Detecting Dependencies between Semantic Verb Subclasses and Subcategorization Frames in Text Corpora

Victor Poznański, Antonio Sanfilippo

SHARP Laboratories of Europe Ltd.
Oxford Science Park, Oxford OX4 4GA
{vp, aps}@prg.oxford.ac.uk

Abstract

We present a method for individuating dependencies between the semantic class of predicates and their associated subcategorization frames, and describe an implementation which allows the acquisition of such dependencies from bracketed texts.

1 Introduction

There is a widespread belief among linguists that a predicate's subcategorization frames are largely determined by its lexical-semantic properties [23, 11, 12]. Consider the domain of movement verbs. Following Talmy [23], these can be semantically classified with reference to the meaning components: MOTION, MANNER, CAUSATION, THEME (MOVING ENTITY), PATH AND REFERENCE LOCATIONS (GOAL, SOURCE). Lexicalization patterns which arise from identifying clusters of such meaning components in verb senses can be systematically related to distinct subcategorization frames.[1] For example, the arguments of a verb expressing *directed caused motion* (e.g. *bring, put, give*) are normally a causative subject (agent), a theme direct object (moving entity) and a directional argument expressing path and reference location (goal), e.g.

(1)	Jackie	will bring	a bottle of retsina	to	the party
	CAUSER		THEME	PATH	GOAL

However, a motion verb which is not amenable to *direct external causation* [13], will typically take a theme subject, with the possible addition of a directional argument, e.g.

(2) The baby crawled (across the room)

Co-occurrence restrictions between meaning components may also preempt subcategorization options; for example, manner of motion verbs in Italian cannot integrate a completed path component and therefore never subcategorize for a directional argument, e.g.

(3)*Carlo ha camminato a casa
Carlo walked home

[1] Following Levin [12] and Sanfilippo [18], we maintain that valency reduction processes (e.g. the causative-inchoative alternation) are semantically governed and thus do not weaken the correlation between verb semantics and subcategorization properties.

These generalizations are important for NLP since they frequently cover large subclasses of lexical items and can be used both to reduce redundancy and elucidate significant aspects of lexical structure. Moreover, a precise characterization of the relation between semantic subclasses and subcategorization properties of verbs can aid lexical disambiguation. For example, the verb *accord* can be used in either one of two senses: *agree* or *give*, e.g.

(4) a The two alibis do not accord
Your alibi does not accord with his
b They accorded him a warm welcome

Accord is intransitive in the *agree* senses shown in (4a), and ditransitive in the *give* sense shown in (4b).

The manual encoding of subcategorization options for each choice of verb subclass in the language is very costly to develop and maintain. This problem can be alleviated by automatically extracting collocational information, e.g. grammar codes, from Machine Readable Dictionaries (MRDs). However, most of these dictionaries are not intended for such processing; their readership rarely require or desire such exhaustive and exacting precision. More specifically, the information available is in most cases compiled manually according to the lexicographer's intuitions rather than (semi-)automatically derived from texts recording actual language use. As a source of lexical information for NLP, MRDs are therefore liable to suffer from omissions, inconsistencies and occasional errors as well as being unable to cope with evolving usage [1, 4, 2, 6]. Ultimately, the maintenance costs involved in redressing such inadequacies are likely to reduce the initial appeal of generating subcategorization lists from MRDs.

In keeping with these observations, we implemented a suite of programs which provide an integrated approach to lexical knowledge acquisition. The programs elicit dependencies between semantic verb classes and their admissible subcategorization frames using machine readable thesauri to assist in semantic tagging of texts.

2 Background

Currently available dictionaries do not provide a sufficiently reliable source of lexical knowlege for NLP systems. This has led an increasing number of researchers to look at text corpora as a source of information [8, 22, 9, 6, 3]. For example, Brent [6] describes a program which retrieves subcategorization frames from untagged text. Brent's approach relies on detecting nominal, clausal and infinitive complements after identification of proper nouns and pronouns using predictions based on GB's Case Filter [16] — e.g. in English, a noun phrase occurs to the immediate left of a tensed verb, or the immediate right of a main verb or preposition. Brent's results are impressive considering that no text preprocessing (e.g. tagging or bracketing) is assumed. However, the number of subcategorization options recognized is minimal,[2] and it is hard to imagine how the approach could be extended to cover the full range of subcategorization possibilities without introducing some form of text preprocessing. Also, the phrasal patterns extracted are too impoverished to infer selectional restrictions as they only contain proper nouns and pronouns.

[2]Brent's program recognizes five subcategorization frames built out of three kinds of constituents: noun phrase, clause, infinitive.

Lexical acquisition of collocational information from preprocessed text is now becoming more popular as tools for analyzing corpora are getting to be more reliable [9]. For example, Basili *et al.* [3] present a method for acquiring sublanguage-specific selectional restrictions from corpora which uses text processing techniques such as morphological tagging and shallow syntactic analysis. Their approach relies on extracting word pairs and triples which represent crucial environments for the acquisition of selectional restrictions (e.g. V_prep_N(*go,to,Boston*)). They then replace words with semantic tags (V_prep_N(PHYSICAL_ACT-to-PLACE)) and compute co-occurrence preferences among them. Semantic tags are crucial for making generalizations about the types of words which can appear in a given context (e.g. as the argument of a verb or preposition). However, Basili *et al.* rely on manual encoding in the assignment of semantic tags; such a practice is bound to become more costly as the text under consideration grows in size and may prove prohibitively expensive with very large corpora. Furthermore, the semantic tags are allowed to vary from domain to domain (e.g. commercial and legal corpora) and are not hierarchically structured. With no consequent notion of subsumption, it might be impossible to identify "families" of tags relating to germane concepts across sublanguages (e.g. PHYSICAL_ACT, ACT; BUILDING, REAL_ESTATES).

3 CorPSE: a Body of Programs for Acquiring Semantically Tagged Subcategorization Frames from Bracketed Texts

In developing CorPSE (Corpus-based Predicate Structure Extractor) we followed Basili *et al.*'s idea of extracting semantically tagged phrasal frames from preprocessed text, but we used the *Longman Lexicon of Contemporary English* (LLOCE [15]) to automate semantic tagging. LLOCE entries are similar to those of learner's dictionaries, but are arranged in a thesaurus-like fashion using semantic codes which provide a linguistically-motivated classification of words. For example, [19] show that the semantic codes of LLOCE are instrumental in identifying members of the six subclasses of psychological predicates described in (5) [12, 11].

(5)

Affect type	Experiencer Subject	Stimulus Subject
Neutral	*experience*	*interest*
Positive	*admire*	*fascinate*
Negative	*fear*	*scare*

As shown in (6), each verb representing a subclass has a code which often provides a uniform characterization of the subclass.

(6)

Code	Group Header	Entries
F1	Relating to feeling	*feel, sense, experience ...*
F140	Admiring and honouring	*admire, respect, look up to ...*
F121	Fear and Dread	*fear, fear for, be frightened ...*
F25	Attracting and interesting	*attract, interest, concern ...*
F26	Attracting and interesting very much	*fascinate, enthrall, enchant ...*
F122	Frighten and panic	*frighten, scare, terrify ...*

Moreover, LLOCE codes are conveniently arranged into a 3-tier hierarchy according to specificity, e.g.

F Feelings, Emotions, Attitudes and Sensations
  F20-F40 Liking and not Liking
    F26 Attracting and Interesting very much
      fascinate, enthrall, enchant, charm, captivate

The bottom layer of the hierarchy contains over 1500 domain-specific tags, the middle layer has 129 tags and the top (most general) layer has 14. Domain-specific tags are always linked to intermediate tags which are, in turn, linked to general tags. Thus we can tag sublanguages using domain-specific semantic codes (as do Basili *et al.*) without generating unrelated sets of such codes.

We assigned semantic tags to *Subcategorization Frame tokens* (SF tokens) extracted from the Penn Treebank [14, 20, 21] to produce *Subcategorization Frame types* (SF types). Each SF type consists of a verb stem associated with one or more semantic tags, and a list of its (non-subject) complements, if any. The head of noun phrase complements were also semantically tagged. We used LLOCE collocational information — grammar codes — to reduce or remove semantic ambiguity arising from multiple assignment of tags to verb and noun stems. The structures below exemplify these three stages.

```
SF token: ((DENY VB)
           (NP (ALIENS NNS))
           (NP (*COMPOUND-NOUN* (STATE NN) (BENEFITS NNS))))

SF type: (("deny" ("C193"-refuse "G127"-reject))
          ((*NP* ("C"-people_and_family))
           (*NP* ("N"-general_and_abstract_terms))))

Disambiguated SF type: (("deny" ("C193"))
                        ((*NP* ("C"))
                         (*NP* ("N"))))
```

3.1 CorPSE's General Functionality

CorPSE is conceptually segmented into 2 parts: a *predicate structure extractor*, and a *semantic processor*. The predicate structure extractor takes bracketed text as input, and outputs SF tokens. The semantic processor converts SF tokens into SF types and disambiguates them.

3.1.1 Extracting SF Tokens

The predicate structure extractor elicits SF tokens from a bracketed input corpus. These tokens are formed from phrasal fragments which correspond to a subcategorization frame, factoring out the most relevant information. In the case of verbs, such fragments correspond to verb phrases where the following simplificatory changes have been applied:

- NP complements have been reduced to the head noun (or head nouns in the case of coordinated NP's or nominal compounds), e.g. `((FACES VBZ) (NP (CHARGES NNS)))`

- PP complements have been reduced to the head preposition plus the head of the complement noun phrase, e.g. `((RIDES VBZ) (PP IN ((VAN NN))))`
- VP complements are reduced to a mention of the VFORM of the head verb, e.g. `((TRY VB) (VP TO))`
- clausal complements are reduced to a mention of the complementizer which introduces them, e.g. `((ARGUED VBD) (SBAR THAT))`

An important step in the extraction of SF tokens is to distinguish passive and active verb phrases. Passives are discriminated by locating a past participle following an auxiliary *be*.

3.1.2 Converting SF Tokens into SF Types

The semantic processor operates on the output of the predicate structure extractor. Inflected words in input SF tokens are first passed through a general purpose morphological analyser [17] and reduced to bare stems suitable for automated dictionary and lexicon searches. The next phase is to supplement SF tokens with semantic tags from LLOCE using the facilities of the ACQUILEX LDB [5, 7] and DCK [17]; LLOCE tags are associated with verb stems and simply replace noun stems.

The resulting SF structures are finally converted into SF types according to the representation system whose syntax is sketched in (7) where: *stem* is the verb stem, *parts* a possibly empty sequence of particles associated with the verb stem, {A ... N } is the set of LLOCE semantic codes, *pform* the head of a prepositional phrase, *compform* the possibly empty complementizer of a clausal complement, and *cat* any category not covered by np-, pp-, sbar- and vp- frames.

(7) SF-type ::= (*stem parts* sem comps)
sem ::= ({A ... N }*)
comps ::= comp*
comp ::= ({ np-frame | pp-frame | sbar-frame | vp-frame | cat-frame })
np-frame ::= (***NP*** sem)
pp-frame ::= (***PP*** *pform* comp)
sbar-frame ::= (***SBAR*** *compform*)
vp-frame ::= (***VP*** *vform*)
cat-frame ::= (***CAT*** *cat*)

3.1.3 Disambiguating SF Types

The disambiguation module of the semantic processor *coalesces* SF types, and reduces semantic tags when verb stems have several codes.

Coalescing merges SF types with isomorphic structure and identical verb stem, combining the semantic codes of NP-frames, e.g.

```
(("accord" ("D101" "N226"))
   ((*PP* TO (*NP* ("C")))))
(("accord" ("D101" "N226"))
   ((*PP* TO (*NP* ("G")))))
(("accord" ("D101" "N226"))
   ((*PP* TO (*NP* ("C" "G")))))
```

⇒

```
(("accord" ("D101" "N226"))
   ((*PP* TO (*NP* ("C" "G")))))
```

This process can be performed in linear time when the input is lexicographically sorted.

We employ two tag reduction methods. The first eliminates equivalent tags, the second applies syntactico-semantic restrictions using LLOCE grammar codes.

More than one LLOCE code can apply to a particular entry. Under these circumstances, it may be possible to ignore one or more of them. For example, the verb *function* is assigned two distinct codes in LLOCE: I28 *functioning and serving*, and N123 *functioning and performing*. Although I- and N-codes may in principle differ considerably, in this case they are very similar; indeed, the entries for the two codes are identical. This identity can be automatically inferred from the *descriptor* associated with semantic codes in the LLOCE index. For example, for a verb such as *accord* where each semantic code is related to a distinct entry, the index gives two separate descriptors:

accord ...
give *v* D101
agree *v* N226

By contrast, different codes related to the same entry are associated with the same descriptor, as shown for the entry *function* below.

function ...
work *v* I28, N123

We exploit the correlation between descriptors and semantic codes in the LLOCE index, reducing multiple codes indexed by the same descriptor to just one. More precisely, the reduction involves substitution of all codes having equal descriptors with a new code which represents the logical conjunction of the substituted codes. This is shown in (8) where "I28+N123" is defined as the intersection of "I28" and "N123" in the LLOCE hierarchy of semantics codes as indicated in (9).

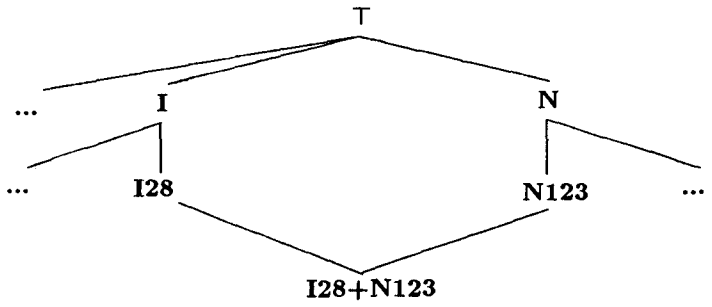
(8)

```
(("function" ("I28" "N123"))
   ((*PP* LIKE (*NP* ("C")))))
```

⇒

```
(("function" ("I28+N123"))
   ((*PP* LIKE (*NP* ("C")))))
```

(9)



The second means for disambiguating SF types consists of filtering out the codes of verb stems which are incompatible with the type of subcategorization frame in which they occur. This is done by using collocational information provided in LLOCE. For example, the verb *deny* is assigned two distinct semantic codes which cannot be reduced to one as they have different descriptors:

deny ...
refuse *v* C193
reject *v* G127

The difference in semantic code entails distinct subcategorization options: *deny* can have a ditransitive subcategorization frame only in the *refuse* sense, e.g.

(10) Republican senator David Lock's bill would permanently $\left\{ \begin{matrix} \text{deny (refuse)} \\ \text{*deny (reject)} \end{matrix} \right\}$ illegal aliens all State benefits

The codependency between semantic verb class and subcategorization can often be inferred by the grammar code of LLOCE entries. For example, only the entry for the *refuse* sense of *deny* in LLOCE includes the grammar code D1 which signals a ditransitive subcategorization frame:

(11) **C193** *verbs*: **not letting or allowing**
deny [D1;T1] ...

G127 *verbs*: **rejecting** ...
deny 1 [T1,4,5;V3] ...**2** [T1] ...

Semantic codes which are incompatible with the SF types in which they occur, such as G127 in (12), can thus be filtered out by enforcing constraints between SF type complement structures and LLOCE grammar codes.

(12)
```
(("deny"("C193" "G127"))
   ((*NP* ("C"))
    (*NP* ("N"))))
```

To automate this process, we first form a set *GC* of compatible grammar codes for each choice of complement structure in SF types. For example, the set of compatible grammar codes *GC* for any SF type with two noun phrase complements is restricted to the singleton set {D1}, e.g.

(13)
```
((stem sem)              ⇒  GC = {D1}
   ((*NP* sem )
    (*NP* sem )))
```

A set of 2-tuples of the form (verb-stem-semantic-code, grammar-codes) is formed by noting the LLOCE grammar codes for each semantic code that could apply to the verb stem. If the grammar codes of any 2-tuple have no intersection with the grammatical restrictions *GC*, we conclude that the associated verb-stem-semantic code is not possible.[3] For example, C193 in the SF type for *deny* in (13) is paired up with the grammar codes {D1;T1} and G127 with {T1,4,5;V3} according to the LLOCE entries for *deny* shown in

[3]This procedure is only effective if the corpus subcategorization information is equally or more precise than the dictionary information. For our corpus, it proved to be the case.

(12). The constraints in (14) would thus license automatic removal of semantic code G127 from the SF type for ditransitive *deny* as shown in (15).

(14)
```
(("deny"  ((C193, {D1, T1})                    ⇒   GC = {D1}
           (G127, {T1, T4, T5, V3}))  )
   ((*NP* ("C"))
    (*NP* ("N"))))
```

(15)
```
(("deny" ("C193"-refuse "G127"-reject))    ⇒  (("deny" ("C193"))
   ((*NP* ("C"))                                  ((*NP* ("C"))
    (*NP* ("N"))))                                 (*NP* ("N"))))
```

It may appear that there is a certain circularity in our work. We use grammar codes to help disambiguate SF types, but it might be argued that the corpus could not have been bracketed without some prior grammatical information: subcategorisation frames. This picture is inaccurate because our SF types provide collocational information which is not in LLOCE. For example, the SF type shown in (16a) captures the use of *link* in (16b); this subcategorization cannot be inferred from the LLOCE entry where no PP headed by *to* is mentioned.

(16) a
```
(("link" NIL ("N"))
   ((*NP* ("C"))
    (*PP* TO (*NP* ("B" "N")))))
```

b The arrest warrant issued in Florida links the attorney to a government probe of the Medhyin drug cartel ...

Indeed, another possible use for our system would be to provide feedback to an on-line dictionary. We also provide a partial indication of selectional restrictions, i.e. the semantic tags of NP complements. Furthermore, text can be bracketed using techniques such as stochastic and semi-automatic parsing which need not rely on exhaustive lists of subcategorisations.

4 Using CorPSE: Emerging Trends and Current Limitations

In testing CorPSE, our main objectives were:

- to assess the functionality of text pre-processing techniques involving automated semantic tagging and lexical disambiguation, and
- to show that such techniques may yield profitable results in capturing regularities in the syntax-semantics interface

In order to do this, we ran CorPSE on a section of the Penn Treebank comprising 576 bracketed sentences from radio transcripts. ¿From these sentences, CorPSE extracted 1335 SF tokens comprising 1245 active VPs and 90 passives. The SF tokens were converted into 817 SF types. The coalescence process reduced the 817 SF types to 583, which are representative of 346 distinct verb stems. The verb stem of 308 of these 583 SF types was semantically ambiguous as it was associated with more than one semantic tag. In some

cases, this ambiguity was appropriate because the semantic codes assigned to the stem were all compatible with the complement structure of their SF type. For example, the verb *call* can occur in either one of two senses, *summon* and *phone*, with no change in subcategorization structure:

(17) a Supper is ready, call the kids
b Call me when you land in Paris

In this case, CorPSE correctly maintains the ambiguity as shown in (18).

(18)
```
(("call" ("G"-summon "M"-phone))
    ((*NP* ("C" "J" "N"))))
```

In other cases, the ambiguity was in need of resolution as some of the verb-stem's semantic codes referred to the same LLOCE entry or were incompatible with the complement structure in the SF type (see §3.1.3). Disambiguation using semantic tag equivalence reduced the ambiguity of 206 types, totally disambiguating 31 stems. Applying collocation restrictions further reduced 38 stems, totally disambiguating 24 of them.

Taking into account that the amount of data processed was too small to use statistical techniques for disambiguation, the results achieved are very promising: we managed to reduce ambiguity in over half the SF types and totally disambiguated 16 percent, thus providing a unique correspondence between semantic verb class and subcategorization frame in 346 cases. Of the remaining 179 SF frames, 106 had verb stems with two semantic codes, 72 had verb stems with 3-5 semantic codes and the verb stem of one SF type had 6. Needless to say, the number of ambiguous SF types is bound to increase as more texts are processed. However, as we accumulate more data, we will be able to apply statistical techniques to reduce lexical ambiguity, e.g. by computing co-occurrence restrictions between the semantic codes of the verb stem and complement heads in SF types.

The table below summarizes some of the results concerning the correlation of semantic codes and subcategorization options obtained by running CorPse on the Penn Treebank fragment. The first column lists the LLOCE semantic codes which are explained in (20). The second column indicates the number of unique subcategorization occurrences for each code. A major difficulty in computing this relation was the presence of certain constituents as arguments that are usually thought of as adjuncts. For example, purpose clauses and time adverbials such as *yesterday, all day, in March, on Friday* had often been bracketed as arguments (i.e. sisters to a V node). Our solution was to filter out inadequately parsed arguments semi-automatically. Certain constituents were automatically filtered from SF types as their status as adjuncts was manifest, e.g. complements introduced by prepositions and complementizers such as *without, as, since* and *because*. Other suspect constituents, such as infinitive VPs which could represent purpose clauses, were processed by direct query. A second problem was the residual ambiguities in SF types mentioned above. These biased the significance of occurrences since one or more codes in an ambiguous SF type could be inconsistent with the subcategorization of the SF type. A measure of the "noise" factor introduced by ambiguous SF types is given in the third column of (19), where ambiguity rate is computed by dividing the number of codes associated with the same complement structure by the number of occurrences of that code with any complement structure. This ambiguity measure allows the significance of the figures in the second column to be assessed. For example, since the occurrences of "E" instances were invariably ambiguous, it is difficult to draw reliable conclusions about

them. Indeed, on referring most of these SF types (e.g. *beat, bolt* and *have*) back to their source texts, the "Food & Drink" connotation proved incorrect. The figures in column 1 were normalised as percentages of the total number of occurrences in order to provide a measure of the statistical significance of the results in the remaining columns. We thus conclude that the results for B, E, H, and I are unlikely to be significant as they occur with low relative frequency and are highly ambiguous. The final three columns quantify the relative frequency of occurrence for VP, SBAR and PP complements in SF types for each semantic code.

(19)

Code	# Occ.	% Ambig	Rel. Freq.	% VP	% SBAR	% PP
A	4	0	1	0	0	0
B	9	44	1	0	0	3
C	72	67	9	15	0	39
D	57	65	7	16	4	44
E	23	83	3	22	0	57
F	42	40	5	10	2	21
G	132	33	17	7	14	28
H	11	82	1	0	0	27
I	27	74	3	4	0	63
J	68	57	9	12	1	35
K	29	69	4	0	0	48
L	33	36	4	21	3	27
M	130	50	16	2	1	52
N	161	44	20	14	4	35

(20)

Code	Explanation
A	Life & Living Things
B	The Body, its Functions & Welfare
C	People & the Family
D	Building, Houses, the Home, Clothes
E	Food, Drink & Farming
F	Feelings, Emotions, Attitudes & Sensations
G	Thought & Communication, Language & Grammar
H	Substances, Materials, Objects & Equipment
I	Arts & Crafts, Science & Technology, Industry & Education
J	Numbers, Measurement, Money & Commerce
K	Entertainment, Sports & Games
L	Space and Time
M	Movement, Location, Travel & Transport
N	General & Abstract Terms

Although the results are not clear-cut, there are some emerging trends worth considering. For example, the low frequency of VP and SBAR complements with code "M" reflects the relatively rare incidence of clausal arguments in the semantics of motion and location verbs. By contrast, the relatively high frequency of PP complements with this code can be related to the semantic propensity of motion and location verbs to take spatial arguments.

The "A" verbs (eg. *create, live* and *murder*) appear to be strongly biased towards taking a direct object complement only. This might be due to the fact that these verbs involve creating, destroying or manipulating life rather than events. Finally, the overwhelmingly high frequency of SBAR complements with "G" verbs is related to the fact that thought and communication verbs typically involve individuals and states of affairs.

We also found interesting results concerning the distribution of subcategorization options among specializations of the same general code. For example, 23 out of 130 occurrences of "M" verbs exhibited an "NP PP" complement structure; 17 of these were found in SF types with codes "M50-M65" which largely characterize verbs of caused directed motion: *Putting and Taking, Pulling & Pushing.* This trend confirms some of the observations discussed in the introduction. It is now premature to report results of this kind more fully since the corpus data used was too small and genre-specific to make more reliable and detailed inferences about the relation between subcategorization and semantic verb subclass. We hope that further work with larger corpora will uncover new patterns and corroborate current correlations which at present can only be regarded as providing suggestive evidence. Other than using substantially larger texts, improvements could also be obtained by enriching SF types, e.g. by adding information about subject constituents.

5 Conclusions

We have provided the building blocks for a system that combines the advantages of free-text processing of corpora with the more organised information found in MRDs, such as semantic tags and collocational information. We have shown how such a system can be used to acquire lexical knowledge in the form of semantically tagged subcategorization frames. These results can assist the automatic construction of lexicons for NLP, semantic tagging for data retrieval from textual databases as well as to help maintain, refine and augment MRDs.

Acknowledgements

Some of the work discussed in this paper was carried out at the Computer Laboratory in Cambridge within the context of the ACQUILEX project. The Penn-Treebank-data used were provided in CD-ROM format by the University of Pennsylvania through the ACL Data Collection Initiative (ACL/DCI CD-ROM I, September 1991). We are indebted to Ian Johnson for helpful comments and encouragement, and to John Beaven and Pete Whitelock for providing feedback on previous versions of this paper. Many thanks also to Ann Copestake and Victor Lesk for their invaluable contribution towards mounting LLOCE on the LDB.

References

[1] Atkins, B., Kegl, J. & Levin, B. (1986) Explicit and Implicit Information in Dictionaries. In *Advances in Lexicology*, Proceedings of the Second Annual Conference of the Centre for the New OED, University of Waterloo, Waterloo, Ontario.

[2] Atkins, B. & Levin, B. (1991) Admitting Impediments. In Zernik, U. (ed.) *Lexical Acquisition: Using On-Line Resources to Build a Lexicon.*, Lawrence Erlbaum Associates, Hillsdale, New Jersey.

[3] Basili, R. and Pazienza, M. T. and Velardi, P. (1992). Computational Lexicography: the Neat Examples and the Odd Exemplars. In *Proc. 3rd Conference on Applied NLP*, Trento, Italy.

[4] Boguraev, B. & Briscoe, T. (1989) Utilising the LDOCE Grammar Codes. In Boguraev, B. & Briscoe, T. (eds.) *Computational Lexicography for Natural Language Processing.* Longman, London.

[5] Boguraev, B., Briscoe, T., Carroll, J. and Copestake, A. (1990) Database Models for Computational Lexicography. In *Proceedings of EURALEX IV*, Málaga, Spain.

[6] Brent, M R. (1991) Automatic Semantic Classification of Verbs from their Syntactic Contexts: An Implemented Classifier for Stativity. In *Proc 29th ACL*, University of California, Berkeley, California.

[7] Carroll, J. (1992). The ACQUILEX Lexical Database System: System Description and User Manual. In *The (Other) Cambridge ACQUILEX Papers*, TR 253, University of Cambridge, Cambridge, UK.

[8] Church, Kenneth and Hanks, Patrick (1989) Word Association Norms, Mutual Information and Lexicography. *Proc 23rd ACL.* pp. 76 - 83.

[9] Church, K and Gale, W. and Hanks, P. and Hindle, D (1991) Using Statistics in Lexical Analysis. In *Lexical Acquisition*, Zernik, Uri, Ed. Erlbaum, Hillsdale, NJ.

[10] Hindle, D. (1990). Noun Classification from Predicate Argument Structures. In *Proc 28th ACL.* pp. 268 - 275.

[11] Jackendoff, R. (1990) *Semantic Structures.* MIT Press, Cambridge, Mass.

[12] Levin, B. (1989) *Towards a Lexical Organization of English Verbs.* Ms., Dept. of Linguistics, Northwestern University

[13] Levin, B. and Rappaport, M. (1991) The Lexical Semantics of Verbs in Motion: The Perspective from Unaccusativity. To appear in Roca, I. (ed.) *Thematic Structure: Its Role in Grammar*, Foris, Dordrecht.

[14] Liberman, M. and Marcus, M (1992) *Very Large Text Corpora: What You Can Do with Them, and How to Do It.* Tutorial notes, 30th ACL, University of Delaware, Newark, Delaware.

[15] McArthur, T. (1981) *Longman Lexicon of Contemporary English.* Longman, London.

[16] Rouvret, A. and Vergnaud, J R. (1980) Specifying Reference to the Subject. In *Linguistic Enquiry*, 11(1).

[17] Sanfilippo, A (1992) A Morphological Analyser for English and Italian. In *The (Other) Cambridge ACQUILEX Papers*, TR 253, University of Cambridge, Cambridge, UK.

[18] Sanfilippo, A (1993) *Verbal Diathesis: Knowledge Acquisition, Lexicon Construction and Dictionary Compilation.* TR SLE/IT/93-11, Sharp Laboratories of Europe, Oxford, UK.

[19] Sanfilippo, A and Poznański, V. (1992) The Acquisition of Lexical Knowledge from Combined Machine-Readable Dictionary Sources. In *Proceedings of the 3rd Conference on Applied Natural Language Processing*, Trento.

[20] Santorini, B. (1991) Bracketing Guidelines for the Penn Treebank Project. Ms. University of Pennsylvania.

[21] Santorini, B. (1991) Part-of-Spech Tagging Guidelines for the Penn Treebank Project. Ms. University of Pennsylvania.

[22] Smajda, F A. and McKeown, K. R. (1990) Automatically Extracting and Representing Collocations for Language Generation. In *Proc 28th ACL.* pp. 252 – 259.

[23] Talmy, L. Lexicalization Patterns: Semantic Structure in Lexical Form. In Shopen, T. (ed) *Language Typology and Syntactic Description 3. Grammatical Categories and the Lexicon*, CUP, 1985.