent type identifier" which will semi-automatically assign Grammatical Function (GF), and Phrase Type (PT) attributes to these FE-marked constituents, eliminating the need for Annotators to mark these.

reviews the skeletal lexical record created by the Vanguard, the annotated example sentences (5.3), and the FEGs extracted from them, and builds both the entries for the lemmas in the Lexical Database (5.2) and the frame descriptions in the Frame Database (5.1), using the Entry Writing Tools (4.2).

Implementation

Data Model

The data structures described above are implemented in SGML.[11] Each is described by a DTD, and these DTDs are structured to provide the necessary links between the components.

Software

The software suite currently supporting database development is an aggregate of existing software tools held together with PERL/CGI-based "glue". In order to get the project started, we have depended on off-the-shelf software which in some cases is not ideal for our purposes. Nevertheless, using these programs allowed us to get the project up and running within just a few months. We describe below in approximate order of application the programs used and their state of completion.

- Frame Description Tool (1.2) (in development) An interactive, web-based tool.
- Lexical Description Tool (1.3) (prototype) An interactive, web-based tool.
- CQP (2.2.1) is a high-performance Corpus Query Processor, developed at IMS Stuttgart (IMS, 1997). The cascade filter, which partitions lemma-specific subcorpora by syntactic patterns, is built using a preprocessor (written in PERL, 2.2.2) which generates CQP's native query language.
- XKWIC (2.3) is an X-window, interactive tool, also from IMS, which facilitates manipulating corpora and subcorpora.
- Subcorpora are prepared for annotation by a program ("arf" for *Annotation Ready Formatter*, 2.2.4) which wraps SGML tags around sentences, target words, comments and other distinguishable text elements. Another program, "whittle" (2.2.3), combines subcorpora in a preselected order, removing very long and very short sentences, and sampling to reduce large subcorpora.
- Alembic (3.2) (Mitre, 1998), allows the interactive markup (in SGML) of text files according to predefined tagsets (3.2.1). It is used to introduce frame element annotations into the subcorpora.
- Sgmlnorm, etc. (from James Clark's SGML tool set) are used to validate and manage the SGML files.
- Entry Writing Tools (4.2) (in development)
- Database management tools to manage the catalog of subcorpora, schedule the work, render the SGML files into HTML for convenient viewing on the web, etc. are being written in PERL. RCS maintains version control over most files.

Conclusion

At the time of writing, there is something in place for each of the major software components, though in some cases these are little more than stubs or "toy" implementations. Nearly 10,000 sentences exemplifying just under 200 lemmas have been annotated; there are over 20,000 frame element tokens marked in these example sentences. About a dozen frames have been specified, which refer to 47 named frame elements. Most of these annotations have been accomplished in the last few months since the software for corpus extraction, frame description, and annotation became operational. We expect the inventory to increase rapidly. If the proportions cited hold constant as the Framenet database grows, the final database of 5,000 lexical units may contain 250,000 annotated sentences and over half a million tokens of frame elements.

[11] Eventually, we plan to migrate to an XML data model, which appears to provide more flexibility while reducing complexity. Also, the FrameNet software is being developed on Unix, but we plan to provide cross-platform capabilities by making our tool suite web-based and XML-compatible.

References

Charles J. Fillmore and B. T. S. Atkins. forthcoming. FrameNet and lexicographic rele-

vance. In *Proceedings of the First International Conference On Language Resources And Evaluation, Granada, Spain, 28-30 May 1998.*

Susanne Gahl. forthcoming. Automatic extraction of subcorpora based on subcategorization frames from a part of speech tagged corpus. In *Proceedings of the 1998 COLING-ACL conference.*

Institut für maschinelle Sprachverarbeitung IMS. 1997. IMS corpus toolbox web page at stuttgart. http://www.ims.uni-stuttgart.de/~oli/CorpusToolbox/.

John B. Lowe, Collin F. Baker, and Charles J. Fillmore. 1997. A frame-semantic approach to semantic annotation. In *Tagging Text with Lexical Semantics: Why, What, and How? Proceedings of the Workshop*, pages 18–24. Special Interest Group on the Lexicon, Association for Computational Linguistics, April.

Mitre. 1998. Alembic Workbench web page at Mitre corp. http://www.mitre.org/resources/ centers/ advanced_info/ g04h/workbench.html.