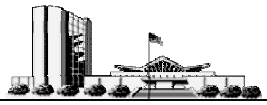
LISTER HILL NATIONAL CENTER
FOR BIOMEDICAL COMMUNICATIONS



TECHNICAL REPORT
LHNCBC-TR-2003-003

Lexical Systems; A report to the Board of Scientific Counselors September 2003

Allen C. Browne
Guy Divita
Chris Lu
Lynn McCreedy
Destinee Nace

U.S. National Library of Medicine, LHNCBC
8600 Rockville Pike, Building 38A
Bethesda, MD 20894

BACKGROUND AND OBJECTIVES

The Lexical Systems Group is a part of the Natural Language Systems (NLS) program at the Lister Hill Center of the National Library of Medicine (NLM), involved with the creation and maintenance of the SPECIALIST lexicon and tools that support and exploit it. The lexicon and its attendant lexical tools are one of the knowledge sources of the Unified Medical Language System (UMLS) and are distributed with the UMLS [1]. The lexical tools are used in the creation and maintenance of the UMLS Metathesaurus and are needed to access the Metathesaurus through its indexes. The lexicon, lexical tools, and other natural language tools developed by the Lexical Systems Group are at the center of NLM's natural language research, providing a foundation for all our natural language processing efforts.

PROJECT SIGNIFICANCE

The complexity of natural language poses a significant barrier to access to biomedical text. Words are the fundamental units of natural language, lying at the intersection of form and meaning. While there are significant generalizations within the lexicons of natural languages, such as the regular inflection rules of English, much lexical knowledge is idiosyncratically related to individual word forms [2, 3]. Any computational system designed to deal with natural language will need a record of this idiosyncratic lexical information. Machine-readable dictionaries (MRD's) have provided computational linguistics with tools to capture some of this information [3], but the usefulness of MRD's for computation is limited by the fact that dictionaries are primarily constructed for the use of humans who are already competent users of natural language. Much of the useful lexical information is implicit. Learner's dictionaries like LDOCE [4] are more likely to make this sort of information explicit but their coverage is necessarily limited to general English. Specialized medical dictionaries, such as Dorland's Dictionary [5], have less explicit syntactic and morphological information than standard English dictionaries, relying even more on the linguistic competence of the user. Extracting information from MRD's is itself a challenging task [6]. The SPECIALIST approach has been to record this information by human input with the assistance of computational tools. This method results in high quality information on which subsequent automatic methods can be based.

The SPECIALIST lexicon has been developed to meet the need for lexical information in the biomedical domain. Its wide coverage provides an important base on which other natural language processing tools are built. The lexicon has provided this essential underpinning to natural language projects at NLM and elsewhere through its wide dissemination within the medical informatics community.

The UMLS lexical tools developed by the Lexical Systems Group exploit the lexicon to provide methods for dealing with lexical variation. The SPECIALIST NLP tools in turn exploit the lexical tools and provide a basis for information retrieval and other NLP tasks inside and outside the NLM [7, 8, 9, 10].

STATUS REPORT

The SPECIALIST lexicon is a large syntactic lexicon that records orthographic, morphological, and syntactic information about biomedical and general English words and terms. The lexicon emphasizes medical technical vocabulary, drawn especially from the UMLS Metathesaurus and Medline abstracts. It also includes the general English vocabulary that constitutes a significant part of biomedical text. General English frequency lists [11, 12] were consulted in the early design of the lexicon and ongoing efforts attempt to morphologically identify verbs, adjectives, and adverbs that might appear in biomedical text.

Since 1990, the lexicon has been built with the help of a team of linguistically trained lexicon consultants using LexBuild, a lexicon-building tool designed to guide users through the process of building lexical entries. The information recorded for each entry is based on linguistically informed human judgment. Lexicon builders consult a variety of tools including general English dictionaries, English learners' dictionaries and medical dictionaries, both online and in paper form [4, 5, 13, 14, 15]. They have online access to both Medline, through PubMed, and the Metathesaurus through the Knowledge Source Server. Lexical judgments are made on the basis of dictionary information when available, on linguistically informed native speaker intuitions and on actual observation of usage. LexBuild assures that lexical entries are consistent and well formed. The lexicon has grown steadily since its inception in 1986. It has been released as a UMLS knowledge source since 1994. Fig. 1 shows the growth of the lexicon since that time. The lexicon contained just over 66,000 entries in its first release. Those entries accounted for almost 112,000 forms (inflectional and spelling variants). The 2004 release will have over 220,000 entries accounting for over 343,000 forms.

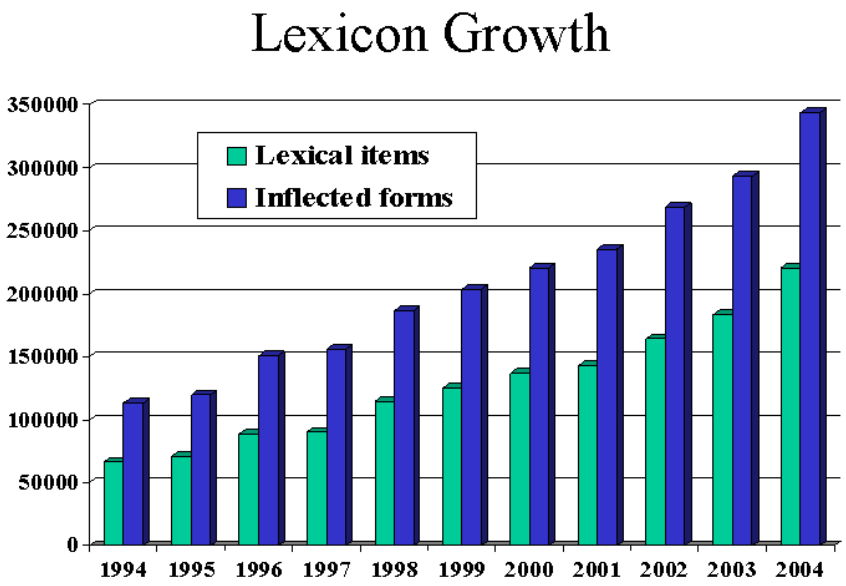


Figure 1 Lexicon Growth Since 1994

The distribution of lexical categories (parts of speech) in the lexicon is shown in Fig. 2. The lexical categories represented in the SPECIALIST lexicon are the standard parts of speech encountered in most grammars of English [16, 17].

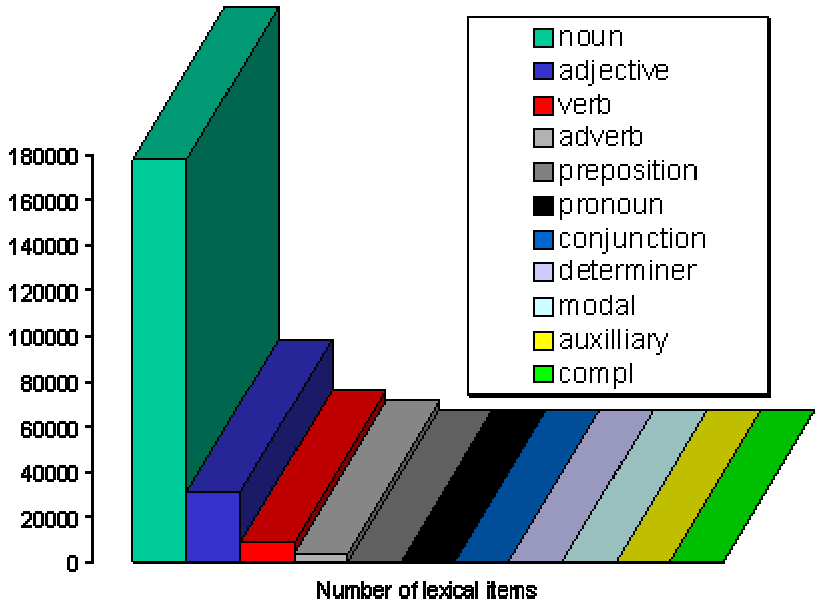


Figure 2 Lexical Category Distribution in the 2004 Release

Many lexical items have more than one part of speech. The majority of lexical entries in the SPECIALIST lexicon are nouns. While nouns predominate in English generally, their number is inflated in our lexicon by a large number of technical terms.

The lexical tools are a suite of programs, built on top of the SPECIALIST lexicon and designed to help users deal with lexical variation and other natural language processing tasks. The lexical tools include Wordind, Norm, and LVG. Wordind is a tokenization tool embodying the UMLS definition of "word". It breaks text or terms into continuous stretches of alphanumeric characters. Wordind is used to create the UMLS word indexes. LVG (Lexical Variant Generation) contains a wide variety of text manipulation facilities aimed at the creation of indexes. Norm is a particular configuration of LVG used to create the UMLS normalized indexes. Both Norm and Wordind are necessary to access the indexes supplied with the UMLS.

The lexical systems group has also developed a set of tools designed to facilitate text processing. These text analysis tools exploit the lexicon and lexical tools. They include a tokenizer, text chunker, lexical lookup and shape recognition tool and a Java implementation of the MetaMap program.

Methods and Procedures

The SPECIALIST Lexicon

The lexicon is built and maintained manually with the help of LexBuild, a web based lexicon building tool that guides lexicon builders through the structure of lexicon records and assures that records are complete and well formed. Each bit of lexical information has been entered by a human lexicon builder. All members of the lexicon building team

have graduate degrees in linguistics and have been trained at NLM for the task of lexicon building. The current implementation of LexBuild uses the Gspell spelling suggestion system to identify possible spelling variants as they are entered. Lexbuild verifies the wellformedness of each entry using the lexical grammar, a BNF representation of the lexical record. Lexicon consultants work off-site and upload records to NLM bi-weekly where they are reviewed and entered into the SPECIALIST lexicon. A new version of LexBuild is being developed which will enter records directly into the lexicon database. A flag will be set when the entry has been reviewed. This will relieve a bottleneck in the system and prevent occasional double entry of lexical items.

SPECIALIST lexical entries are frame structures of slots and fillers. Each record has a base form indicated by a 'base=' slot, a category indicated by a 'cat=' slot and an entry unique identification number (EUI) indicated by an 'entry=' slot. Other slots appear in the records for specific categories. Lexical entries include information on orthography, morphology, and syntax. The lexicon has records for single and multiword lexical items, since English lexical items often contain more than one orthographic word. Natural language is fundamentally oral and many languages lack a writing system (orthography). Orthography is secondary to this natural system and imposes distinctions not found in the language itself. "Ice cream" for example is certainly a word of English although it is spelled with two orthographic words [2, 20]. To avoid confusion between the several senses of "word" we use the term "lexical item" to cover both single and multi-word lexical units. The SPECIALIST lexicon records multi-word items when the meaning is not a function of the meanings of the constituent orthographic words. "Heart attack", for example, has a specific meaning not covered by the construction of "heart" and "attack". Occurrence in dictionaries and synonymous acronyms are indicators that a phrase may be a lexical item.

Orthography deals with the spelling of words. In the lexicon the possible spellings of a word are indicated in the 'base=' and 'spelling_variant=' slots. When an item has only one spelling, its citation form is indicated in the 'base=' slot. If there are alternate spellings they will appear in the 'spelling_variant=' slot. The 'base=' slot is just one of the variants which serves as preferred name for the record. The form in the 'base=' slot need not be the most common or correct spelling. Only legitimate spelling variants are recorded. Spelling and typographic errors are not. These variants are handled by the Gspell spelling suggestion system. The spelling conventions of contemporary English are less settled than is generally believed. Studies of spelling variation in standard dictionaries show that many words have more than one spelling in a single dictionary and that the standard dictionaries often list different spellings for the same word [18, 19]. Major variables affecting spelling of English lexical items include hyphenation and spacing in compounds, as well as capitalization. Spelling variation due to national and regional standards of English is also common. For example, the spelling "oedema" is more likely to be seen in British English than American English. The records below for "edema" and "referable" illustrate SPECIALIST lexical records and show how spelling variants are recorded.

```
{base=edema
spelling_variant=oedema
entry=E0024504
        cat=noun
        variants=uncount
        variants=reg
        variants=glreg
}

{base=referable
spelling_variant=referrible
spelling_variant=referrable
entry=E0052409
        cat=adj
        variants=inv
        position=attrib(3)
        position=pred
        compl=pphr(to,np)
        stative
        nominalization=referability|noun|E0219410
}
```

The form in the 'base=' slot is the form used to name the record; it need not be preferred over the other spelling variants. The lexicon makes no judgment as to which, if any, spelling variant is most correct. The inflection codes are applied to all of the spelling variants.

Inflectional morphology deals with the forms taken by words in different syntactic contexts. For example "nucleus" and "nuclei" are inflectional forms (singular and plural) of the same word. Derivational morphology deals with the formation of words from other words by processes such as suffixation and prefixation [20, 21].

Inflectional morphology is indicated in the lexical entry with the 'variants=' slot. The fillers of the 'variants=' slot indicate the inflectional pattern of the lexical item. Nouns, verbs, adjectives, and adverbs are inflected in English. Although there are regular patterns of inflection, English displays great variety and idiosyncrasy in its inflectional system. Nouns, for example, are broadly divided into count and uncount nouns. Uncount nouns are nouns that do not have plurals and can appear without an article e.g. 'mud', 'smallpox' or 'potassium'. The lexicon marks these nouns 'variant=uncount'. Count nouns generally have both a singular and plural form and must appear with an article. Nouns that inflect according to the regular English pattern of inflection are marked 'variants=reg'. The "variants=reg" slot in the record for 'edema' above indicates that it can occur with the regular "-s" suffix, that is as 'edemas'. The rule applies to both spellings, so 'oedemas' is also the plural of this lexical item. Those items that inflect according to a selected set of Greco-Latin patterns are marked 'variants=glreg' for Greco-Latin regular. The plurals 'oedemata' and 'edemata' are produced by the Greco-Latin regular rules. In addition to the regular and Greco-Latin regular patterns there are fixed plural nouns that have no singular form e.g. 'police'. These nouns are marked 'variants=plur'. In some nouns like 'committee' the singular form can agree with

singular or plural verbs. These collective nouns are recorded as group nouns in the lexicon. Many nouns are simply irregular. Irregular plurals are represented by ‘variants=irreg||’ as shown in the record for ‘larynx’ below.

```
{base=larynx
entry=E0036919
        cat=noun
        variants=irreg|larynges|
        variants=reg
}
```

Inflectional rules in English are basically phonological in nature. Even the spelling rules make use of phonological information [22]. These rules are reflected in the SPECIALIST lexicon as purely orthographic rules; for example the rule for adding “s” to form the third person singular of verbs or the plural of nouns requires “es” following a sibilant. The regular inflection rule used in the SPECIALIST lexicon has to mention the letters “s”, “z”, “x”, “sh” and “ch” which generally represent sibilants. The word in the word “patch”, “ch” represents a sibilent sound and “patch” therefore listed as regular. In “stomach” the “ch” does not represent a sibilant. “Stomach” is therefore listed as an irregular noun.

Verb and adjective inflections are treated similarly. Some verbs double their final consonant before the past tense and past participle suffix ‘ed’. The past tense of “format” is “formatted”. This regular doubling pattern in indicated by ‘variants=regd’.

```
{base=format
entry=E0028590
        cat=verb
        variants=regd
        tran=np
}
```

Because of the wide variety and unpredictability of inflectional patterns, a lexicon is required to record this information.

The SPECIALIST lexicon records information specific to a lexical item. Derivational morphology, which deals with relations between separate items, is generally not recorded in the lexicon itself, with one exception. Nominalization is a special instance of derivational morphology, which is represented directly in the SPECIALIST lexicon because of its pervasiveness in technical writing and its syntactic importance. Some verbs and adjectives have synonymous nouns derived from them by the process of nominalization. The noun ‘anticipation’ is the nominalization of the verb ‘anticipate’.

```
{base=anticipate
entry=E0009453
        cat=verb
        variants=reg
        intran
        tran=np
        tran=fincomp(o)
```

```
        tran=whinfcomp:arbc
        tran=whfincomp
        tran=ingcomp:subjc
        nominalization=anticipation|noun|E0009455
}
```

Nominalizations have a ‘nominalization_of=’ slot containing a cross reference to the record of the verb or adjective. Nominalizations share the meaning and predicational structure of the verb or adjective they nominalize.

```
{base=anticipation
entry=E0009455
        cat=noun
        variants=uncount
        variants=reg
        compl=pphr(of,np)
        compl=pphr(by,np)
        nominalization_of=anticipate|verb|E0009453
}
```

Verbs and adjectives, which have nominalizations, also have ‘nominalization=’ slots cross-referencing them to their nominalizations.

Other types of derivational morphology are captured in an ancillary derivational morphology file and in a set of derivational morphology rules. The derivational variant facts file contains pairs of derivationally related words and their parts of speech.

Derivational facts
treatment|noun|treat|verb prohibition|noun|prohibitive|adj cell lineage|noun|cell line|noun photochemotherapeutic|adj|photochemotherapy|noun pharmacotherapeutic|adj|pharmacotherapy|noun

Figure 3 Derivational Facts

The derivational rules use a regular expression syntax to indicate the pattern of related words. The ‘$’ in the suffixation rule below indicates the end of the word.

Derivational Rules
e.g. alienation|alienate ation$\|noun\|ate$|verb ration|rate; station|state

Figure 4 Derivational Rules

Known exceptions to rules are recorded with the rules themselves. The derivational component of LVG applies these rules to heuristically generate possible derivational variants. This system of positive examples, rules and exceptions provides an adaptable way to use derivational morphology to recognize words not in the lexicon. Until

recently, the derivational morphology system dealt only with suffixation rules. A prefixation component has now been implemented. Further development will be needed to expand the list of prefixation rules and develop the list of prefixation facts.

Many medical and technical words are neoclassical compounds formed from Latin and Greek roots with connecting vowels [20, 23, 24]. A list of neoclassical combining forms categorized as prefixes, combining forms and terminals is distributed with the lexicon. Each form has a short English gloss to indicate its meaning.

Traditionally, morphological variation, both inflectional and derivational, has been computationally dealt with using stemmers that attempt to remove common suffixes from words to discover a common stem [25]. These systems are fast and computationally inexpensive but they are error-prone. Our method of recording positive instances, using heuristic rules only when dealing with unknown instances and recording exceptions eliminates many errors in finding morphologically related strings.

The lexicon's syntactic information includes verb complements and adjective positions. Traditionally verbs are categorized as intransitive or transitive depending on whether they take noun phrase complements (objects). The SPECIALIST lexicon recognizes other sorts of objects and expands the list to include ditransitive, linking, and complex transitive verb complementation patterns. These verb complementation patterns provide important syntactic information. They provide the structure of verb phrases that, in turn, are the skeletons of sentence structure. This is an example of the projection principle [26, 27], which says that representations at all syntactic levels are "projected" from the lexicon. The code ditran=np,pphr(with, np) licenses a verb phrase like "treated the patient with the drug".

```
{base=treat
entry=E0061964
        cat=verb
        variants=reg
        intran
        tran=np
        tran=pphr(with,np)
        tran=pphr(of,np)
        ditran=np,pphr(to,np)
        ditran=np,pphr(with,np)
        ditran=np,pphr(for,np)
        cplxtran=np,advbl
        nominalization=treatment|noun|E0061968
}
```

Until recently, number words have not appeared in the lexicon. Words like "five", "ten", "thousand", and "trillion" do not fit into the category system of the SPECIALIST lexicon. Number expressions like "five thousand three hundred and eight" function in English as determiners. Number words are the building blocks of number expressions. There is an infinite number of number expressions but a finite number of number words. The 2003 release of the SPECIALIST lexicon includes a separate file of lexical entries for number words. These records indicate inflection (cardinal, ordinal, denominator forms) and

contain the features necessary to determine the grammaticality and numerical value of a number expression. The syntax adopted for number expressions closely follows the analysis of Bauer and Huddleston [28].

The Lexical Tools

The UMLS lexical tools are a set of programs that exploit the SPECIALIST lexicon to deal with the lexical variation inherent in natural language. They aid in pattern matching and in the creation of word and term indexes. The lexical tools distributed with the UMLS include Wordind, LVG and Norm. Wordind is a tokenizer that implements the UMLS definition of "word", a sequence of non-alphanumeric characters. LVG (Lexical Variant Generation) is a suite of tools that provide methods to abstract away from lexical variation. Norm is a particular configuration of LVG used to create the normalized indexes in the UMLS. These tools have a command line interface, which reads pipe-fielded input from Standard in and writes output to Standard out. They are also available as Java API's and through a web based GUI tool.

Wordind is used to create the word and normalized word indexes in the UMLS. It breaks text into words, lowercases and removes white space and punctuation. Users of the word indexes should use Wordind to tokenize their queries in order to assure consistency with the indexes.

Norm abstracts away from alphabetic case, punctuation, word order, possessives and inflectional variation. The output of norm is all lower case, with each word in its morphological base form, without "'s", without punctuation, and with words sorted in alphabetic order. By reducing words to their base form all the inflectional variants can be matched to the same index entry. Word order sorting is needed to match terms like "cancer, liver" and "liver cancer". Many terms in medical vocabularies are inverted for alphabetization. LVG can uninvert terms inverted around a single comma but many terms are ambiguously inverted. Word order sort handles all varieties of inversion. Norm is used to create the normalized string and word indexes. Users need to use Norm when they access those indexes and queries need to be normalized for compatibility with the indexes. As an example, each of the terms below appear in the UMLS Metathesaurus, and they all normalize to the same string "disease hodgkin". In practice this means that a user of a system using normalized indexes generated by Norm could begin with any of those variants and find all the others. A variant of Norm is used to generate the lexical variant classes (LUI's) in the Metathesaurus so all those forms will fall into the same concept.

UMLS Metathesaurus Strings normalized as "**disease hodgkin**"
•Hodgkin Disease **•HODGKINS DISEASE** **•Hodgkin's Disease** **•Disease, Hodgkin's** **•HODGKIN'S DISEASE** **•Hodgkin's disease** **•Hodgkins Disease** **•Hodgkin's disease NOS** **•Hodgkin's disease, NOS** **•Disease, Hodgkins** **•Diseases, Hodgkins** **•Hodgkins Diseases** **•Hodgkins disease** **•hodgkin's disease** **•Disease;Hodgkins** **•Disease, Hodgkin**

Figure 5 Normalization

The 2004 release of Norm will additionally deal with spelling variation, using the 'base=' form to represent all the spelling variants. It will convert non-ASCII characters to ASCII, split ligatures, and remove diacritic marks.

LVG is a suite of tools designed to allow users to do pattern matching or build indexes. It offers a large menu of tools to manipulate input text for indexing. These transformations, called flow components, can be arranged into flows so that the output of one is the input of the next. LVG can execute multiple flows in parallel. The transformations range from lowercasing through base forming and removing punctuation to recursively finding derivationally related words and synonyms. Some LVG flow components are shown below.

Some LVG flow components
n no-op l lowercase u uninvert g genitive marker removal s spelling variants w sort words that make up term in ASCII ascending order p remove punctuation i generate inflectional variants b reduce term to base form(s) B reduce each 'word' to its base form(s) d generate derivational variants t remove stop words

Figure 6 LVG Flow Components

LVG implements the morphological rules (inflectional and derivational) described above to generate derivational and inflectional variants for input words and terms. It can uninvert inverted terms like “cancer, lung” into “lung cancer” or as in Norm, sort the constituent words in alphabetical order.

Gspell is a spelling suggestion program that uses several algorithms to find words similar to a candidate misspelled word. Spell checking algorithms are well known and widely used [29, 30, 31]. Unlike other spelling suggestion systems, Gspell deals with multi-word and multi-token input so that “noncontributory” could be a suggestion for “non contributory” or “non-contributory”. Gspell is not only useful for spelling correction, it is also used in LexBuild to detect potential legitimate spelling variants.

The Gspell output below shows suggestions for the misspelling “anonomous”. The number in the third field indicates the edit distance between the suggestion and the input spelling and the number in the fourth column is used to rank suggestions.

Example Gspell output
anonomous|anonymous|1.0|0.8734230160180236|NGrams| anonomous|allonomous|2.0|0.5819672267388108|NGrams| anonomous|autonomous|2.0|0.5819672267388108|NGrams| anonomous|anadromous|3.0|0.2958160192082048|NGrams| anonomous|analogous|3.0|0.2958160192082048|NGrams| anonomous|anomalous|3.0|0.2958160192082048|NGrams| anonomous|anonymously|3.0|0.295816019208248|NGrams| anonomous|anonymes|3.0|0.2958160192082048|Metaphone| anonomous|anonyms|3.0|0.2958160192082048|Metaphone| …

Figure 7 Gspell Output

As an example of the utility of the lexical tools consider the Lister Hill Center’s ClinicalTrials.gov web site. If a ClinicalTrials.gov user types the misspelling “osteoparoses” in the type-in window she will be shown two possible correct spellings, “osteoporoses” and “osteopetroses” suggested by Gspell. When the user chooses “osteoporoses” she will be shown clinical trials involving “osteoporosis”. This mapping between the singular and plural forms of “osteoporosis” comes from the lexicon via LVG and Norm.

Text Processing Tools

The Lexical Systems Group also produces a set of text processing tools. These tools include a set of NLP tools, a Java implementation of MetaMap and a spelling suggestion tool. The NLP tools comprise a tokenizer, a lexicon look-up term recognizer and a phrase parser. These programs are embedded so that the parser includes the term recognizer and the term recognizer includes the tokenizer. They are available as Java API’s.

The tokenizer breaks free and structured text into sections (paragraphs), sentences and tokens. Sentences are recognized with a regular expression method based on the work of Grefenstette and Tapanainen [32]. The tokenizer can deal with free text, HTML, and Medline abstracts. The example below shows a sentence and its tokens. The first field is

the name of the Java object, which this line of output represents and the second field is an identification number for the token. The third and fourth fields indicate the location of first and last characters of the token. The blank fields at the end of each line represent information to be filled in with subsequent processing.

Sample tokenizer output

```
Sentence|0|0|50|Tests for African tick-borne fever were negative.

Token|0|0|4|0|Tests|||
Token|1|6|8|1|for|||
Token|2|10|16|2|African|||
Token|3|18|21|3|tick|||
Token|4|22|22|3|-|||
Token|5|23|27|4|borne|||
Token|6|29|33|5|fever|||
Token|7|35|38|6|were|||
Token|8|40|47|7|negative|||
Token|9|48|48|8|.|||
```

Figure 8 Tokenizer Output

The term recognizer recognizes lexical items in text by look-up in the SPECIALIST lexicon and by regular expression. In the example output below the multi-token item “African tick-borne fever” is recognized based on its occurrence in the lexicon. Part of speech from the lexicon is added to the records and positional information from the tokenizer is retained.

Sample term recognizer output

```
Lexical Element|0|LEXICON|noun|Tests|0|4
Lexical Element|1|LEXICON|prep|for|6|8
Lexical Element|3|LEXICON|noun|African tick-borne fever|10|33
Lexical Element|8|LEXICON|aux|were|35|38
Lexical Element|9|LEXICON|noun|negative|40|47
…
```

Figure 9 Term Recognizer Output

The NP parser chunks free text into phrases based on the parts of speech derived from the SPECIALIST lexicon. A statistical tagger can be interposed at this point to disambiguate lexical items having more than one part of speech; otherwise the parser uses simple heuristics to resolve ambiguous parts of speech.

Example Phrase Chunker output

```
Phrase|0|0|4|Tests|Tests|1
Phrase|1|6|33|for African tick-borne fever|African tick-borne fever|5
Phrase|2|35|38|were||1
Phrase|3|40|48|negative|negative|1
…
```

Figure 10 Phrase Chunker Output

Tokens, lexical items, phrases and sentences are Java objects with pointers to their constituent parts so that information from one level of analysis is available at other levels. One output format of the NLP tools is the prolog structure shown below.

Prolog Output Structure

```
[
    prep([lexmatch([for]),
          inputmatch([for]),
          tag(prep)]),
      head([lexmatch(['African tick-borne fever']),
          inputmatch(['African',tick,'-',borne,fever]),
          tag(noun)])],
[
    aux([lexmatch([were]),
          inputmatch([were]),
          tag(aux)])],
[
    head([lexmatch([negative]),
          inputmatch([negative]),
          tag(noun)]),
    punc([inputmatch(['.'])])]
```

Figure 11 Prolog Structured Output

The SPECIALIST NLP tools are available as an independent tool set and they embedded within the MMTx, the Java implementation of MetaMap, a program that maps terms to the UMLS Metathesaurus.

Project Plan

Continued development of the SPECIALIST lexicon is needed for the continued success of natural language processing at NLM. While the lexicon has reached a level of critical mass so that it is useful for many natural language tasks in medical informatics, there is still medical vocabulary not yet covered by the lexicon. The growth of both Medline and the Metathesaurus and the growth of medical vocabulary require continued lexicon growth. The lexicon building effort will continue to emphasize words found in the UMLS Metathesaurus and in Medline abstracts. Common English vocabulary will also be emphasized by comparing Metathesaurus and Medline wordlists with available dictionaries and spell checking lists. Quality assurance and maintenance are also required.

Lexical items have to be reviewed and corrected continually by the lexicon building team.

UNICODE and UTF-8 will soon be the UMLS standard and the lexical tools are being reworked to handle UNICODE input. The lexicon and lexical tools have until recently been strictly 7-bit ASCII. UTF-8 will allow the lexicon to deal with English words that have diacritic marks not found in 7-bit ASCI such as “résumé”, “cliché” and “déjà vu”. Medical text also includes personal names spelled with non-ASCII diacritics. The next release of the lexical tools will be able to handle UTF-8. LVG and Norm will include a transformation to strip diacritic marks and a transformation to convert any UTF-8 character to 7-bit ASCII.

Design efforts are beginning on a Java based statistical tagger to supplement the NLP tools. This facility will apply statistical techniques to disambiguate parts of speech in context. Currently available taggers are difficult to integrate into the system and are not freely redistributable. One design feature of the new tagger will be the ability to work with multi-words so that it will be able to exploit the multi-word information in the lexicon.

An effort is underway to gather corpora of medical text for experimentation and statistical generalization. These data may be needed to train the statistical tagger and will provide a test bed for other statistical NLP techniques. Bootstrapping methods are under consideration to alleviate the difficulty and inaccuracy of human tagging.

Summary

The SPECIALIST lexicon represents an extensive accumulation of medical English lexical information that is exploited by the lexical tools and NLP tools. These resources are available to researchers at the Lister Hill Center and to the medical informatics community and provide a foundation that supports further natural language processing.

References

1 McCray AT, Srinivasan S, Browne AC. Lexical Methods for Managing Variation in Biomedical Terminologies. In: Proceedings of the 18th Annual Symposium on Computer Applications in Medical Care, 1994, 235-239.

2 McCray AT, The Nature of Lexical Knowledge. Methods of Inf Med, Vol 37 No. 4-5, 1996.

3 Guthrie L, Pustejovsky J, Wilks Y, and Slator BM. The Role of Lexicons in Natural Language Processing, Communications of the ACM Vol. 39 No. 1 1996.

4 Procter P (ed). Longman Dictionary of Contemporary English. Longman Group Limited, 1978.

5 Taylor EJ (ed.) Dorland's Illustrated Medical Dictionary. W.B. Saunders Company, 1988.

6 McCray AT, and Srinivasan S. Automated Access to a Large Medical Dictionary: Online Assistance for Research and Application to Natural Language Processing. Computers and Biomedical Research 23, 1990

7 Hauser S, Browne A, Thoma G, McCray A, Lexicon Assistance Reduces Manual Verification of OCR Output. The 11th IEEE Symposium on Computer-Based Medical Systems.

8 Tuttle MS, Olson NE, Keck KD et al. Metaphrase: An Aid to the Clinical Conceptualization and Formalization of Patient Problems in Heathcare Enterprises. Meth Info Med 1998:37(4-5):373-83.

9 Johnson SB. A Semantic Lexicon for Medical Language Processing, JAMIA, June 1999, 6 (3) 205-218

10 Mills EM, Wilcke JR and Bender HS. Use of the Metathesaurus and SPECIALIST lexicon of the Unified Medical Language System, Lexical Matching and Domain-Specific Free-Text to Identify Undocumented Vocabulary. JAMIA Suppl S 1998 1043-1043.

11 Kucera H, and Francis WN. Computational Analysis of Present-Day American English. Providence: Brown University Press, 1967.

12 Carroll, JB. The American Heritage Word Frequency Book. Boston: Houghton Mifflin, 1971.

13 Sinclair J (ed) Collins Cobuild English Dictionary. Collins 1987.

14 Pease RW (ed) Websters Medical Desk Dictonary, Merriam-Webster Inc. 1996.

15 Merriam Webster Unabridged Online Dictionary, http://unabridged.merriam-webster .com.

16 Quirk R, Greenbaum S, Leech G, Svartvik J. A Comprehensive Grammar of the English Language. London: Longman Group Limited, 1985; 1515-1585.

17 Huddleston, R and Pullum G eds. The Cambridge Grammar of the English Language, Cambridge University Press, 2002, 1842 pages.

18 Emery DW. Variant Spellings in Modern American Dictionaries. National Council of Teachers of English, 1973; 130 pages.

19 Deighton, LC, A Comparative Study of Spellings in Four Major Collegiate Dictionaries, Hardscrabble Press, 1972; 144 pages.

20 Bauer L. English Word Formation. Cambridge: Cambridge University Press, 1983; 311 pages.

21 Marchand H. The Categories and Types of Present-Day English Word-Formation. Munich: C.H. Beck, 1969; 545 pages.

22 Cummings, DW American English Spelling, Johns Hopkins University Press 1988.

23 Pacak M, Norton LM, Dunham G. Morphosemantic Analysis of –it is Forms in Medical Language. Meth Inform 1980; 19:99-105.

24 McCray AT, Browne AC, and Moore DL. The Semantic Structure of Neo-Classical Compounds. In: Proceedings of the Twelfth Annual Symposium on Computer Applications in Medical Care, 1998, 165-8

25 . Porter MF. An algorithm for suffix stripping. Programming 1980;14:130-137.

26 Chomsky N Lectures on Government and Binding. Foris Publications 1982.

27 Szabolcsi A Combinatory Grammar and Projection from the Lexicon. In Sag IA, and Szabolcsi A (eds) Lexical Matters. Center for the Study of Languages and Information. 1992: 241-68.

28 Hudddleston R and Bauer L. Lexical word-formation, In: Huddleston, R and Pullum G eds. The Cambridge Grammar of the English Language, Cambridge University Press, 2002, 1842 pages.

29 Peterson JL. Computer programs for detecting and correcting spelling errors. Communications of the ACM 1980;23(12):676-687. Page 15

30 Bently J, A Spelling Checker, Programming Pearls, Communications of the ACM 28:5:456-462, 1985.

31 Kukich K, Techniques for Automatically Correcting Words in Text ACM Computing Surveys, 24:4:377-439, 1992.

32 Grefenstette G, and Tapanainen P. What is a word, What is a sentence? Problems of Tokenization. Rank Xerox Research Centre, Grenoble Laboratory, France, 1994.

Allen C. Browne

Information Research Specialist

National Library of Medicine
8600 Rockville Pike,
Bethesda, Maryland 20894

Education:

Georgetown University, Washington D.C.	B.S.	1975-1977	Linguistics
Georgetown University, Washington D.C.	M.S.	1977-1979	Sociolinguistics
Georgetown University, Washington D.C.		1979-1980	Sociolinguistics

Experience:

Computational Linguist, Lister Hill National Center for Biomedical Communications, National Library of Medicine, 1987 - present

Consultant on Grammar Analysis, IBM Watson Research Center 1985.

Document Processor, Informatics General Corporation, Health and Information Services Division 1984-1985.

Research Assistant, Association of American Universities, Resources for Language and Area Studies Project, 1984-1985

Instructor in Linguistics, Georgetown University,1979-1983

Fellowships and Grants:

Georgetown University Fellowship, 1979-1982.

Publications:

"The Lexical Properties of the Gene Ontology (GO)", A. T. McCray, A. C. Browne and O. Bodenreider, Proceedings of the The AMIA 2002 Symposium, pp 504-508.

"Evaluating UMLS Strings for Natural Language Processing" by A. T. McCray, O. Bodenrieder, J.D. Malley and Allen C. Browne, Proceedings of the The AMIA 2001 Symposium, pp 448-452.

“Lexicon assistance reduces manual verification of OCR output” by S.E. Hauser, A.C. Browne, G.R. Thoma, and A.T. McCray. in the Proceedings of the 11th IEEE Symposium on Computer Based Medical Systems. 1998 pp. 90-95.

“A Modular Text Processing System Based On the SPECIALIST Lexicon and Lexical Tools” by A.C. Browne, G. Divita, V. Nguyen, and V.C. Cheng. Poster presented at the AMIA 1998 Annual Symposium, November 7-11, Orlando FL.

“Evaluating Lexical Variant Generation to Improve Information Retrieval” by G. Divita, A.C. Browne, and T.C. Rindflesch. Proceedings of the AMIA 1998 Annual Symposium, November 7-11, Orlando FL.

“The UMLS Knowledge Source Server: A versitile Internet-Based Research Tool” by A.T. McCray, A.M. Razi, A.K. Bangalore, A.C. Browne, P.Z. Stavri. In Proceedings of the 1996 AMIA Fall Symposium. 1996

“Lexical Methods for Managing Variation in Biomedical Terminologies” by A.T. McCray, S. Srinivasan and A.C. Browne Proceedings of the Thirteenth Annual Symposium on Computer Applications in Medical Care. 1994.

“Exploiting a large thesaurus for information retrieval” by A.R. Aronson, T.C. Rindflesch and A.C. Browne. In Proceedings of RIAO. 1994. (with A. Aronson and T. Rindflesch)

“UMLS knowledge for biomedical language processing” by A.T. McCray, A.R. Aronson, A.C. Browne, A. Razi, T.C. Rindflesch, S. Srinivasan In Bulletin of the Medical Library Association 81. 1993.

The SPECIALIST Lexicon, NLM Technical Report No.: NLM-LHC-93-1. by A.C. Browne National Library of Medicine. 1993.

The SPECIALIST Natural Language Processing System. by A.T. McCray, A.C. Browne, S. Srinivasan, A.R. Aronson, T.A. Waldspurger, and I. Pufahl NLM Technical Report No.: NLM-LHC- 90-02, National Library of Medicine, 1990. (with A. McCray, S. Srinivasan, A. Aronson, T. Waldspurger and I. Pufahl)

“The Semantic Structure of Neo-Classical Compounds” by A.T. McCray A.C. Browne and D..L. Moore in the Proceedings of the Twelfth Annual Symposium on Computer Applications in Medical Care 1988

“The Role of the Lexical Knowledge in Biomedical Text Understanding” by A.T. McCray, J.L. Sponsler, B. Brylawski, and A.C. Browne in the Proceedings of the Eleventh Annual Symposium on Computer Applications in Medical Care. 1987

“Univocal Or -- Again” by A.C. Browne. In Linguistic Inquiry, Volume 17, Number 4, Fall 1986.

“Inductive Inference and Pragmatic Explanation” by A.C. Browne. In Proceedings of the 18th Regional Meeting of The Chicago Linguistics Society. 1982

Guy Divita

Computer Scientist

under contract from *Management Systems Designers, Inc.*

Lister Hill National Center for Biomedical Communications

National Library of Medicine

8600 Rockville Pike

Bethesda, MD 20894

divita@nlm.nih.gov

(w) (301) 435-3249

Mr. Divita works for the National Library of Medicine's (NLM) Lister Hill Center as a member of the Natural Language Systems group and a member of the Lexical Systems group. Mr. Divita has experience in natural language processing applied to information retrieval systems. He has implemented and now maintains the SPECIALIST NLP tools and text-to-concept mapping tools. He has researched and developed multi-strategy spelling suggestion algorithms. He has researched text normalization techniques for improving precision and recall. Mr. Divita has broad experience with the Unified Medical Language System (UMLS) and is part of the team that maintains one of the UMLS knowledge sources, the SPECIALIST Lexicon. He built and maintained the suite of lexical tools that is used to index the Metathesaurus, a concept centric thesaurus composed of a variety of controlled medical vocabularies. Mr. Divita has 18 years experience in application development, in Java and C on Unix and PC platforms.

Mr. Divita earned an MS in Computer Science from the George Mason University (1992) and a BS in Computer and Information Systems from the statistics department from The George Washington University (1985).

Mr. Divita has worked for the last nine years as contractor from Management Systems Designers, Inc. to NLM. (3/94-Present). Prior to his tenure at the Library, Mr. Divita worked as a Senior Software Analyst for Intergraph, Inc (1990-1994). While working for Intergraph, he helped develop two large scale GIS products, an environmental resource management system and an electronic nautical chart maintenance system. While working on the chart maintenance system, he was a team leader of a group of four developers.

Mr. Divita started his career as a statistician under contract to the NHTSA for Automated Sciences Group (1985-1990). He worked on congressionally mandated studies of accidents involving heavy trucks. He worked as a systems administrator and application developer and manager of systems administrators while at NHTSA. He oversaw the system administration of the machines used for the Auto Safety Hotline.

Publications

"A Spelling Suggestion Technique for Terminology Servers." **Guy Divita**, Allen C. Browne, Tony Tse, May L. Cheh, Russell F. Loane. A Poster. 2000 AMIA Fall Symposium, Los Angeles, CA.

"Analysis of Biomedical Text for Chemical Names: A Comparison of Three Methods." Dr. W. John Wilbur, Dr. George F. Hazard, Jr., **Guy Divita**, James G. Mork , Dr. Alan R. Aronson and Allen C. Browne. 1999 AMIA Fall Symposium Proceedings, 176-180.

"Evaluating Lexical Variant Generation to Improve Information Retrieval." **Guy Divita**, Allen C. Browne, and Thomas C. Rindflesch. In Chute, C. (ed) Proceedings of the AMIA Fall Symposium, 775-779. 1998.

"A Modular Text Processing System Based On the SPECIALIST Lexicon and Lexical Tools." A Poster. Allen C. Browne, **Guy Divita**, Van Nguyen, and Vincent C. Cheng. In Chute, C. (ed) Proceedings of the AMIA Fall Symposium, 982, 1998.

"Conducting the NLM/AHCPR Large Scale Vocabulary Test: A Distributed Internet Based Experiment." Dr. Alexa T. McCray, May L. Cheh, Anantha K. Bangalore, P. Zoe Stavri, Amir M. Razi, and **Guy Divita**. In Masys, D. (ed.) Proceedings of the AMIA Fall Symposium, 560-564, 1997.

Professional memberships

Association for Computational Linguistics (1997-present)

American Medical Informatics Association (1996-1998,2003)

Recent Courses Completed

Seminar in the Organization of Knowledge. University of Maryland, College of Information Studies (LBSC 772). Used Sowa's Knowledge Representation as the text book. Instructor: Rebecca Green. (Summer, 2001)

Advanced NLP: Theory and Practice. University of Maryland, Computer Science School. (CMSC 828) Instructors: Bonnie J. Dorr, Rebecca Hwa. (Spring, 2002)

Text Summarization, Georgetown University, Linguistics Department (Ling 461) Instructor: Inderjeet Mani. (Fall, 2002)

Senior System Architect, Chris J. Lu Ph.D.

Summary

Strong skill set includes: tuning software development process, architecture specification, analysis, design, implementation, and mentoring. Excellent engineering background and abundant experience in integrating hardware/software.

More than ten years experience in system integration, simulation modeling, computer graphics, GUI libraries design, API/CASE tool, and Web applications development. Have more than 40 technical research publications in international journals, conference proceedings, and technical reports, such as in IEEE, ASME, and IASTED. Areas of expertise include Web applications, natural language processing, virtual prototyping system development, large-scale system automation, full life cycle methodology for large scale systems, and 3-D real time simulation.

Education

Ph.D., Computer Integrated Manufacturing and Design, ME, University of Maryland, 1995

M. S., Computer Integrated Manufacturing and Design, ME, University of Maryland, 1988

B. S., Computer Integrated Manufacturing and Design, ME, National Taiwan University, 1984

Summary of Computing Skills

Languages:	Java, C, C++, PROC, MS Visual C++, JSP, JavaScript, and HTML
Hardware:	Hardware A, Hardware B, Hardware C
OS:	UNIX, Window NT, and Window 2000
Software:	MySql, Oracle, InstantDB, MS Access, SCRT, etc..

Professional Experience

August, 2003 – Present **MSD, Inc.**

Senior System Architect

Works on THE SPECIALIST Lexicon project at the National Library of Medicine. My major responsibilities include the annual release and technical support of Lexical Tools, maintain Lexical Tools web sites, and developing LexBuild system.

April, 2002 – August, 2003 **Aquilent, Inc.**

Senior System Architect

Served as a senior system architect and responsible for several natural language projects at National Library of Medicine (NLM). My responsibilities included performing feasibility study on new project, defining project scope, and documenting requirements, designing system architecture, establishing software development processes, implementing the core library, testing, deploying, and maintain the developed system. Major projects include Lexical Tools, LexCheck, and LexBuild. These projects are web based application developed in **Java** using **Servlets, JSP, JDBC** with **MySql** and **Instant DB**, and run on **UNIX** (**Sun Solaris, Apache and TomCat web server**).

August, 1998 – March, 2002 **Commerce One, Global Services**

Senior System Architect

Worked as a senior system architect in several projects at National Library of Medicine (NLM). My responsibilities included defining and exercising all phases of software development life cycle. Major projects include Lexical Tools, UMLS-Assistant System (UA), Software Change Request Tool (SCRT), and Demo Server tool (DS). All these systems are web based applications and developed in **Java** using **Servlets**, **Applets, CGI, JDBC** with **MySql** and **Instant DB**, and run on **UNIX** (**Sun Solaris, Apache web server**).

March, 1997 – July, 1998 **TRW, System Integration Group**

Senior Software Engineer

Served as a senior software engineer in Earth Data Observation System (EDOS) project. My tasks include software development and technical support for integration testing on varieties of computer software components (CSCs). These CSCs are developed in **C** and run on **UNIX** (**AIX** and **SGI**).

September, 1988 **–** March, 1997 **Advanced Technology & Research Corporation**

Senior System/Software Engineer

Served as a senior technical leader on the projects of a Virtual Prototyping Tool. This tool is a 3-D CASE tool employing **Object-Oriented** method provides a virtual testbed to validate control software of automated systems. This tool was developed in **C**/**C++** and ran on **UNIX**/**Window NT** platforms, respectively. USPS (United States Postal Service) utilized this tool to develop and analyze the material handling system of Mail Consolidating Center (MCC) at Harrisburg, PA. In addition, NIST (National Institute of Standards and Technology) and GM (General Motor) used this tool to develop control software of machine tool (KT-800).

Planning Manager/Software Architect

Managed international projects for the development of Visualization, Analysis, and Simulation Tool (VAST). These projects were funded by Tjing-Ling Industrial Research Institute and assigned to NTU (National Taiwan Univ.), UMCP (Univ. of Maryland, College Park), and ATR. This tool was successfully developed in **C** on **UNIX** (SGI**)** and presented in IRF Conference.

Research Engineer

Served as a research engineer in the development of a seamless, life-cycle tool set - CSAT (Control, Simulation, and Analysis Tool), for USPS. This tool set provided a powerful system for the design, simulation, analysis, training, and maintenance for large scale systems. This tool was developed in **C** and ran on **UNIX** (SGI) platform. USPS used this tool to design and analyze facilities at Richmond, VA; San-Diego, CA; Sacramento, CA; and Kansas, KS. My major responsibilities and technical accomplishments included:

September, 1986 - August, 1988 **University of Maryland, College Park**

Research Assistant

Developed advanced numerical algorithms, Cheater Homotopy and Generic Homotopy methods, to solve the classic Forward and Inverse Kinematics problems, such as Puma Robot design, path generation, and coupler-point curve synthesis. These algorithms were implemented in **FORTAN** on **VAX** 750 platform.

Lynn A. McCreedy
426 North Patrick Street
Alexandria, Virginia 22314
(703) 549-3845 (home)

Professional Objective

To conduct independent research and participate in team research projects on linguistic issues. My principal analytical focus is on the analysis of texts and face-to-face interaction, to elucidate the relevance of specific aspects of linguistic code to their more general uses on stylistic, discourse-organizational and social levels.

Education

Ph.D. in sociolinguistics, Georgetown University, May 1983. Dissertation, titled "Aspects of Reference, Cohesion, and Style in Three Genres of Navajo Texts," defended "with distinction," December 1982. Comprehensive examinations passed "with distinction," fall 1978. Major in sociolinguistics, minors in theoretical linguistics and anthropology. **M.S. in Linguistics, Georgetown University**, May 1977. Major in sociolinguistics, minor in general linguistics. **B.A. cum laude, Western Michigan University**, April 1973. Majors in anthropology and German, minor in linguistics. Participated in General Education Honors Program.

Writing/Research

Computational Lexicography: Analysis and encoding of lexical items, considering morphological, syntactic and graphemic information for incorporation into a natural language processing system, which is designed to capture both the regularities of general English and specialized biomedical text. The resulting lexicon, a major knowledge source of the Unified Medical Language System, is perhaps the largest such database currently available for natural language processing, and is used by both biomedical and natural language processing professionals. August 1996-present; June-October, 1995.

Classroom Discourse Analysis: As a Research Associate of the University of Maryland's Institute for the study of Exceptional Children and Youth, in the Department of Special Education, I conducted a micro-analysis of video- and audiotaped classroom interaction in special education and regular elementary classes from an urban school system in the eastern U.S. The analysis identified aspects of academic talk and students' dialect shifting. In collaboration with Dr. Carolyn Adger of the University of Maryland and Ms. Jennifer Detwyler of the Center for Applied Linguistics, Washington, D.C. March-August, 1993.

Classroom Q/A Analysis: Designed and conducted a detailed discourse-interactional analysis of teacher/student communication in Q/A sequences during elementary math and science lessons. Other researchers on this project were Dr. Carmen Simich-Dudgeon (principal investigator, currently at Indiana University) and Dr. Mary Schleppegrell (currently at UC Davis); a project of the Center for Language Education and Research (CLEAR), at the Center for Applied Linguistics, Washington, D.C. Concluded May 31, 1988.

Publications and Conference Presentations:

"The Effect of Role and Footing in Classroom Verbal Interaction on Students' Oral Academic Language." In *Kids Talk: Language Practices of Older Children*, edited by S. Hoyle and C. Adger. New York: Oxford University Press, 1998.

"Promoting the Development of Oral Academic Language Competence: the Value of Cooperative Learning."

Coauthor with Jennifer Detwyler. Presented at the presession of the Georgetown University Round Table on Languages and Linguistics, Washington, D.C., March, 1993.

"Conducting Verbal Reviews," in *The Multicultural Classroom: Readings for Content-Area Teachers*, edited by P.A. Richard-Amato and M. A. Snow. London: Longman, 1992. Coauthor with Carmen Simich-Dudgeon and Mary Schleppegrell.

"An Educology of Classroom Discourse: How Teachers Produce Coherence in Classroom Discourse Through Managing Topics, Interactive Tasks and Students," *International Journal of Educology* vol. 4, no. 2: 1990. Coauthor with Carmen Simich-Dudgeon.

"Cohesion and Discourse Structure in Three Genres of Navajo Discourse," in *Studies in Athabaskan Linguistics*, edited by E.D. Cook and K. Rice. The Hague: Mouton de Gruyter, 1988.

Helping Limited English Proficient Children Communicate in the Classroom: A Handbook for Teachers. Coauthor with Carmen Simich-Dudgeon and Mary Schleppegrell. National Clearinghouse for Bilingual Education, Program Information Guide Series No. 9, Winter 1988/89.
"Managing Topics, Interactive Tasks, and Students: How Teachers Produce Coherence in Classroom Discourse." Coauthor with Carmen Simich-Dudgeon. Poster presentation at the Annual Meeting of the American Educational Research Association, New Orleans, Louisiana, April 1988.

"Showing Off Your Smarts: Displaying Verbal Knowledge During Math and Science Lessons in Elementary School." Coauthor with Carmen Simich-Dudgeon. Presented at the Annual Meeting of the American Educational Research Association, Washington, DC, April 1987.

"Academic Language Talk: Metacognitive Awareness of Teachers and Students." Coauthor with Carmen Simich-Dudgeon. Presented at the Annual Meeting of the American Anthropological Association, Philadelphia, PA, November 1986.

"Pronouns and Style in Navajo Narratives." Presented at the Annual Meeting of the American Anthropological Association, Denver, CO, November 1984.
of the Linguistic Society of America, July 1982.

Teaching Experience

Lecturer of English Linguistics, George Mason University. Summers 1985, 1986, 1989; spring 1986.
Linguistics Instructor in the English Department of Goucher College, Towson MD. Spring 1983.

Professional Organizations

Linguistic Society of America, Society for the Study of the Indigenous Languages of the Americas.

DESTINEE NACE TORMEY
226A Harrison Lane
Princeton, New Jersey 08540
E-mail: dln4@georgetown.edu

EDUCATION

May 2003	M.S., Georgetown University, Computational Linguistics Master's Research Paper Title: "A Corpus Analysis of Levin's Verb Classes and Alternations for *Bring* and *Take*."
May 1999	B.A., magna cum laude, Millersville University of Pennsylvania, French and English

WORK EXPERIENCE

July 2002 to present	Lexical Developer for the SPECIALIST Lexicon, National Library of Medicine, National Institutes for Health
August 2001 to July 2003	Coordinator for the Georgetown University Writing Program, Department of English, Georgetown University
August 1999 to August 2001	Typesetter (11/99 to 8/01) and Proofreader (8/99-11/99), Merrill Corporation, San Francisco, California

CURRENT RESEARCH INTERESTS

- Statistical Natural Language Processing
- Part-of-speech tagging
- Lexical Semantics
- Corpus Linguistics

COMPUTER LANGUAGES

- Perl
- C
- Java (currently learning)

FOREIGN LANGUAGES

- French: fluent
- Spanish: reading and spoken knowledge

PROFESSIONAL ORGANIZATIONS

- Association for Computational Linguistics

FELLOWSHIPS AND AWARDS

- Georgetown Fellowship, awarded 2002
- Ralph J. Hyson Memorial Award for excellence in French studies, 1999
- Frank R. Heavner Memorial Award for excellence in Linguistics, 1999
- Student Research Grant, 1998
- Melville Scholarship, 1995-1999
- SICO Scholarship, 1995-1999