Susanne GAHL, University of California at Berkeley

Automatic Extraction of Subcategorization Frames for Corpus-based Dictionary-building

Abstract

This paper presents a method for automatically extracting subcorpora isolating different subcategorization frames for nouns, adjectives, and verbs in the 100 mi. word BNC. The tool is being used in the FrameNet project, an NSF-funded project that is involved in producing a database and tools for dictionary-building, based on the principles of Frame Semantics. The subcorpora are used (1) to facilitate the selection of corpus lines illustrating the full range of semantic and syntactic combinatory possibilities of a given lemma, (2) to determine relative frequencies of different syntactic contexts of each lemma in the database. The database thus created, which will be human- and computer-readable, will be a rich resource for lexicographers, as well as for researchers in lexicology and natural language processing.

Keywords: dictionary-building, corpus linguistics, subcategorization extraction, Frame Semantics

1. Introduction

1.1. The FrameNet project

The set of tools described in this paper form part of the FrameNet project[1] conducted at the University of California.[2] The end product of the FrameNet project is a database consisting of (1) a list of semantic frames that are necessary to describe the meanings of words in 13 different semantic domains (health care, chance, perception, communication, transaction, time, space, body, motion, life stages, social context, emotion, and cognition), and (2) a database of 5000 lexical entries. Each entry contains frame-semantic, combinatory, and probabilistic descriptions of a lexical unit, at the level of lexical semantics and syntactic subcategorization, and describes the linking of semantic Frame Elements to syntactic units. The purpose of the subcorpus extraction within the FrameNet project is to provide the lexicographers with corpus examples of each syntactic configuration a given lemma can occur in. Annotators then select from these syntactically-based subcorpora sentences that illustrate the ways in which Frame Elements can be syntactically realized. The workflow from the initial linguistic specification through the selection and annotation of corpus lines, to the final entry is described in the next section.

1.2. FrameNet work flow

The flowchart in figure 1 below describes the overall workflow of the FrameNet project, from the initial frame description, through subcorpus extraction, to selection and annotation of corpus lines, to the preparation of the final entry. In the sections that follow, we will describe each step in greater detail. A fuller description can be found in Lowe et al. (1998).

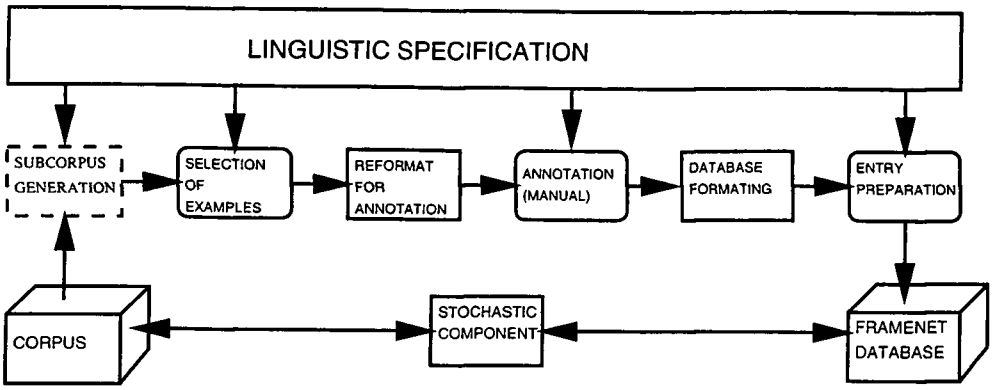

Figure 1: FrameNet workflow diagram

1.2.1. Linguistic Specification

The initial phase of linguistic specification covers two areas: We first prepare an initial description of a Frame, including a list of Frame Elements. For example, one Frame in the "health" domain is the "allergy" Frame. Elements of this Frame include the Protagonist (i.e. the person afflicted with an allergy), and the Trigger (the substance or process triggering the allergic response). Based on the list of Frame Elements, a set of tags is prepared for use in the annotation process. Secondly, for each word in each domain, we prepare an initial description of the syntactic contexts in which the word can be expected to appear. This initial specification of expected patterns is based on (a) machine-readable and print dictionaries, such as Levin (1993), the COMLEX syntactic database (cf. Macleod et al. 1994), Hornby (1989) and others, and (b) a preliminary inspection of corpus data.

The initial description is what determines which syntactic patterns are extracted from the corpus. Concordance lines that do not match any of the patterns that were specified in the initial description are saved in a set of specially marked separate subcorpora, for further inspection by the lexicographers. Where these "remainder" subcorpora are unduly large, lexicographers can revise the linguistic specification and determine whether the remainder subcorpora should be re-submitted to the automatic extraction process.

1.2.2. Subcorpus generation

We first create a subcorpus consisting of all the concordance lines from the BNC that contain each given lemma. This 'lemma-subcorpus' is then partitioned into smaller subcorpora according to the specifications in the initial description of the word. We will describe the subcorpus generation process in greater detail in section 2 below.

1.2.3. Selection of examples

The next step in the process is the selection of the corpus lines that will form part of the final lexical entry. The results of the subcorpus extraction are submitted to the lexicographers, who then select corpus lines to be included in the database.

1.2.4. Annotation

Lexicographers select examples and each Frame Element that is overtly instantiated in the sentences. A more detailed discussion of this process, and examples of annotated corpus lines, can be found in Lowe et al. (1998).

1.2.5. Entry preparation

Based on the initial linguistic specification and the annotated corpus lines, lexicographers then prepare the final entry for each lexical unit. The final entries will contain information on each frame, and on the ways in which Frame Elements can be instantiated with each lemma.

1.2.6. Stochastic information

The stochastic component of the database will include estimated probabilities of the various combinatorics of each lexical unit, based on frequency distributions of the syntactic patterns, and on the distribution of Frame Elements in the contexts for each lemma. For example, our preliminary results indicate that the verbs *cure* and *heal*, although similar to some degree in the syntactic complementation patterns they allow, differ with regard to the relative frequencies with which their accompanying Frame Elements in the "healing" Frame are instanstiated. The entries for individual words state the full combinatorial possibilities of the word, and the ways in which frame elements can be instantiated, with examples from the corpus, as in the entry for the noun *allergy* in figure 2. (Note: only a subset of the examples in the entry is shown here.)

Frame:	Allergy		
Frame Elements:	Protagonist (Prot) Trigger (Trig)		
Protagonist can be realized as		FEG	Examples
• argument of support verbs HAVE, ACQUIRE, GIVE			
Pat has an allergy		Prot Subj NP	1
• possessive determiner of target			
Pat's allergy		Prot Gen Poss	3, 4
• prepositional object			
allergy in children		Prot Comp PP	2
Trigger can be realized as			
• prepositional object			
allergy to milk		Trig Comp PP	2, 4
• noun modifier in compound with target as head			
milk allergy		Trig Mod N	1, 3

Examples:

1. On top of all that, **Copper** has always had a **dust allergy** and he got very congested so we had to give him powders to keep his lungs clear.
2. **Allergies to wood dust** can develop **in staff** and consideration should be given to using dust-free sawdust and to the staff wearing masks when handling the dry bedding.
3. Peter immediately explained **Carol's codeine allergy** but the doctor replied: 'But she's had some already."
4. If we wished to test a new theory about **Napoleon's allergy to snuff**, say, it would not make sense to examine look-alikes of Napoleon's clothing.

Figure 2: An entry for *allergy*

The information included in the FrameNet database goes beyond that contained, e.g., in the COMLEX database and in WordNet. Most importantly, unlike COMLEX or WordNet, the FrameNet database links syntactic and semantic information about words.

2. Subcorpus extraction

In this section, we will describe the subcorpus extraction process in greater detail. More information can be found in Gahl (1998).

The extraction tool consists of a set of batch-files for use with CQP (Corpus Query Processor) (CQP), which is part of the IMS corpus workbench (cf. Christ 1994 a, b). CQP is a general corpus query processor for complex queries over annotated information types, including part-of-speech tags, morphosyntactic tags, lemmas and sentence boundaries. The corpus queries are written in the CQP corpus query language, which uses regular expressions over part-of-speech tags, lemmas, morphosyntactic tags, and sentence boundaries. For details, see Christ (1994a.).

2.1. Subcorpus extraction for nouns and adjectives

The extraction tool is used to create syntactic subcorpora for nouns, adjectives, and verbs. For all three classes of words, we first create a lemma-based subcorpus which is then partitioned into smaller subcorpora according to the syntactic environments the lemma is found in. The searches apply in a cascading fashion. That is to say, the lines matching each query are removed from the lemma subcorpus, and the remainder is then submitted to the next query or set of queries.

For nouns and adjectives, we are able to extract prepositional, clausal, infinitival, and gerundial complements. For adjectives, we further isolate prenominal uses. For nouns, separate subcorpora are created for complements following the head noun, as well as for compounds. In addition, the tool accommodates searches for compounds and for possessor phrases (*the participants' answers to the question posed by the chair, my milk allergy*). Even though these categories are not tied to the syntactic subcategorization frames of the target lemmas, they often Frame Elements (Fillmore 1982, Lowe et al. 1998).

As an example, some of the subcorpora for the noun *answer* are listed in table 1, along with examples of phrases found in each subcorpus:

description	subcorpus name	match (from corpus)
possessor phrase + target noun	[answer-N-poss]	*the participants' answers*
target noun as head of compound	[answer-N-cmpd]	*classroom answers*
poss. phrase + noun compound	[answer-N-posscmpd]	*his test answers*
target noun + PPby complement	[answer-N-ppby]	*an answer by the Minister of State*
target noun + PPfrom complement	[answer-N-ppfrom]	*an answer from Ellen*
target noun + PPof complement	[answer-N-ppof]	*the answer of a gentleman*
target noun + PPto complement	[answer-N-ppto]	*the answer to using this type of fabric*

Table 1: selected subcorpora for the noun *answer*

Some of the subcorpora listed above overlap with one another. For example, the answer-N-poss subcorpus, which contains matches like *the participants' answers*, also contains instances of *answer* followed by a prepositional phrase complement, as in *the participants' answers to our questions*. Before subcorpora get passed on to the selection-and-annotation

process, subcorpora that overlap in this way are intersected with one another. The resulting intersections and complements are then saved separately before being passed on to the lexicographers. As a result, the final subcorpora that are passed on to selection-and-annotation are always mutually exclusive. This prevents the lexicographers from accidentally selecting the same line more than once.

2.2. Subcorpus extraction for verbs

The extraction process for verbs proceeds in two stages. During the first stage, the lemma-subcorpus is queried for syntactic patterns involving 'displaced' arguments, e.g. WH-movement, tough-movement, and passives. The resulting subcorpora are homogeneous with respect to major constituent order, which simplifies the subsequent searches for complementation patterns considerably. For example, the string *hit by a car* in a passive use of the transitive verb *hit*, as in *he got hit by a car*, might otherwise erroneously be classified by the system as containing an intransitive use of *hit*. More generally, most 'movement' contexts falsely match search strings for intransitives, unless steps be taken to filter out such contexts. For example, the - somewhat simplified - query shown in table 2 below finds passive sentences involving coordination structures, such as *this condition can be effectively treated and cured.*

query expression	description	corpus match	
[(lemma="be|being|get")&(word!="s")&(pos!="NN1|NN2")]	passive auxiliary	*been*	*be*
[(class!="c")|(class="c"&pos="PUQ")|(word=",")]{0,4} [pos="VVN|VVD|VVD-VVN|AJ0-VVN|ADJ0-VVD"] [pos="AVP"]? [(((pos="PUQ")|(word=","))&(class="c"))|(class!="c")]{0,3}	past participle (obligatory), modifiers (optional)	*ameliorated*	*treated*
[word="or"|word="and"|word="but"|word=","|word="rather than"|word="if"] [(pos!="VVN|VVD|VBB|VBD|VBG|VBI|VBN|VBZ|VDB|VDD|VDG|VDI|VDM|VDZ|VHB|VHD|VHG|VHI|VHN|VHZ|VM0|VVB|VVG|VVI|VVZ|AT0|DPS|DT0|DTQ|PNI|PNP|PNQ")|(pos="PNQ"&word=".*ever")]{0,3}	complements and adjuncts of first verb in the coordination structure (optional), conjunction (obligatory), negation (optional), modifiers (optional)	*but not*	*for it and*
[lemma="cure"& pos="VVB|VVD|VVG|VVI|VVN|VVZ|AJ0-VVN|AJ0|VVD|AJ0-VVG|NN1-VVB|NN1-VVG|NN2-VVZ|VVD-VVN"&pos="VVN"&pos!="AJ0"][pos!="AJ0|AJC|AJS|AT0|CRD|DPS|DT0|DTQ|NN0|NN1|NN2|NP0|ORD|PNI|PNP|PNQ|PNX|VVG|VVD"]	past participle of target lemma	*cured*	*cured*

Table 3: a regular expression matching passives in coordination structures

A more complete representation of the matches found by this particular query is found in figure 3 below, which shows the results of the same query, viewed in Xkwic, another tool in the IMS corpus workbench (cf. Christ 1994b).

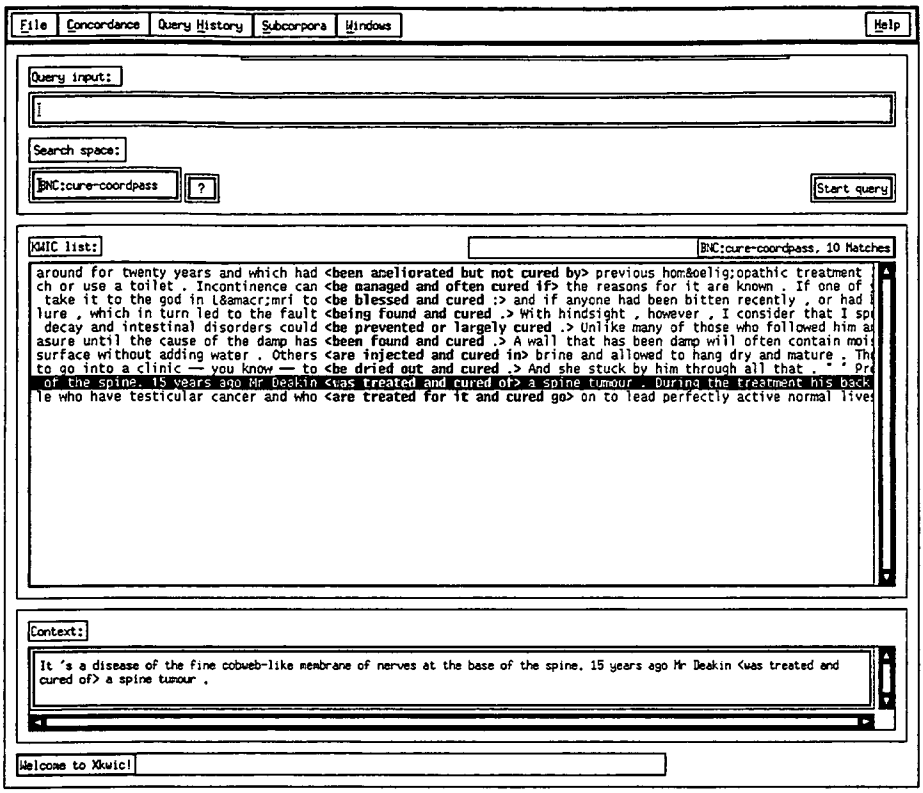

Figure 3: Xkwic view of *cure* in passives in coordination structures

During the second stage of the extraction process for verbs, we isolate syntactic contexts not involving movement phenomena. The resulting subcorpora are based on categories similar to those used in the COMLEX database. For example, we distinguish simple transitives, ditransitives, prepositional phrases, VP- and clausal-type complements. A fuller list of the verb frames that are currently searchable is given in figure 4 below, along with an example of each pattern. The categories we are using are roughly based on those used in the COMLEX syntactic dictionary (Macleod et al. 1995).

intransitive	'worms wiggle'	pp	'look at the picture'
np	'kiss me'	pp_pp	'turned from a frog into a prince'
np_np	'brought her flowers'	Pvping	'responded by nodding her head'
np_pp	'replaced it with a new one'	Pwh	'wonder about how it happened'
np_Pvping	'prevented him from leaving'	intrans. part.	'touch down', 'turn over'
np_pwh	'asked her about what it all meant'	np_particle	'put the dishes away', 'put away the dishes'
np_vpto	'advised her to go'	particle_pp:	'run off with it'
np_vping	'kept them laughing'	particle_wh:	'figured out how to get there'
np_sfin	'told them (that) he was back'	vping	'needs fixing'
np_wh	'asked him where the money was'	sfin	'claimed (that) it was over'
		sbrst	'demanded (that) he leave'
np_ap	'considered him foolish'	vpto	'agreed to do it over'
np_sbrst	'had him clean up'	directquote	'no, said he', '"no", 'he said', 'he said: "no"'
ap	'turned blue'		
		adverb	'behave badly'

Figure 4: Searchable complement types for verbs

3. The macroprocessor

The cqp tool can be used with a macroprocessor[3] that allows the user to specify in a simple input file which subcorpora are to be created for a given lemma. The macroprocessor also returns the number of matches found in each subcorpus. This information will be used in the stochastic component of the project, in which estimated probabilities for each pattern will be computed.

4. Further applications of the SC extraction tool

Besides its application in the FrameNet routine, the extraction tool is also being used to select stimuli for use in psycholinguistic experiments on probabilistic parsing effects. We are currently testing aphasic speakers' sensitivity to lexical subcategorization preferences (Gahl, in preparation). Previous studies on lexical biases (or "valence probabilities") were based on psychological norming studies, such as Connine et al. (1990) or on corpora that are far smaller than the BNC, such as the Treebank corpus (Marcus et al. 1993). The lack of information on lexical biases based on larger corpora represents a serious methodological problem in psycholinguistic research which we are hoping to address in developing the extraction tool.

5. Conclusion

We have presented an overview of a tool for extracting corpus lines illustrating subcategorization patterns of nouns, verbs, and adjectives, and for determining the frequency of these patterns. The tools are currently being used as part of the FrameNet project, as well as in a psycholinguistic investigation of aphasics' speakers sensitivity to lexical valence preferences (Gahl, in preparation). An overview of the FrameNet project can be found at http://www.icsi.berkeley.edu/ ~framenet.

6. Notes

[1] I would like to thank Ulrich Heid of IMS-Universität Stuttgart for his help and support at all stages of this project. I would also like to thank the members of the FrameNet project for much valuable feedback and continued support, and Judith Eckle-Kohler (IMS Stuttgart) for useful comments on an earlier draft of this paper. Very special thanks go to the ever-helpful Collin Baker.

[2] Under NSF grant IRI 96 18838. The Principal Investigator is Charles J. Fillmore. The project is housed in the International Computer Science Institute in Berkeley, CA. An overview of the wholeproject can be found at: http://www.icsi.berkeley.edu/ framenet/

[3] Our macroprocessor was developed by Collin Baker (UC Berkeley Linguistics) and Douglas Roland (U of Colorado, Boulder).

7. References

Christ, O. (1994a). *The IMS Corpus Workbench Technical Manual.* Institut für maschinelle Sprachverarbeitung, Universität Stuttgart.

Christ, O. (1994b) *The XKwic User Manual.* Institut für maschinelle Sprachverarbeitung, Universität Stuttgart.

Eckle, J. & U. Heid. (1996). Extracting raw material for a German subcategorization lexicon from newspaper text. In *Proceedings of the 4th International Conference on Computational Lexicography COMPLEX'96*, Budapest, Hungary.

Fillmore, C. J. (1982). Frame Semantics. In *Linguistics in the morning calm*, Hanshin Publishing Co., Seoul, South Korea, pp. 111-137.

Fillmore, C. F. & B. T. Atkins. (1992). Towards a frame-based lexicon: The semantics of RISK and its neighbours. In A. Lehrer & E. F. Kittay (eds.), *Frames, Fields, and Contrasts*, pp. 75-102.

Gahl, S. (1998). Automatic extraction of subcorpora based on subcategorization frames from a part-of-speech tagged corpus. ms., ICSI-Berkeley.

Gahl, S., (in preparation). A usage-based model of aphasic sentence comprehension. UC Berkeley doctoral dissertation.

Hornby, A. S. (1989). *Oxford Advanced Learner's Dictionary of Current English. 4th edition.* Oxford University Press, Oxford, England.

Levin, B. (1993). *English Verb Classes and Alternations.* University of Chicago Press.

Lowe, J. B., Fillmore, C. J. & Baker, C. (1998). *The FrameNet project.* ms., ICSI-Berkeley.