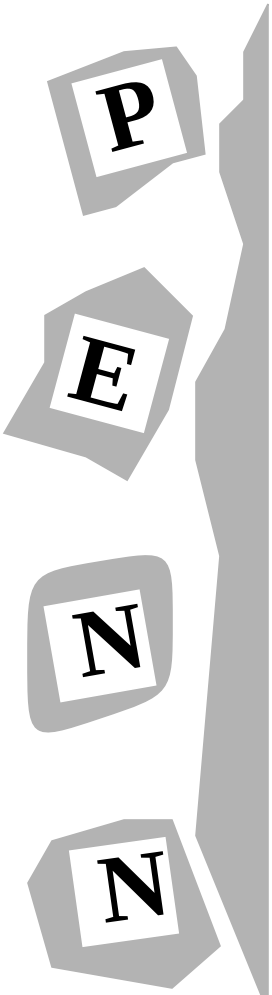
The Institute For Research In Cognitive Science

A Freely Available Syntactic Lexicon for English

by

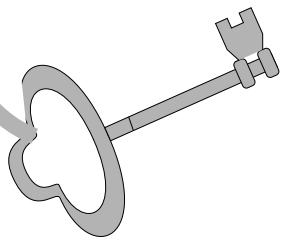
Dania Egedi
IRCS
Patrick Martin
SRA



University of Pennsylvania
3401 Walnut Street, Suite 400C
Philadelphia, PA 19104-6228

April 1995
(originally published in August 1994)

Site of the NSF Science and Technology Center for
Research in Cognitive Science



University of Pennsylvania
Founded by Benjamin Franklin in 1740

IRCS Report 95-11

Appears in the *Proceedings of the International Workshop on Sharable Natural Language Resources*, Nara, Japan, August 1994, pp. 123-130

A Freely Available Syntactic Lexicon for English

Dania Egedi and Patrick Martin*
Institute for Research in Cognitive Science
University of Pennsylvania
Philadelphia, PA 19104-6228, USA

{egedi,martin}@unagi.cis.upenn.edu

Abstract

This paper presents a syntactic lexicon for English that was originally derived from the Oxford Advanced Learner's Dictionary and the Oxford Dictionary of Current Idiomatic English, and then modified and augmented by hand. There are more than 37,000 syntactic entries from all 8 parts of speech. An X-windows based tool is available for maintaining the lexicon and performing searches. C and Lisp hooks are also available so that the lexicon can be easily utilized by parsers and other programs.

1 Introduction

One of the central needs of any wide-coverage parser is a large lexicon that contains the syntactic information for various lexical items. The creation of such a lexicon has traditionally been a very large and daunting task and most universities have shied away from it, leaving the creation of wide-coverage parsers to commercial institutions that could afford the time and personnel to devote to the creation of such a lexicon. The release of several machine-readable dictionaries (MRDs) into the public domain has opened new possibilities to grammar developers at research institutions, but the task did not become trivial. The problem of creating large scale lexicons changed from the tiresome, painstaking task of trying to develop individual word lists for various syntactic phenomena to the task of 'simply' extracting the information from the on-line dictionaries. This, however, has not turned out to be as simple or straight-forward as researchers may have hoped. Machine readable dictionaries present numerous problems in terms of errors and inconsistencies in the various components of the lexical entries, making extraction quite difficult. Many researchers abandon the extraction process altogether because it consumes too many scarce resources.

Although a number of researchers have extracted information out of the various dictionaries available, the resulting lexicons have not, in general, been made freely available to the NLP research community. In at least some cases ([Carroll and Grover, 1989], [Guthrie *et al.*, 1993]) this is due to licensing restrictions on the source dictionaries. In response to the related problems of duplication of effort and non-availability of needed lexicons, there are currently several on-going projects to create syntactic lexicons and make them generally available.

- The Proteus Project at New York University is developing the Comlex Syntactic Dictionary from scratch for release as one of the lexical resources in COMLEX (available through the Linguistic Data Consortium) [Macleod *et al.*, 1994].
- The IITLEX project at Illinois Institute of Technology has an on-going project

*Currently at SRA, Arlington, VA, 22201 USA; martinp@sra.com

to extract and release the information in the Collins English Dictionary, along with information from various other word lists that will include both syntactic and semantic information. That system is still under development, however, and currently uses an expensive relational database package, a drawback which they plan to correct. [Conlon, 1994]

The syntactic lexicon described here contains approximately 37,000 entries extracted from the *Oxford Advanced Learner's Dictionary of Current English* [Hornby, 1974] and the *Oxford Dictionary for Current Idiomatic English* [Cowie and Mackin, 1975]. It is available via FTP in both an ASCII and a database format. The database format uses a UNIX hash table facility [Seltzer and Yigit, 1991] that is freely distributed, and comes with an X-windows based interface for modifying the database and doing searches. C and Lisp hooks to allow other programs to use the database are also included.

2 Syntactic Lexicon

The syntactic lexicon has entries for 8 part-of-speech categories: Adjective, Adverb, Complementizer, Conjunction, Determiner, Noun, Preposition, and Verb. Each entry consists of the following required and optional fields:

- INDEX field (required) – the uninflected form under which the lexical item is compiled in the database;
- ENTRY field (required) – contains all of the lexical items associated with the INDEX[1];
- POS field (required) – gives the part-of-speech for the lexical item(s) in the ENTRY field;
- FRAME field (required) – contains the syntactic information about that entry;
- FS field (optional) – the Feature Structure field may provide additional information about the FRAME field.
- EX field (optional) – may be used for any number of example sentences.

Note that lexical items may have more than one entry in the database (e.g. *have*) and that they may select the same FRAME field more than once, using the FS to capture lexical idiosyncrasies (e.g. *map*). Table 1 shows selected entries from the database.

INDEX:	have
ENTRY:	have
POS:	Verb
FRAME:	Auxiliary_Verb
FS:	Goes_on_Infinitive
EX:	John has to go to the store.
INDEX:	have
ENTRY:	have
POS:	V
FRAME:	Transitive_Verb
FS:	Non-Ergative
EX:	John has a problem.
INDEX:	map
ENTRY:	map out
POS:	Verb Verb_Particle
FRAME:	Transitive_Verb_Particle
INDEX:	map
ENTRY:	map
POS:	Noun
FRAME:	Base_Noun
	Noun_Determiner_required
	Noun_Modifier
FS:	wh−, reflexive−
INDEX:	map
ENTRY:	map
POS:	Noun
FRAME:	Noun_Determiner_not_required
FS:	wh−, reflexive−, plural

Table 1: Selected Syntactic Database Entries

Because the syntactic database is part of the XTAG project [Doran *et al.*, 1994], a on-going project to develop a wide-coverage parser for English (see Section 7), some entries in the syntactic lexicon reflect specific XTAG analyses. In fact, the graphical interface for the syntactic lexicon (described in Section 4) can run in

[1] For example, a verb particle construction would be INDEXed under the verb, but would contain both the verb and the verb particle in the ENTRY field.

two modes - **xtag** and **verbose**. Tables 1, 2, and 3 were all generated in **verbose** mode.

The vast majority of lexical items in the database fall into just 3 categories - Adjectives, Nouns, and Verbs. These three categories plus Adverbs are presented in more detail in the following subsections.

2.1 Adjectives

There are 3,303 lexical adjectives in the database, of which 80 are 'Proper Name' adjectives, such as *Chinese* and *American*. Adjectives have 5 frames that they can select, which are listed below. Possible values for the FS field are **wh**− and **wh**+.

- **Base adjective**: All adjectives.
- **Modifying adjective**: Adjectives that can occur in direct modification contexts. Ex. *the* **Chinese** *man.*
- **Predicative adjective**: Adjectives that can occur as the complement of a predicative verb. Ex. *John was* **happy**.
- **Predicative adjective w/ sentential complement**: Adjectives that can occur as the complement of a predicative verb and that take a sentential complement. Ex. *John was* **happy** *that Mary left Bill.*
- **Predicative adjective w/ sentential subject**: Adjectives that can occur as the complement of a predicative verb and that take a sentential subject. Ex. *That John loves Mary is* **great***!*

2.2 Nouns

Nouns are by far the largest category in the syntactic database, accounting for well over 50% of the entries. Proper nouns and pronouns both have the part-of-speech Noun. Proper names, such as *Danielle* and *Nicholas* are not well-represented in the database, but geographic names, particularly places in England, generally are[2]. The frames for nouns are similar in many ways to the frames for adjectives, since nouns can modify other nouns and occur in predicative sentences. Other frames provide information about the use of the noun with determiners when forming noun phrases. The frames for noun are presented below:

- **Base noun**: All nouns.
- **Noun Phrase with Determiner**: Nouns that can take a determiner when forming a noun phrase. Ex. *a* **man**; **a* **jealousy**
- **Noun Phrase without Determiner**: Nouns that can appear without a determiner when forming a noun phrase. Ex. **envy**; ***plant**
- **Modifying noun**: Nouns that can modify other nouns. Note that not all nouns can modify other nouns. Proper nouns in general cannot modify other nouns, and specific lexical items may be restricted as well. Ex. **basketball** *game*; ***John** *car*
- **Noun with sentential complement**: Nouns that take sentential complements. Ex. *the* **fact** *that Mary loves John...*
- **Predicative noun**: Nouns that can occur as the complement of a predicative verb. Ex. *John was a* **man**.
- **Predicative noun w/ sentential subject**: Nouns that can occur as the complement of a predicative verb and that take a sentential subject. Ex. *That John loves Mary is a* **crime**.

Because this lexicon is used in the XTAG system, the lexicon often indicates precise syntactic behavior, rather than simply placing a general label on a lexical item. For the class of nouns, this is seen in the specification of nouns with respect to their co-occurrence with determiners. Instead of assigning a general label as as 'common noun' or 'mass noun', the noun frames explicitly indicate whether certain forms of the noun can appear with or without a determiner. However, since the syntactic database is indexed on root forms only, the morphology of the lexical item is not available. Instead, the FS field is used to indicate any restrictions on a particular use of a lexical item. For example, in Table 1, the noun *map*

[2] This reflects the origin of the dictionary from which the lexicon was originally extracted.

occurs twice. The first time that it appears, it selects the **Noun_Determiner_required** frame. The feature structures associated with it indicates only that the noun is not a wh-word, and that it is not reflexive. No restrictions are made with respect to its morphology. In contrast, the second entry, which selects the **Noun_Determiner_not_required** has **plural** as part of its FS. This indicates that the noun for this frame is restricted to its plural form. Hence *map* can only occur with a determiner, but *maps* is free to occur both with or without one. Nouns that belong to the class of so-called 'mass nouns' would not have the **plural** restriction on the entry that selects the **Noun_Determiner_not_required** frame, thereby indicating that the singular form is also allowed to occur without a determiner.

2.3 Verbs

Verbs, with their varied subcategorization frames, are perhaps the most interesting lexical items in a syntactic lexicon. There are over 8100 verbs (not including auxiliary verbs) that make up almost 9000 entries in the database. There are 19 different frames that the verbs can select, including transitive, intransitive, sentential complement, sentential subject, verb particle constructions (transitive and intransitive), double objects with shifting, double objects without shifting, and light verb constructions.

As with the nouns, the FS field is used to provide a more concise format for specifying the frames for each lexical item. For the verbs, the FS field is used to specify the difference between ergative and non-ergative transitive verbs, as can be seen in the *have* entry in Table 1, and is also used heavily for further differentiating the frames for verbs that take sentential complements. There are two frames for sentential complements - **Sentential_Complement** and **NP_and_Sentential_Complement**. Either of these can occur with the feature structures **Infinitive_Complement**, **Indicative_Complement**, or **Predicative_Complement**. This reduces the number of values for FRAME that are necessary to cover all of the possible lexical environments, and also allows for easier searches across categories. To find all the verbs that take infinitive complements, one can simply search on the **Infinitive_Complement** feature structure, rather than having to specify each frame that could fill this role. Table 2 shows some values for various verbs that take sentential complements.

INDEX:	want
ENTRY:	want
POS:	Verb
FRAME:	Sentential_Complement
FS:	Infinitive_Complement
EX:	Dan wants to finish this paper.
INDEX:	want
ENTRY:	want
POS:	Verb
FRAME:	NP_and_Sentential Complement
FS:	Infinitive_Complement
EX:	Dan wants Al to finish this paper.
INDEX:	think
ENTRY:	think
POS:	Verb
FRAME:	Sentential_Complement
FS:	Indicative_Complement
EX:	Dan thought that the paper was done.
INDEX:	think
ENTRY:	think
POS:	Verb
FRAME:	Sentential_Complement
FS:	Infinitive_Complement
EX:	Doug thought to clean the kitchen.
INDEX:	think
ENTRY:	think
POS:	Verb
FRAME:	Sentential_Complement
FS:	Predicative_Complement
EX:	Dan thought Carl a jerk.

Table 2: Verbs with Sentential Complements

2.3.1 Auxiliary verbs

The lexical entries for auxiliary verbs are very closely tied to the XTAG analysis, which orders the auxiliary verbs based on their morphological forms. Each entry in the lexicon is restricted via the FS field to only a certain form of the auxiliary verb (**present**, **past**,

ppart, etc), which also indicates what other forms that it can go on[3]. Table 3 shows the entries for the auxiliary verbs for the sentence *John should have been waiting.*

INDEX:	should
ENTRY:	should
POS:	Verb
FRAME:	Auxiliary_Verb
FS:	Indicative, Present, Goes_on_Base
INDEX:	have
ENTRY:	have
POS:	Verb
FRAME:	Auxiliary_Verb
FS:	Base, Goes_on_Past_Participle
INDEX:	be
ENTRY:	be
POS:	Verb
FRAME:	Auxiliary_Verb
FS:	Past_Participle, Goes_on_Gerund

Table 3: Example Auxiliary Verb Entries

2.4 Adverbs

A syntactic lexicon for adverbs is particularly useful because adverbs are so idiosyncratic as to where they can occur in a sentence. Although there are only 169 adverbs in the syntactic lexicon, but there are 15 different FRAME values that they can select. These include basic adverb, pre and post verb phrases, pre and post sentences, pre and post adjective, pre-adverb, pre-preposition, pre-noun, etc. Table 4 shows some selected adverb entries.

INDEX:	ahead
ENTRY:	ahead
POS:	Adverb
FRAME:	Base_Adverb
	Post-VP
	Pre-PP
INDEX:	essentially
ENTRY:	essentially
POS:	Adverb
TREES:	Base_Adverb
	Pre-VP
	Pre-S
	Post-S
INDEX:	even
ENTRY:	even
POS:	Adverb
FRAME:	Base_Adverb
	Pre-VP
	Pre-Adj
	Pre-Noun
	Pre-PP
INDEX:	very
ENTRY:	very
POS:	Adverb
FRAME:	Base_Adverb
	Pre-Adj
	Pre-Adv

Table 4: Some Adverb Lexical Entries

3 File Formats

The information in the syntactic database is available both in an ASCII 'flat' file, and a hashed database format. The ASCII file contains one entry per line, and each field is clearly marked. This format is easily usable by various UNIX^{tm} utilities such as *grep* and *awk*, and it can be easily parsed by custom programs.

The hashed database format is very useful for programs that need quick access to the information in the database. Each entry is indexed under the INDEX key, and a single call to the database for a particular index returns all of the entries that share that index. This makes it particular useful for parsers. The database uses an encoding scheme for the POS, FRAME, and FS fields, which condenses the space required for the database and shortens the search time for non-index fields. All of the entries for a given lexical index can be retrieved in 1.6 msecs, on average.

4 Interface

Although the format of the flat file is excellent for various file utilities programs, and

[3]For a more detailed description of this and other XTAG analyses, please see the XTAG Technical Report [The XTAG Project, 1994].

the database format works well for retrieving entries quickly, neither is particularly well-suited for human readability. The X-windows interface[4] for the syntactic database allows users to easily look at the database. Searching is available not only on the INDEX under which the lexical item is stored, but also on all other fields, with the exception of the EX field. Searches may also be done on combinations[5] of fields. For instance, one could search on POS = **Noun** and FS = **wh+** to find the set of all wh+ nouns (*what*, *who*, *whom*, *which*, *when*). Figure 1 shows the interface after a search has been done on the index *need*. All of the entries with that index are listed in a scroll window, which can be browsed through using the **Next** and **Previous** buttons, or specific entries can be clicked on, and the entire record will show in the upper window. The results of searches can be saved to a file to create smaller 'custom' lexicons. In addition to searching the database, users can also easily add, delete and modify individual entries, tailoring the syntactic database to fit their needs. Users may also delete all entries found in a given search, and we hope to add the capacity to modify a entire set of entries in the future.

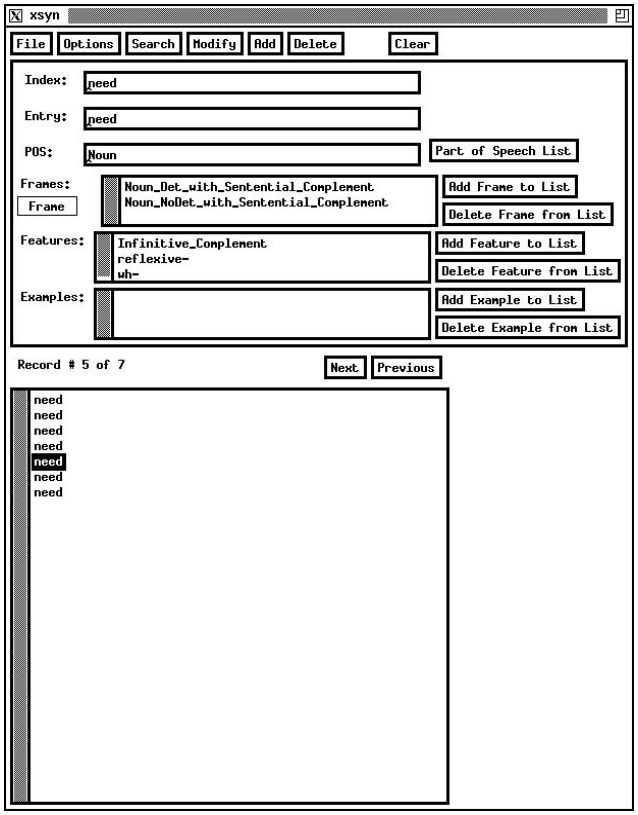


Figure 1: Result of a search on the index *need*

Corpus	Number of Hits	Total # of Words	Percent Hit
WSJ	1974528	2462557	80.18%
Brown	799904	991008	80.72%
IBM	60944	68800	88.58%
ATIS	10156	13791	73.64%

Table 5: Percentage of Hits for various corpora

5 Statistics

Statistics were gathered on the coverage of the syntactic lexicon on the IBM, ATIS, WSJ, and Brown corpora. These corpora were chosen because they have been tagged and hand corrected by the TreeBank project [Santorini, 1990]. The data in Table 5 show the coverage of the lexicon on various corpora. A lexical item/part-of-speech pair is counted as a **hit** if the lexical item is in the syntactic lexicon with the indicated tag. No attempt was made to determine if the lexicon had the correct frame needed to parse the sentence.

Because the syntactic lexicon contains only the root form of lexical entries, the inflected form was first looked up in the morphology database [Karp *et al.*, 1992] to retrieve the root form, and then that was used for the syntactic lexicon. Items that were not found in the morphological database were counted against the syntactic lexicon, as the morphology database is a superset of the syntactic database[6]. The statistics in Table 5 are over all word occurrences in the corpora[7], so words that occur frequently are given more weight.

Not surprisingly, nouns and proper nouns[8]

[4]The interface uses the MIT Athena Toolkit, which is distributed with the standard MIT X release.

[5]We hope to add expand this in the future to include full regular expression searches.

[6]Because these databases are being used in an actual parser, an attempt was made some time ago to make ensure that all words in the syntactic lexicon appear in the morphological database. Although the databases may have diverged slightly since then, it should not be statistically significant.

[7]Numbers and the genitive marker ('s) were taken out before the statistics were compiled.

[8]Although we do not distinguish nouns and proper nouns in the syntactic lexicon, the TreeBank tags do

Corpus	Number of Non-hits	Percent Proper N	Percent Nouns	Percent Adj	Percent Adv	Percent Verbs
WSJ	488029	43.8%	30.7%	13.8%	5.7%	1.3%
Brown	191104	26.2%	40.6%	14.8%	7.4%	1.8%
IBM	7856	17.1%	56.9%	11.3%	2.8%	2.5%
ATIS	3635	67.4%	14.0%	1.6%	0.6%	2.4%

Table 6: Percentage of missing words for various Parts of Speech

comprise the largest category of words missed, followed by adjective, adverbs, and verbs. Table 6 shows the percentage of each of these categories in the list of items not found. Again, this is a percentage of word occurrences in the corpora.

As Table 6 indicates, the majority of the missing items are either nouns or proper nouns (66.8% - 81.4%). This is not surprising, nor particularly distressing, as nouns tend to be the easiest items to 'guess' information about. Verbs, which tend to be the hardest, are reasonably well-covered in this lexicon. The number of adjectives not covered, however, seems fairly high, and we plan to add a number of those missing to the syntactic lexicon.

6 Future Work

The lexicon in its present form does not provide a mechanism to specify preferences of lexical items for certain syntactic structures. As part of future enhancements to the lexicon we hope to associate probabilities with each entry. The probabilities will reflect the affinity of the lexical item for the syntactic structure associated with that entry. These probabilities will be computed from parsed corpora.

It has been observed quite conclusively in recent work in lexicography that certain combinations of words co-occur more often than would be expected if they corresponded to arbitrary usages of the individual words. Collocational information has been shown to be of immense use in pruning the search space for a parser. We hope to eventually extract collocational information from the corpora and make it a part of the syntactic lexicon.

make this distinction, and it seemed useful to continue this distinction for this part of the analysis.

7 Related Work

The syntactic lexicon was developed as part of the XTAG project [Doran *et al.*, 1994] at the University of Pennsylvania under the direction of Dr. Aravind Joshi. The XTAG system is a wide-coverage parser and grammar for English based on the Tree Adjoining Grammar (TAG) formalism [Joshi *et al.*, 1975]. The English grammar consists of 3 sections - a morphology database, a syntactic database, and a tree grammar. Together with a parser and an X-windows interface, they comprise the XTAG system. Both the morphology [Karp *et al.*, 1992] and syntactic databases are available separately. The entire XTAG system is also freely available to the NLP research community. Information about the entire XTAG system and FTP instructions may be obtained by writing xtag-request@linc.cis.upenn.edu.

8 Computer Platform

The syntactic lexicon and accompanying interface were developed on the Sun SPARC station series, as were the other tools mentioned in Section 7. All of the XTAG tools, including the syntactic lexicon and interface, are freely available without limitation through anonymous FTP to **ftp.cis.upenn.edu**. The syntactic lexicon and accompanying programs together require about 9MB of space (for both the ASCII and DB versions of the lexicon). Please send mail to lex-request@linc.cis.upenn.edu for current FTP instructions or for more information.

References

[Carroll and Grover, 1989] Carroll, J. and Grover, C. (1989). The Derivation of a Large Computational Lexicon for English from LDOCE. In B. Boguraev and E. Briscoe (eds.) *Computational Lexicography for Natural Language Processing*, Harlow, UK: Longman, 177-134.

[Conlon, 1994] Conlon, S. Pin-Hgern. (1994). The IIT Lexical Database: Dream and Reality. In A. Zampolli, N. Calzolari, M. Palmer (eds.), *Current Issues in Computational Linguistics: In Honour of Don Walker*, Giardini with Kluwer.

[Cowie and Mackin, 1975] Cowie, A.P. and Mackin, R., eds. (1975). *Oxford Dictionary of Current Idiomatic English*, Volume 1, Oxford University Press, London.

[Doran *et al.*, 1994] Doran, C., Egedi, D., Hockey, B.A., Srinivas, B., Zaidel, M. (1994). XTAG System - A Wide Coverage Grammar for English. In *Proceedings of the* 16^{th} *International Conference on Computational Linguistics (COLING '94)*, Kyoyo, Japan, August.

[Guthrie *et al.*, 1993] Guthrie, L., Rauls, V., Luo, T., Bruce, R. (1993). *LEXICAD/CAM.* Technical Report MCCS-93-259, Computing Research Laboratory, New Mexico State University.

[Hornby, 1974] Hornby A. S., ed. (1974). *Oxford Advanced Learner's Dictionary of Current English*, Third Edition, Oxford University Press, London.

[Joshi *et al.*, 1975] Joshi, A., Levy, L. and Takahashi, M. (1975). Tree Adjunct Grammars. *Journal of Computer and System Sciences.*

[Karp *et al.*, 1992] Karp, D., Schabes, Y., Zaidel, M., Egedi, D. (1992). A Freely Available Wide Coverage Morphological Analyzer for English. In *Proceedings of the* 15^{th} *International Conference on Computational Linguistics (COLING '92)*, Nantes, France, August.

[Macleod *et al.*, 1994] Macleod, C., Grishman, R. and Meyers, A. (1994). Creating a Common Syntactic Dictionary of English. In *Proceedings of the International Workshop on Sharable Natural Language Resources*, Nara, Japan, August.

[Santorini, 1990] Santorini, B. (1990). *Part-of-Speech Tagging Guidelines for the Penn TreeBank Project.* Technical report MS-CIS-90-47, Department of Computer and Information Science, University of Pennsylvania.

[Seltzer and Yigit, 1991] Seltzer, M. and Yigit, O. (1991). A new hashing package for UNIX. In *USENIX*, Winter.

[The XTAG Project, 1994]
The XTAG Project (1994). *A Feature-Based, Lexicalized Tree Adjoining Grammar (FB-LTAG) for English.* Manuscript, University of Pennsylvania. In Progress.