

# A domain-independent semantic tagger for the study of meaning associations in English text

George DEMETRIOU  
 Department of Computer Science,  
 University of Sheffield,  
 SHEFFIELD S1 4DP, England  
[g.demetriou@dcs.shef.ac.uk](mailto:g.demetriou@dcs.shef.ac.uk)

Eric Steven ATWELL  
 School of Computer Studies,  
 University of Leeds,  
 LEEDS LS2 9JT, England  
[e.s.atwell@leeds.ac.uk](mailto:e.s.atwell@leeds.ac.uk)

## Abstract

A comparison of semantic tagging with syntactic Part-of-Speech tagging leads us to propose that a domain-independent semantic tagger for English corpora should *not* aim to annotate each word with an atomic ‘sem-tag’, but instead that a semantic tagging should attach to each word a set of semantic primitive attributes or features. These features should include:

- lemma or root, grouping together inflected and derived forms of the same lexical item;
- broad subject categories where applicable;
- selectional restrictions where applicable;
- a meaning definition, stated in terms of a restricted Defining Vocabulary, and processed to remove stoplist-words and repetitions.

A semantic tagger meeting this description can be derived from the Longman Dictionary of Contemporary English, if combined with a robust lemmatiser; allowing automated semantic tagging of large English corpora such a LOB and BNC.

## 1 Introduction: comparison with PoS-tagging

A number of grammatical part-of-speech (PoS) taggers for English are available, and there is broad consensus on domain-independent grammatical categories or tags for English (and other languages); at least in the category and feature types, if not their labels. For illustration, Table 1 (from Atwell et al 2000) shows how an example sentence from the IPSM Corpus (Sutcliffe et al 1996), ‘Select the text you want to protect’, is tagged according to several alternative PoS-tagging schemes and vertically aligned. There is variation in the underlying nomenclature, for example POW uses **M** for **main** verb, **H** for noun **head** of noun-phrase; but there is clearly consensus on the broad syntactic categories verb, article, noun, pronoun, particle, punctuation. The main substantive difference between PoS-tagsets is the degree of delicacy, the range of subcategory features added to the tags. For example, Brown and LOB tagging schemes give the same tag, **VB**, to each of the three verb ‘select’, ‘want’, and ‘protect’; whereas the ICE PoS-tagging scheme distinguishes between **imperative**, **present**, and **infinitive** features for verbs.

	Brown	ICE	LLC	LOB	PARTS	POW	SEC	UPenn
select	VB	V(montr,imp)	VA+0	VB	verb	M	VB	VB
the	AT	ART(def)	TA	ATI	art	DD	ATI	DT
text	NN	N(com,sing)	NC	NN	noun	H	NN	NN
you	PPSS	PRON(pers)	RC	PP2	pron	HP	PP2	PRP
want	VB	V(montr,pres)	VA+0	VB	verb	M	VB	VBP
to	TO	PRTCL(to)	PD	TO	verb	I	TO	TO
protect	VB	V(montr,infin)	VA+0	VB	verb	M	VB	VB
.	.	PUNC(per)	.	.	.	.	.	.

**Table 1.** An example sentence tagged according to eight rival PoS-tagging schemes

There is no such consensus for lexical semantics: there is no agreed semantic tagset. The Expert Advisory Group on Language Engineering Standards has proposed standards for syntactic annotation (EAGLES 1996), but not for domain-independent semantic annotation. Current systems are application-specific, for example the semantic tagset of (Wilson and Rayson 1993) was tailored to applications in market research and advertising; and/or not capable of accurately handling unconstrained English.

## 2 What to require of semantic tagging

Let us consider what properties to expect of a domain-independent semantic tagset; what should a semantically-tagged corpus look like?

### 2.1 Grouping words by semantic behaviour

One obvious analogy to draw from PoS-tagging is that tags should group together words with the same semantic combinational behaviour: words that “mean the same thing” should have the same tag. There is broad consensus on a small number (c10) of PoS categories, where all words in a category share the same (more or less) syntactic combinational behaviour. However, there is no obvious equivalent set of ten or so semantic categories. Roget’s Thesaurus offers 1000 semantic categories; but arguably this categorisation is too fine-grained, and the underlying classification system has not won universal acceptance by (computational) linguists.

### 2.2 Grouping words by lemma

One semantic categorisation principle that has general acceptance is that different inflected and derived forms of a lexical item share the same root meaning. So, a semantic tagger needs a lemmatiser to find the root or lemma for each input word; and the lemmas themselves could form (part of) the semantic tags.

### 2.3 Capturing semantic combinational properties

PoS-tag categories are meant to capture or characterise syntactic combinational properties of words; it follows that (domain-independent) semantic tags should encapsulate semantic combinational properties of words. Grammatical tag combination patterns are local: the main constraints on a PoS-tag are its immediate

neighbours, allowing most PoS-tag associations to be captured in (something like) a bigram model. In contrast, words far apart can have meaning associations, and one word can have meaning associations with many other words in its context; so a simple bigram or n-gram model is too simplistic and not appropriate for semantic tagging.

### 2.4 A tag as a bundle of semantic features

A grammatical tag is usually thought of as an atomic symbol, particularly in n-gram models; for example, the LOB PoS-tagset has 134 tags, so the PoStag-bigram model used in the LOB corpus tagging program (Leech et al 1983, Johansson et al 1986) used a bigram matrix of 134\*134 tag-pair probabilities. However, underlying most tags is a combination of syntactic features. For example, the PoS-tag NN (as used in LOB and Brown tagged corpora) or N(**com,sing**) (used in the ICE corpora) may be treated as a single category or state in Markov models, but is actually a bundle of three syntactic features : singular + common + noun. Some features cut across broad PoS-categories; for example, singular/plural applies to verbs, nouns, pronouns, determiners (and possibly to other categories in other languages).

We need to recognise that a "semantic tag" need not be atomic, but can be a set of semantic primitive features; and there is no reason why the semantic tag of a word should be restricted to just two or three semantic primitive features. This bundle of features should include the root or lemma (see subsection 2.2) and other semantic primitive features which capture meaning and semantic links to other words.

### 2.5 Dictionary meaning definitions

If we accept a semantic tag can be a bundle of semantic primitive features, then a dictionary meaning definition can serve as a semantic tagging. A Machine Readable Dictionary (MRD) such as the Longman Dictionary of Contemporary English (LDOCE) can be converted into a domain-independent lexical knowledge base, used to assign a set of semantic features to each word in Corpus texts. Semantic associations between words can be measured in terms of overlap of these features, allowing long-distance semantic associations to be explored.

### 3 Deriving a semantic tagger from a MRD.

Natural Language Understanding (NLU) researchers started using Machine Readable Dictionaries (MRDs) in the mid 80s (see Lesk 1986) in the hope that online dictionaries might provide a way out of the knowledge acquisition bottleneck. In relation to NLU, examples of MRD exploitation include research in word sense disambiguation (Lesk 1986, Guthrie et al 1991, Cowie et al 1992, Demetriou 1993, Bruce and Wiebe 1994), knowledge acquisition and organisation (Binot and Jensen 1987, Calzolari and Picchi 1988, Wilks et al 1989, Guo 1995), information retrieval (Wallis 1993), information extraction (Cowie et al 1993), text coherence (Kozima and Furugori 1993), meaning associations and speech recognition (Demetriou 1997).

Using a dictionary as a knowledge source for formal lexical semantics is an attractive option since the word meanings in standard dictionaries represent sense distinctions made by professional lexicographers. The provision of on-line dictionaries and text corpora offers the possibility of enormous savings in time and human resources for constructing large scale knowledge bases. The problem, however, has changed from one of how to construct knowledge to that of knowledge utilisation i.e. how to make the available knowledge really useful and efficient for large-scale NLP applications. If a formal lexical semantic model can be extracted from a MRD, this can be used in a domain-independent semantic tagger.

### 4 The LDOCE as a Knowledge Base

The machine readable file of LDOCE is particularly appropriate as a MRD source of semantic information. LDOCE is well-known in lexicography for having employed a restricted set of words, the Longman Defining Vocabulary (LDV), in all word-sense definitions. The LDV is effectively a set of semantic primitives from which all other meanings can be constructed. The following section discusses how the semantic knowledge was "extracted" from the machine readable file and was transformed to a usable knowledge base.

### 4.1 From MRD to LKB

The online dictionary was passed through the following pre-processing stages:

(I) the lisp online version was filtered to remove all typesetting codes. For example, the entry for "doctor" now becomes (the sign "#" is the definition separator):

```
[doctor/1= a person holding one of the highest
degrees given by a university (such as a PHD , DSC ,
DLITT , etc.) # a person whose profession is to
attend to sick people (or animals) # [infml] a person
who repairs the stated things ; REPAIR MAN #
[AmE] DENTIST # [BrE infml] being treated by a
doctor (for) #
doctor/2= to give medical treatment to # to repair #
[derog] to change for some purpose # [derog] to
change in a dishonest way # [euph] to make (esp . an
animal) unable to breed ; NEUTER/3]
```

(II) all different senses of a word were joined and duplicate semantic primitives (i.e. words appearing more than once in this set) were eliminated (the ordering of the words is not important any more in the definition). Abbreviations such as "derog", "euph", "esp." etc. were also removed. All distinct wordforms in the definitions were conflated to their roots or basic concepts, for example, "repairs" to "repair", "treated" to "treat" etc.; a stemmer was developed with the use of a lemmatisation lexicon of about 95,000 words and a number of stemming rules to strip off a set of 112 suffixes and 74 affixes of the words. The actual conflation achieved was about 62% (more on the development of the stemmer and the lemmatisation lexicon can be found in later sections). A stopword list of 32 very common function words was also eliminated from the database as it was decided that they should not take in the semantic overlap; these words were also excluded as headwords in the database. The entry of "doctor" then appears as (the ordering of the words is not important any more):

```
[doctor # DLITT DSC PhD animal attend breed
change degree dentist dishonest doctor give high hold
make man medical neuter people person profession
purpose repair sick some state such thing treat
treatment unable university way who whose]
```

Total number of headwords (concepts)	35,926
Total number of words in definitions	413,377
Total number of distinct attributes (including headwords)	30,955
Average length of meaning definitions	11.5
Minimum length	1
Maximum length	231
Standard deviation of length	11.1

**Table 2:** Lexical Knowledge Base statistics

Some statistics about the transformed database, which has now become a Lexical Knowledge Base (LKB), are given in Table 2. ‘length of meaning definitions’ equates to the number of semantic primitive features in semantic tag; we see that, whereas a PoS-tag may typically be a combination of 2 or 3 syntactic primitive features, a semantic tag based on the LDOCE LKB is typically a bundle of 11 or 12 semantic primitive features. The actual number depends crucially on what is excluded in the stop list; the ‘doctor’ example above would have been shorter if our stoplist had included [who, whose, such, some].

It is important to note that although about 2220 LDV primitives were principally used to define the senses of the words in the LDOCE, the number of primitives or attributes in our LKB is much higher than that because of cross-reference words in sense definitions and the fact that in our database the headwords used can be defining primitives of the concepts they represent i.e. ‘doctor’ is the lemma of the concept ‘doctor’ but it also used in the definition of doctor’ itself. This effectively increases the size of the defining vocabulary to nearer the size of the dictionary headword list.

#### 4.2 Other forms of meaning relatedness in LDOCE

Apart from meaning definitions, the LDOCE database provides information about two more kinds of semantic relatedness. The first is the so called ‘selection restrictions’ or ‘preferences’ in the linguistic literature. The second is information about the specific discourse the meanings of the words refer to. Both are explained in the following two subsections.

##### 4.2.1 Selection preferences

These are used to specify semantic restrictions on either (a) the subject or object noun of a verb and (b) the noun modified by an adjective and

are realised via a semantic hierarchy of nouns. There are 35 such restrictions; from those only 33 were used in our experiments (codes 1,2 and Z were excluded, as too general). These markers (called ‘box codes’ in LDOCE terminology) encode semantic information at a more abstract level than that of defining primitives in sense definitions.

For example, the meaning of "emissary" (= *a person who is sent with an official message, often secret, or who is sent to do special work, often unpleasant*) includes the "human" (b\_H) attribute i.e.

[emissary = b\_H message official often person secret send special unpleasant who work]

Here ‘b\_’ indicates a ‘box’ code and ‘H’ specifies the human attribute. The same code also appears in the meanings of "rational" and "think" indicating the relatedness of the meanings of those words with a noun with human characteristics.

##### 4.2.2 Thematic or subject categories

The specific discourse or domain that a sense of a word can occur in is encoded by a subject coding scheme in LDOCE. These codes have been compiled to specify subdivisions of the senses of the words in the dictionary according to thematic or subject categories. For example, "bronchitis" has a thematic relation to "medicine & biology" (main subject) and also to "histology" (detailed subject) i.e.

[bronchitis # b\_T b\_X s\_MDZH tube bronchial inflammation illness]

Here ‘s\_’ indicates a subject code, ‘MD’ specifies medicine, Z is used to specify subdivisions and ‘H’ is for histology. There are 125 main subject areas and 212 subdivisions below them.

## 5 Results on vocabulary and text coverage

For a semantic tagger, the implementation of a wide coverage automated word stemming system has been of primary importance because of:

- the need to conflate the words to their root forms in the LDOCE sense definitions in order to improve the semantic pattern matching;
- the need to identify the root form of a word in running text in order to retrieve its semantic information by looking it up in the LKB.

Coverage was evaluated by checking the percentage of words that can be found in the lexicon (so that their roots can be retrieved). There are two kinds of statistics for evaluating the coverage of the lexicon (with or without the application of the stemmer):

- one that uses the number of distinct wordforms in text ("vocabulary type" coverage). This answers the question *"how many of the different word types in language are covered by the system?"*; vocabulary type coverage usually gets low percentages when the text is large (65% is typical).
- one that uses the total number of words in text (the probability of finding a root for a word token - "real text token" coverage); this answers the question *"how many of the word-tokens in a text are expected to be covered by the system?"*; real text token coverage gets better results than vocabulary type coverage

since a lexicon should be able to handle the most frequent words correctly.

The term 'unstemmed words' will be used for those words that were not found in the lemmatisation lexicon either before or after the application of the stemmers. When only the lemmatisation lexicon and lexical lookup is used, the coverage will be referred as *'Lexicon-only'*. When the lexicon is used in conjunction with the stemmer, this will be referred as *'Lexicon + stemmer'* coverage.

During the compilation of the lexicon the British National Corpus (BNC) was not available. For testing and evaluation the LOB corpus (1 million words) was used and the lexicon (with or without the stemmer) was found quite efficient in providing the root forms of the words (see Table 3). At that stage, the lexicon was found more than sufficient for our experiments and no further enhancements were needed. When the BNC (100 million words) became available the lexicon was tested using the BNC wordlist.

As can be seen from Table 3 the results are still quite respectable although not as good as those for the LOB corpus (but one has to take into account that the BNC is two orders of magnitude larger than LOB). Only word strings that consisted of alphabetical characters (plus hyphens) were used for these tests. Separate coverage statistics are given when the proper names are excluded from consideration. This is so because proper names are difficult to model using standard lexical resources

Lemmatisation Lexicon (95,000 words)					
		Vocabulary type coverage (%)		Real text token coverage (%)	
		Lexicon only	Lexicon + stemmer	Lexicon Only	Lexicon + stemmer
	All words	66.2	81.6	95.71	96.96
<b>LOB</b>	Excluding proper names	84.4	96.4	98.85	99.79
	All words	24.3	55.2	92.91	95.84
<b>BNC</b>	Excluding proper names	43.4	58.0	95.24	97.77

**Table 3:** Lemmatisation and stemming coverage results

and they are usually dependent on the domain of use. When the proper names were excluded from consideration, the performance of the system (lexicon + stemmer) went up to about 98% for the BNC and nearly 100% percent for the LOB (real text token coverage). This is a very respectable result if we take into account that no extra manual additions were made to the lexicon.

The results indicate that the vocabulary type coverage is very good for small corpora such as LOB but not equally good for large corpora such as the BNC. When the stemmer is used we can see more than 100% improvement in the vocabulary type coverage for all word types in the BNC and more than 33% when the proper names are excluded. The above results are generally better than those of other lemmatisers (although direct comparisons cannot be made because the systems were tried on different corpora).

Subsequent experiments estimated the coverage of the LDOCE on language by limiting the lookup operation to those root forms in the lemmatisation lexicon that are included in LDOCE. The results are in Table 4.

These results indicate that a lexicon based on LDOCE provides good coverage. A bigger lexicon would not contribute much to finding the roots of the words in real text. The results show minimal improvements in real text token coverage (i.e. less than 1% - note, however, that the actual error is reduced by about a quarter). This is good news for those who, for one reason or another, still want to use the 1978 version of LDOCE (note, however, that this version of LDOCE is

about 20 years old and a new dictionary should provide better type coverage).

## 6. Conclusion

We propose that a domain-independent semantic tagger for English corpora should *not* aim to annotate each word with an atomic Sem-tag, but instead that a semantic tagging should attach to each word a set of semantic primitive attributes or features. These features should include

- the lemma or root, grouping together inflected and derived forms of the same lexical item;
- broad thematic categories or subject codes, where available;
- selectional restrictions or box codes, where available;
- a meaning definition, stated in terms of a restricted Defining Vocabulary, and processed to remove stoplist-words and repetitions.

A semantic tagger meeting this description can be derived from the Longman Dictionary of Contemporary English, if combined with a robust lemmatiser. The lemmatiser described above meets this requirement: the high text token coverage allows automated semantic tagging of large English corpora such a LOB and BNC. These semantic tags capture meaning association between words independently of domain or application.

		Lexicon: LDOCE entries only (80,000 entries)			
		Vocabulary type coverage (%)		Real text token coverage (%)	
		Lexicon only	Lexicon + stemmer	Lexicon only	Lexicon + stemmer
<b>LOB</b>	All words	65.2	81.1	95.57	96.90
	Excluding proper names	83.6	96.0	98.74	99.77
<b>BNC</b>	All words	23.3	54.3	92.58	95.60
	Excluding proper names	42.6	56.8	95.05	97.67

**Table 4:** Lemmatisation and stemming coverage results for words in LDOCE LKB

## Acknowledgements

Our thanks go to Longman for allowing access to the LDOCE Machine Readable version for research.

## References

- Atwell, Eric, George Demetriou, John Hughes, Amanda Schiffrin, Clive Souter, and Sean Wilcock. 2000. A comparative evaluation of modern English corpus grammatical annotation schemes. *ICAME Journal* 24, 7-23.
- Binot J.-L. and Jensen K. 1987. A Semantic Expert Using an Online Standard Dictionary. In: *Proceedings of IJCAI-87*.
- Bruce R. and Wiebe J. 1994. Word-sense Disambiguation Using Decomposable Models. In: *Proceedings of ACL-94*.
- Calzolari N and Picchi E. 1988. Acquisition of semantic information from an on-line dictionary. In: *Proceedings of COLING-88*, 87-91.
- Cowie J, Guthrie J and Guthrie L. 1992. Lexical Disambiguation Using Simulated Annealing. In: *Proceedings of COLING-92*, 359-365.
- Cowie J, Wakao T., Guthrie L. and Jin W. 1993. The Diderot Information Extraction System. In: *Proceedings of the PACLING-93 Pacific Association for Computational Linguistics Conference*, 23-32.
- Demetriou, George. 1993. Lexical Disambiguation Using Constraint Handling In Prolog. In: *Proceedings of the European Chapter of the Conference of the Association for Computational Linguistics (EACL)*, Utrecht, Holland, 431-436.
- Demetriou, George. 1997. *Lexical semantic information processing for large vocabulary human-computer speech communication*. PhD thesis, School of Computer Studies, Leeds University.
- EAGLES (1996), WWW site for European Advisory Group on Language Engineering Standards, <http://www.ilc.pi.cnr.it/EAGLES96/home.html>
- Guo C. M. 1995. *Machine Tractable Dictionaries: Design and Construction*, Ablex, New Jersey.
- Guthrie J., Guthrie L, Y. Wilks and H. Aidinejad. 1991. Subject dependent co-occurrence and word sense disambiguation. In: *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics* 146-153.
- Johansson, Stig, Eric Atwell, Roger Garside and Geoffrey Leech. 1986. *The Tagged LOB corpus: users' manual*. Bergen University, Norway: ICAME, The Norwegian Computing Centre for the Humanities. Available from <http://www.hit.uib.no/icame/lobman/lob-cont.html>
- Kozima H. and Furugori T. 1993. Similarity between Words Computed by Spreading Activation on an English Dictionary. In: *Proceedings of EACL' 93* 232-239.
- Leech, Geoffrey, Roger Garside and Eric Atwell. 1983. The automatic grammatical tagging of the LOB corpus. *ICAME Journal* 7:13-33.
- Lesk M. 1986. Automatic sense disambiguation using MRDs: how to tell a pine cone from an ice cream cone. In: *Proceedings of the 1986 SIGDOC*, 24-26.
- Sutcliffe, Richard, Heinz-Detlev Koch and Annette McElligott (eds.). 1996. *Industrial parsing of software manuals*. Amsterdam: Rodopi.
- Wallis P. 1993. *Analysing Text using an On-Line Dictionary: Morphology and Compound Words*, RMIT Comp. Science Technical Report, available at ftp site: phobos.kbs.citri.edu.au
- Wilks Y., Fass D., Guo C. M., McDonald J., Plate T. and Slator B. 1989. A tractable machine dictionary as a resource for computational semantics. In: *Computational Lexicography for Natural Language Processing*, B. Boguraev and T. Briscoe (eds), Longman.
- Wilson A. and P. Rayson. 1993. Automatic content analysis of spoken discourse. In: *Corpus-Based Computational Linguistics*, C. Souter and E. Atwell (eds), Rodopi Press, 215-222.