Automatic extraction of subcorpora based on subcategorization frames from a part-of-speech tagged corpus

Susanne GAHL
UC Berkeley, Department of Linguistics
ICSI
1947 Center St, Suite 600
Berkeley, CA 94704-1105
gahl@icsi.berkeley.edu

Abstract

This paper presents a method for extracting subcorpora documenting different subcategorization frames for verbs, nouns, and adjectives in the 100 mio. word British National Corpus. The extraction tool consists of a set of batch files for use with the Corpus Query Processor (CQP), which is part of the IMS corpus workbench (cf. Christ 1994a,b).
A macroprocessor has been developed that allows the user to specify in a simple input file which subcorpora are to be created for a given lemma.
The resulting subcorpora can be used (1) to provide evidence for the subcategorization properties of a given lemma, and to facilitate the selection of corpus lines for lexicographic research, and (2) to determine the frequencies of different syntactic contexts of each lemma.

Introduction

A number of resources are available for obtaining subcategorization information, i.e. information on the types of syntactic complements associated with valence-bearing predicators (which include verbs, nouns, and adjectives). This information, also referred to as *valence information* is available both in machine-readable form, as in the COMLEX database (Macleod et al. 1995), and in human-readable dictionaries (e.g. Hornby 1989, Procter 1978, Sinclair 1987). Increasingly, tools are also becoming available for acquiring subcategorization information from corpora, i.e. for inferring the subcategorization frames of a given lemma (e.g. Manning 1993).

None of these resources provide immediate access to corpus evidence, nor do they provide information on the relative frequency of the patterns that are listed for a given lemma.

There is a need for a tool that can (1) find evidence for subcategorization patterns and (2) determine their frequencies in large corpora:

1. Statistical approaches to NLP rely on information not just on the range of combinatory possibilities of words, but also the relative frequencies of the expected patterns.
2. Dictionaries that list subcategorization frames often list expected patterns, rather than actual ones. Lexicographers and lexicologist need access to the evidence for this information.
3. Frequency information has come to be the focus of much psycholinguistic research on sentence processing (see for example MacDonald 1997). While information on word frequency is readily available (e.g. Francis and Kucera (1982)), there is as yet no easy way of obtaining information from large corpora on the relative frequency of complementation patterns.

None of these points argue against the usefulness of the available resources, but they show that there is a gap in the available information.

To address this need, we have developed a tool for extracting evidence for subcategorization patterns from the 100 mio. word British National Corpus (BNC). The tool is used as part of the lexicon-building process in the FrameNet project, an NSF-funded project aimed at creating a lexical database based on the principles of Frame Semantics (Fillmore 1982).

1 Infrastructure

1.1 Tools

We are using the 100 mio. word British National Corpus, with the following corpus query tools:

- CQP (Corpus Query Processor, Christ (1994)), a general corpus query processor for complex queries with any number and combination of annotated information types, including part-of-speech tags, morphosyntactic tags, lemmas and sentence boundaries.
- A macroprocessor for use with CQP that allows the user to specify which subcorpora are to be created for a given lemma.

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). The queries essentially simulate a chunk parser, using a regular grammar.

1.2 Coverage

A list of the verb frames that are currently searchable is given in figure 1 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).

In our queries for nouns and adjectives as targets, we are able to extract prepositional, clausal, infinitival, and gerundial complements. In addition, the tool accomodates searches for compounds and for possessor phrases (*my neighbor's addiction to cake, my milk allergy*). Even though these categories are not tied to the syntactic subcategorization frames of the target lemmas, they often instantiate semantic arguments, or, more specifically, Frame elements (Fillmore 1982, Baker et al. forthcoming).

1.3 Method

1.3.1 Overview

We start by creating a subcorpus containing all concordance lines for a given lemma. We call this subcorpus a lemma-subcorpus. The extraction of smaller subcorpora from the lemma subcorpus then proceeds in two stages. During the first stage, syntactic patterns involving 'displaced' arguments (i.e. 'left isolation' or 'movement' phenomena) are extracted, such as passives, tough movement and constructions involving WH-extraction. The result of this procedure is a set of subcorpora that are homogeneous with respect to major constituent order. Following this, the remainder of the lemma-subcorpus is partitioned into subcorpora based on the subcategorization properties of the lemma in question.

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

figure 1: searchable complement types for verbs

1.3.2 Search strategies: positive and negative queries

For the extraction of certain subcategorization patterns, it is not necessary to simulate a parse of all of the constituents. Where an explicit context cue exists, a partial parse suffices. For example, the query given in figure 2 below is used to find [_ NP VPing] patterns (e.g. *kept them laughing*). Note that the query does not positively identify a noun phrase in the position following the target verb.

encoding	description	example
[$search_by]	target lemma	kept
[pos!="V.*|CJC|CJS|CJT|PRF|PRP|PUN"] {1,5}		them
[pos ="VVG|VBG|VDG|VHG"]	gerund	coming
within s;	within a sentence	

figure 2: A query for [_NP VPing]

1.3.3 Searches driven by subcategorization frames

Applying queries like the one for [NP VPing] "blindly", i.e. in the absence of any information on the target lemma, would produce many false hits, since the query also matches gerunds that are not subcategorized. However, the information that the target verb subcategorizes for a gerund dramatically reduces the number of such errors.

The same mechanism is used for addressing the problems associated with prepositional phrase attachment. The general principle is that prepositional phrases in certain contexts are considered to be embedded in a preceding noun phrase , unless the user specifies that a given preposition is subcategorized for by the target lemma. For example, the *of*-phrase in a sequence Verb - NP - *of* - NP is interpreted as part of the first NP (as in *met the president of the company*), unless we are dealing with a verb that has a [_NP PP*of*] subcategorization frame, e.g. *cured the president of his asthma.*

1.3.4 Cascading queries

The result of each query is subtracted from the lemma subcorpus and the remainder submitted to the next set of queries. As a result, earlier queries pre-empt later queries. For example, concordance lines matching the queries for passives, e.g. *he was cured of his asthma* are filtered out early on in the process, so as to avoid getting matched by the queries dealing with (active intransitive) verb + prepositional phrase complements, such as *he boasted of his achievements.*

Another example of this type of preemption concerns the interaction of the query for ditransitive frames (*brought her flowers*) with later queries for NP complements. A proper name immediately followed by another proper name (e.g. *Henry James*) is interpreted as a single noun phrase except when the target lemma subcategorizes for a ditransitive frame[1]. An analogous strategy is used for identifying noun compounds. For ditransitives, strings that represent two consecutive noun phrases are queried for first. Note that this method crucially relies on the fact that the subcategorization properties of the target lemma are given as the input to the query process.

2 Examples

2.1 NPs

An example of a complex query expression of the kind we are using is given in figure 3. The expression matches noun phrases like "the three kittens", "poor Mr. Smith", "all three", "blue flowers", "an unusually large hat", etc.

```
([pos = "AT0|CRD|DPS|DT0|ORD|CJT-DT0|CRD-PNI"]* [pos = "AV0|AJ0-AV0"]* [pos = "AJ0|AJC|AJS|AJ0-AV0|AJ0-NN1|AJ0-VVG"]* [pos="NN0|NN1|NN2|AJ0-NN1|NN1-NP0|NN1-VVB|NN1VVG|NN2-VVZ"])
|([pos = "AT0|CRD|DPS|DT0|ORD|CJT-DT0|CRD-PNI"]+ [pos = "AV0|AJ0-AV0"]* [pos = "AJ0|AJC|AJS|AJ0-AV0|AJ0-NN1|AJ0-VVG"]+)| ([pos = "AT0|CRD|DPS|DT0|ORD|CJT-DT0|CRD-PNI"]* [pos = "AV0|AJ0-AV0"]* [pos = "AJ0|AJC|AJS|AJ0-AV0|AJ0-NN1|AJ0-VVG"]* [pos = "NP0|NN1-NP0"]+)|([pos = "AJ0|AJC|AJS"]* [pos = "PNI|PNP|PNX|CRD-PNI"])
```

figure 3. A regular expression matching NPs

2.2 Coordinated passives

As an example of a query matching a 'movement' structure, consider the query for coordinated passives, given in figure 3 below. The leftmost column gives the query expression itself, while the other columns show

[1] Inevitably, this strategy fails in some cases, such as "I'm reading Henry James now" (vs. "I read Henry stories."

concordance lines found by this query. The target lemma is the verb *cure*:

[(lemma = "be|being|get") & (word != "s") & (pos != "NN1|NN2")]	[(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}	[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}	[lemma = "cure" & pos="VVB|VVD|VVG|VVI|VVN|VVZ|AJ0-VVN|AJ0|VVD| AJ0-VVG|NN1-VVB |NN1-VVG|NN2-VV Z|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"]
been	ameliorated	but not	cured
be	prevented	or largely	cured
be	managed	and often	cured
be	treated	for it and	cured

figure 4. A query for passives in coordination structures

3 The macroprocessor

A macroprocessor has been developed[2] that allows the user to specify in a simple input file which subcorpora are to be created for a given lemma.
The macroprocessor reads the the number of matches for each subcategorization pattern into an output file. A sample input file for the lemma *insist* is given in figure 5 below.

```
        lemma:      insist
        ------------------------------------------------
                        CQP Search Definition
        search_by:    lemma                  save_text:    no
        POS:          verb                   save_binary:  yes
np:             (y/n) n                      pp:           (_list_ prepositions) on
np_np:          (y/n) n                      ping:         (_list_ prepositions) on
np_ap:          (y/n) n                      pwh:          (_list_ prepositions) on
np_pp:          (_list_ prepositions) none   particle:     (y/n) n
np_ping:        (_list_ prepositions) none   np_particle:  (y/n) n
np_pwh:         (_list_ prepositions) none   particle_pp:  (y/n) n
np_vpto:        (y/n) n                      particle_wh:  (y'n) n
np_vping:       (y/n) n                      ap:           (y/n) n
np_sfin:        (y/n) n                      directquote:  (y/n) y
np_wh:          (y/n) n                      sfin          (y/n) y
np_sbrst:       (y/n) n                      sbrst:        (y/n) y
```

figure 5 Input form for macroporcessor

4 Output format

The subcorpora can be saved as binary files for further processing in CQP or XKWIC, an interactive corpus query tool (Christ 1994) and as text files. The text files are sorted, usually by the head of the first complement following the target lemma.

5 Limitations of the approach

Our tool relies on subcategorization information as its input. Hence it is not capable of automatically learning subcategorization frames, e.g. ones that are missing in diction-

[2] Our macroprocessor was developed by Collin Baker (ICSI-Berkeley) and Douglas Roland (U of Colorado, Boulder).

aries or omitted in the input file. The tool facilitates the (manual) discovery of evidence for new subcategorization frames, however, as potential complement patterns are saved in separate subcorpora. Indeed, this is one of the ways in which the tool is being used in the context of the FrameNet project.

Some of the technical limitations of the existing tools result from the fact that we are working with an unparsed corpus. Thus, many types of 'null' or 'empty' constituents[3] are not recognized by the queries. Ambiguities in prepositional phrase attachment are another major source of errors. For instance, of the concordance lines supposedly instantiating a [_NP PP*with*] frame for the verb *heal*, several in fact contained embedded PPs (e.g. [_NP], as in *heal [children with asthma]*, rather than [_NP PP*with*], as in *healing [arthritis] [with a crystal ball]*),

Finally, the search results can only be as accurate as the part-of-speech tags and other annotations in the corpus.

7 Future directions

Future versions of the tool will include searches for predicative (vs. attributive) uses for adjectives and nouns. For verbs, the searches will be expanded to cover the entire set of complementation patterns described in the COMLEX syntactic dictionary.

Conclusion

We have presented an overview of a set of tools for extracting corpus lines illustrating subcategorization patterns of nouns, verbs, and adjectives, and for determining the frequency of these patterns. The tools are currently used as part of the FrameNet project. An overview of the whole project can be found at: http://www.icsi.berkeley.edu/~framenet.

Acknowledgements

This work grew out of an extremely enjoyable collaborative effort with Dr. Ulrich Heid of IMS Stuttgart and Dan Jurafsky of the University of Boulder, Colorado. I would like to thank Doug Roland and especially the untiring Collin Baker for their work on the macroprocessor. I would also like to thank the members of the FrameNet project for their comments and suggestions. I thank Judith Eckle-Kohler of IMS-Stuttgart, JB Lowe of ICSI-Berkeley and Dan Jurafsky for comments on an earlier draft of this paper.

[3] Our system is able to identify passive structures, tough-movement, and a number of common left isolation constructions, i.e. constructions involving 'traces' or movement sites.

References

Baker, C. F., Fillmore, C. J. and Lowe, J. B (forthcoming). *The Berkeley FrameNet project.* Proceedings of the 1998 ACL-COLING conference.

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.

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

Francis, W. N. and Kucera, H. (1982) Frequency Analysis of English Usage: Lexicon and Grammar. Houghton Mifflin, Boston, MA.

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

MacDonald, M. C. (ed.) (1997) Lexical Representations and Sentence Processing. Erlbaum Taylor & Francis.

Macleod, C. and Grishman, R. (1995) COMLEX Syntax Reference Manual. Linguistic Data Consortium, U. of Pennsylvania.

Manning, Christopher D. (1993). Automatic Acquisition of a large subcategorization dictionary from corpora. Proceedings of the 31st ACL, pp. 235-242.

Procter, P. (ed.). (1989) *Longman Dictionary of Contemporary English.* Longman, Burnt Mill, Harlow, Essex, England.

Sinclair, J. M. (1987) *Collins Cobuild English Language Dictionary.* Collins, London, England.