Researchers have studied lexical phenomena for years. Here, the vital role of lexicons, a key component of NLP systems, is traced through a variety of projects.

Louise Guthrie

James Pustejovsky

Yorick Wilks

Brian M. Slator

The Role of Lexicons in Natural Language Processing

Dictionaries and computation are two subjects not often brought together in the same article nor even the same proposition. This article explores the growing relations between these two entities and, in particular, investigates whether what is found in traditional dictionaries can be of service to those concerned with getting computers to process and understand natural human languages.

Patrick Suppes used to say he would be interested in machine translation only when programs could process a whole book. In fact, they already could and did when he said that 25 years ago, but his remark was in the right spirit. Although machine translation, by methods now considered superficial by most who work in artificial intelligence (AI), processed large volumes of text by the mid-1960s, AI researchers who aimed for something deeper—for a modeling of what most people would call "understanding" a text—wrote programs that processed only a few sentences (a fact often disguised in their theses, papers, and books.) In a moment of great honesty five years ago, a group of AI researchers of natural language processing (NLP) admitted in public (in answer to a question by Bran Boguraev) how many words there really were in the vocabularies of their systems [23]. Of the answers, the average was 36, a figure often taken to be a misprint when it appears, though it was all too true. Remedies for this situation, which had little to do with science or engineering as normally understood, have been to move to a larger scale and

more empirical methods: These have ranged from the use of neural or connectionist networks in language understanding to a return to statistical methods for the analysis of texts, as well as attempts to increase the size of lexicons for computational systems by attempting to extract meaning (together with syntactic and other kinds of grammatical information) from existing dictionaries available in machine-readable form (machine-readable dictionaries, henceforth MRDs).

But did MRDs have content that could be made available in some automatic manner to computational linguistic programs operating on texts? One could also wonder, "Can existing dictionaries yield information that helps understand texts with computers?" where "yield" has been taken to mean subjecting the dictionary to the same kind of analysis (or parsing) techniques that were applied to ordinary texts, so as to get information from them. In what follows, "dictionary" has its typical meaning of a printed wordbook for human readers, which may also be available as an MRD. "Lexicon" will refer to the component of a NLP system that contains information (semantic, grammatical) about individual words or word strings.

Almost everyone concedes that dictionaries contain a great deal of knowledge about the world, and that the distinction between knowledge of language and knowledge of the world is shaky at best. But the basic question remains as to whether information that can be extracted from dictionaries, automatically or by humans, is also the kind that meets the AI demand for computer-usable, or tractable, world knowledge.

By a curious historical serendipity, computational interest in dictionaries revived at just the moment when some publishers, particularly in Britain, were making dictionary formats—the formats for their readers, that is, not just the format on computer tapes—more purely formalized in regard to both syntax and the vocabulary used to define word senses. Dictionaries like the *Longman Dictionary of Contemporary English* (LDOCE, published by Longman Group Ltd. in 1978 and reissued subsequently) and later the *Oxford Advanced Learner's Dictionary of Current English* (OALD, first published by Oxford University Press in 1942) and *Collins' COBUILD Dictionary of the English Language,* (first published in 1987) seem almost to have been designed for computational linguistic investigation, despite the fact their radically new formats were designed for learners of English with only limited facility in the language.

Early Work

Dictionaries are special texts whose subject matter is a language, or a pair of languages in the case of a bilingual dictionary. The purpose of dictionaries is to provide a wide range of information about words—etymology, pronunciation, stress, morphology, syntax, register—to give definitions of senses of words, and, in so doing, to supply knowledge not just about language, but about the world itself.

The earliest attempts to compute over a whole dictionary of substantial size were made by Olney and Revard in the group headed by Bob Simmons at Systems Development Corp. in the late 1960s. *Webster's Seventh New Collegiate Dictionary* was keypunched twice (to avoid errors) onto paper tape from paper text. These first efforts were therefore almost certainly the only ones for which the dictionary was put in machine-readable form primarily for the purpose of computational exploration, as opposed to the work's being a mere by-product of the existence of the printer's machine-readable form, as is now universally the case. What they did with that vast amount of information was chiefly to explore word frequencies in definitions, itself an enormous computing task with the hardware resources then available.

The first structural work in computational lexicology was done by Amsler and White [1], under the direction of Simmons, and hence has strong continuity with the work just described. What Amsler and White did was to establish, partly by hand but by potentially computable methods, a taxonomic structure for a representative subset of the *New Merriam-Webster Pocket Dictionary*, published by Pocket Books. They were the first to point out, in a clear, procedural way, that dictionaries are ideally structured for taxonomic organization, since the defining words themselves all appear in the headword list that are the dictionary itself (a fact that had long been noticed for thesauri, as we will discuss). This work used the existing division of wordsenses in Webster's Seventh to define a taxonomy—work that in some sense was the dual of earlier Sparck Jones work that used an existing taxonomy to define senses.

Sparck Jones [21] described a thesaurus as classifying word usages by their appearance under main heads, of which there are about 1,000 in *Roget's Thesaurus*, which were themselves hierarchically classified by him. She broadly identified a word's appearance under a head as pinning down a sense of it (though this can then become an initial definition of sense, not a description).

Sparck Jones took classic information retrieval (IR) techniques, applied them to the thesaurus and produced "clumps" of words with common features. This was ground-breaking work in empirical semantics, even if it rested on an unempirical initial notion of synonymy (which could now be remedied) and suffered the abiding problem of all automatic classification techniques; that the clumps, when found, did not have names. Sparck Jones' work in 1964 was the

first explicit link from IR to NLP, a link that was then buried until her dissertation was finally published 20 years after it was written. The relevance of this work for NLP is that it can be seen as establishing one empirical basis for the crucial notion of wordsense, a matter that has remained highly disputable.

Taxonomies Implicit in Dictionary Definitions

The type of information extraction pioneered by Amsler depended on the fact that dictionary definitions of nouns are normally written in such a way that one can identify, for each headword (the word being defined), a *genus term* (a word more general than the headword), which is related via an IS-A relation. The information following the genus term, the *differentia,* serves to differentiate the headword from other headwords with the same genus term. For LDOCE:

> *knife—a blade fixed in a handle, used for cutting as a tool or weapon.*

Here "blade" is the genus term of the headword "knife," and "fixed in a handle, used for cutting as a tool or weapon" yields the differentia. In other words, a "knife IS-A blade" (genus) distinguished from other blades by the features of its differentia. The standard technique for creating a semantic network from noun definitions is to identify the IS-A relationships of this type. But for this information to be useful, it is necessary to distinguish the sense of "blade" used in the network from other possible senses (by ruling out a blade of grass, a propeller blade, and a sharp amusing fellow). We can view the process of building a semantic hierarchy of IS-A links from noun definitions as twofold: (1) to find the genus term (or terms) in the definitions and (2) to disambiguate it.

Amsler and White [1] began with the concordances of Olney (created at the SDC Corp. in 1969) and an implementation of a tree-growing algorithm by Amsler. They worked with computers much less powerful than the desktops of today to create a semantic network (sometimes called a "tangled hierarchy") of noun senses from the *New Merriam Webster Pocket Dictionary* (published in 1964). Their techniques were partially automated, but required a great deal of human intervention in the disambiguation process. Nonetheless, their work was among the most diligent and complete in the field, and descriptions of it have provided the backbone of all that followed. Unfortunately, the actual network of word senses never became publicly available for further research. Other researchers in NLP began to replicate their work for other dictionaries, using better computers and more automation. Amsler coined the term "computational lexicology" for this new generation of research, and showed the research community that information useful for NLP could be extracted from MRDs.

The 1980s saw a flurry of work that automated the construction of semantic networks using the machine-readable version of LDOCE. LDOCE was a very appealing source of information for researchers, and this, in addition to the cooperation of Longman's with the research community, has made it the most widely used English dictionary for language processing.

LDOCE is a full-sized dictionary designed for learners of English as a second language. It contains over 76,000 sense definitions based on nearly 40,000 spelling forms. The definitions have a restricted syntax and make use of only 2,000 words from the dictionary itself. The machine-readable version of LDOCE also contains "box" and "subject" codes not found in the book version. The box codes use a set of primitives such as "abstract," "concrete," and "animate," organized into a type hierarchy, which are used to assign type restrictions on nouns and adjectives, as well as on the arguments of verbs.

The subject codes use another set of primitives organized in a hierarchy consisting of main headings, such as "engineering," with subheadings, like "electrical." These primitives are used to classify words by their subject; for example, one sense of "current" is classified as "geology-and-geography," while another sense is marked "engineering/electrical."

Extracting Information from LDOCE

Noel and the group at the University of Liege (including, at various times, Michiels, Fontenelle, Mullenders, and Vanandroye) have done extensive studies of the structure of LDOCE (see, for example, Michiels, Mullenders, and Noel in [5]). Michiels's 1982 thesis was the first large-scale exploration of the LDOCE database. Parsers have been developed that parse the definition text of LDOCE (see, for example, Vossen; Wilks, Fass, Guo, McDonald, Plate and Slator, in [9]).

There have been successful efforts, with various dictionaries, to create networks from the genus terms of noun definitions in an automatic way, but LDOCE is the only English dictionary that has been used for disambiguating the terms of the network automatically.

Nakamura and Nagao in [6] constructed a network of spelling forms from LDOCE automatically and created a database of LDOCE definitions. Although they did not attempt automatic disambiguation of the terms, they identified patterns of words in the definitions and created links corresponding to them.

Using the semantic category markings in the machine-readable version of LDOCE, the NMSU group (for example, Guthrie, Slator, Wilks, and Bruce in [7]; Bruce and Guthrie in [8]) developed automatic techniques for selecting the genus term in an LDOCE definition and disambiguating it relative to the senses in LDOCE. A hierarchy of 39,000 nouns

and phrases defined in LDOCE was constructed using these results. Analysis of this hierarchy shows that it is relatively shallow (the median depth of a node is two levels down). The Cambridge group (see Alshawi; Boguraev and Briscoe in [9]), developed similar, but not identical, heuristic procedures that, to a great extent, automate the task of developing a hierarchy of word senses.

The ACQUILEX project in Europe (see Calzolari and Bindi in [7]) is developing disambiguated networks for a range of dictionaries and has provided extensive studies on the structure of LDOCE. Vossen in [9] has created complete networks of noun senses for both LDOCE and the Dutch Van Dale dictionary using a technique for disambiguation that combines information from both dictionaries with information from the Van Dale bilingual Dutch-English dictionary. Copestake and Briscoe (personal communication), at Cambridge University use heuristics based on the structure of LDOCE together with information from the *Longman Lexicon of Contemporary English* (published in 1981) in their disambiguation algorithm.

Building a Lexical Database

The appeal of using on-line dictionaries in the construction of formal computational lexicons is intuitively obvious: dictionaries contain information about words, and lexicons need such information. If automated procedures could be developed for extracting and formalizing lexical data on a large scale from existing on-line resources, NLP systems would have ways of capitalizing on much of the lexicographic effort embodied in the production of reference materials for human consumption.

Not surprisingly, one approach to scaling up the lexical components of natural language systems prototypes to enable them to handle realistic texts has been to turn to existing machine-readable forms of published dictionaries. On the assumption that they not only represent (trivially) a convenient source of words but also contain (in a less obvious and more interesting way) a significant amount of lexical data, recent research efforts have shown that automated procedures can be developed for extracting and formalizing explicitly available, as well as implicitly encoded, information—phonological, syntactic, and semantic—from MRDs.

Research in the area reflects a change in view. Whereas early efforts for utilizing dictionary data were aimed primarily at what had been explicitly stated in the entries (see [22]), comparatively recent developments have focused on carrying out much more detailed analysis of the sources, with a view to uncovering information that turns out to be systematically, albeit implicitly, represented by dictionary entry content, dictionary structure, and lexicographic conventions.

Research has progressed:

- From simple word-list construction to part-of-speech extraction (e.g., noun, verb) and subcategory acquisition (e.g., common noun, transitive verb) (see Byrd, Calzolari, Chodorow, Klavans, Neff, and Rizk in [2]) and to acquiring information about control and logical type of predicates (see [2, 9]);
- From constructing simple taxonomies for verbs and nouns [1] to fleshing out semantic networks (see Alshawi in [9]) to building semantically sound lexical hierarchies (see Beckwith, Fellbaum, Gross, and Miller in [25]);
- From acquiring semantic features (e.g., selectional restrictions such as "animate" and "abstract") for lexical disambiguation (Byrd et al. in [2]) to deriving empirical evidence of the existence of semantically coherent word-sense clusters (see Slator in [25] and Wilks et al. in [16]); and
- From "sprouting" (in Byrd's parlance) networks of lexical relations between words (by creating genus links between word senses) to refining such networks to reflect word-sense distinctions (see Guthrie et al. in [7]) and to populating richer lexical structures that introduce an additional dimension to the notion of lexical relation and promote more flexible interpretation of the notion of word sense [17].

When the lexical information that has been collected by these or other means is presented in a structured format accessible by programs that process language, we refer to it as a lexical database.

Database Approaches to the Machine-Tractable Dictionary

Reports on the concept and construction of lexical databases began to appear in the early 1980s. A fundamental step in creating a lexical database from an MRD is the parsing of the dictionary entries (including the special codes, diacritical marks, and arcania that are a part of the special formating of the dictionary) to allow the information to be recorded in the appropriate data structure for subsequent processing. Since the MRDs were original tapes from which the paper versions of the dictionaries were printed, much of the information they contain consists of instruction designed to produce a document pleasing to the eye. The parsing of these MRDs turned out to be a formidable task

The parsing of the dictionary entries just described does not attempt to extract meaning from the string of words that constitute the text of the definition. The utilization of their semantic information is the next step in the construction of the lexical database. Large-scale lexical databases have been implemented by several researchers (see [11] and Nakamura and Nagao in [6]).

Byrd et al. in [2] describe a large inventory of tools for manipulating MRDs and give an impressive

list of MRDs over which these tools are to work. These include a database management scheme for handling lexical data, a menu-driven query interface to the databases, tools to build network structure from definitions (the sprouting mentioned earlier), utilities for finding recurring phrases in text, and morphological analyzers for definition text. Extensive study of the structure of MRD entries was also carried out at IBM, resulting in the Dictionary Entry Parser (DEP) system (see Boguraev, Byrd, Klavans, and Neff in [25]). The purpose of this device is to transduce MRD entries into a lexical database without loss of information.

The Illinois school of Evens and colleagues has concentrated on analysis (collected in [11]) of Webster's Seventh, parsing it originally with the Linguistic String Parser of Sager and then with analysis programs of their own design. Aside from "taxonomy and set-membership relations," the Illinois school has been characterized by the study of "defining formulae" and the construction of what they call the "relational lexicon." In [11], they describe the design and implementation of a project to convert an MRD into a lexical database (LDB) in relational database form. They report that in 1988 their LDB contained syntactic and semantic information for about 50,000 words and was implemented using commercial database software (the Oracle Relational Database Management System)

The New Mexico State University (NMSU) lexical research program combines the following:

1. A system for creating a (statistically based) Pathfinder network (see, for example, Fowler and Dearholt in [18]) of clustered word senses, in which the clusters are based on cooccurrence statistics;
2. A hand-coding initiative due to Guo (see [12]), so as to find the defining senses of the dictionary; and
3. A dictionary parsing system [19] that processes definitions as well as using the semantic and pragmatic codes of LDOCE to disambiguate the genus terms of definitions.

These three parallel efforts pursued at NMSU all take as input the forms of information given on the LDOCE tape (English definitions, syntax codes, and subject and pragmatic codes, among others) and provide, respectively:

1. A clustered network of LDOCE words whose clusters correspond to empirically derived senses (see McDonald, Plate, and Schvaneveldt in [18]);
2. A formalized set of definitions of sense entries in a nested predicate form, where the predicates are a "seed set" of senses, half the size of the existing controlled vocabulary of LDOCE. This would be a Fregean compositional formalization of LDOCE [12]; and
3. Frame-like structures containing (in addition to the LDOCE syntactic codes) a formalization of the English definitions using predicates that are English words (not senses) from the controlled vocabulary [19].

The appeal of using on-line dictionaries in the construction of formal computational lexicons is intuitively obvious: **dictionaries contain information about words, and lexicons need such information.**

The major difficulty in making LDOCE tractable is the fact that ambiguous English words comprise the definitions that must be tagged explicitly with their senses in LDOCE definitions (as in the knife and blade example). The SPIRAL procedure [24] cycles information through Guo's seed senses, Slator's LDOCE parser, the Genus Disambiguator, and Plate and McDonald's distributional network so as to yield a sense-tagging of the definition words in frames.

Recent projects include the construction of a lexical database of lexical semantic objects. The system, based on the text-specific lexicon provider of Slator, allows complicated queries and provides data to a workbench for lexicon builders, which allows lexicons to be created in the format needed for a particular system.

Several dictionary-related projects have been undertaken in the Netherlands, originally under the auspices of the Automatic Scanning system for Corpus Oriented Tasks (ASCOT) project at the University of Amsterdam. Akkerman in [9] made a very thorough and revealing comparison between the grammar coding schemes of LDOCE and OALD. These two MRDs have the most elaborate grammatical information of any on-line resource, and it turns

out that the bulk of their codings can be mutually translated back and forth. Other projects that make use of the ASCOT lexicon include TOCSA (ASCOT reversed), which builds a grammar of English using the Extended Affix grammar formalism, and the PARSCOT project which aims to reimplement the Linguistic String Project grammar of Naomi Sager.

Vossen, Meijs, and den Broeder in [9] describe how they participate in the LINKS project, a research program using LDOCE that has been pursued in various forms at the University of Amsterdam since 1986. The goal of the project is the construction of a semantic database intended to serve the needs of various other systems for parsing and text analysis. The NMSU SPIRAL project had similar aims [13, 20].

Lexical Disambiguation

Lexical ambiguity is pervasive in most forms of text, including dictionary definitions themselves, as we noted. For those engaged in transforming MRDs into machine-tractable dictionaries (MTDs), it follows that the language in MRD definitions needs to be disambiguated and interpreted at the word-sense level before we can benefit fully from the information implicit in them. Here we describe an NMSU approach to this in lexical definitions, that involves tagging the words with the appropriate sense from the dictionary (LDOCE) itself.

Researchers in NLP and information retrieval need automatic methods of disambiguation, considered more generally. In machine translation, for example, one must disambiguate the input text in order to yield a correct translation to another language. Information retrieval systems might be more effective if they were able to disambiguate the words in a query and in stored documents, and corpus analysis and lexicography could become more automated if words in the corpora were disambiguated.

The concept of disambiguation, however, is nebulous. Humans often cannot agree about which of a given collection of senses is being used in a particular sentence. Lexicographers themselves do not agree about the number of senses for a given word, nor about the way a word's use should be divided into senses. Research to automate the task of disambiguation (for example, Cowie, Guthrie and Guthrie; Yarowsky in [8]; Dagan, Itai, and Schwall; Guthrie, Guthrie, Wilks, and Aidinejad in [4]; McDonald, Plate, and Schvaneveldt in [18]; Veronis and Ide; and Zernik and Jacobs in [7]; and Wilks et al. in [16]), even though it has received much attention recently, is still in a primitive stage and the granularity of the senses used in experiments varies. Senses are often chosen in an ad hoc way, usually with fewer and broader senses than would be found in most standard dictionaries. For these experiments, it suffices to make sense distinctions adequate to the purpose (to sufficiently capture the distinctions in the texts being processed), which usually means eliminating the anachronistic or rare senses often found in standard dictionaries.

Several effective procedures have been reported for the lexical disambiguation necessary for choosing the correct translation of a word in general (and not just dictionary) texts. Brown et al. in [4], among others, derive the set of translations of a given word, as well as a model of the corresponding context, for each translation from automatically aligned bilingual corpora in French and English. This information is then used to determine the correct translation of the word in a new text. This definition of senses is certainly well motivated and sufficient for automatically disambiguating words with respect to word forms in another language, but the data does depend on the existence of large parallel texts of that type and seems to be applicable only to machine translation, where equivalence in another language can serve as an effective sense specification. Dagan, Itai, and Schwall in [4] report similar experiments for the translation of German and Hebrew into English.

Yarowsky in [8] reports a disambiguation scheme that sense-tags words with their corresponding Roget categories (one of 1,042 categories). For each category, contexts representative of that category are gathered from the 10-million-word Grolier's Encyclopedia (50 words on each side of a word in the category). Salient words are identified statistically, and weights are determined for these salient words. The weights are then used to predict the category of a polysemous word in a new text.

Another technique for disambiguation, reported by Zernik and Jacobs in [7], relies on training text that has been sense-tagged by hand; it also incorporates information about the word from its morphology and part-of-speech tagging. They report on experiments in which the three words "interest," "stock," and "bond" are disambiguated with respect to a restricted set of senses. For example, for a certain corpus, they are able to confidently parse with four senses of "interest," which is a significant reduction from the 10 entries LDOCE provides.

Many researchers have found that a standard dictionary, with its distinctions made by professional lexicographers, is still the most attractive option for use in disambiguation. The use of standard dictionaries for the purpose can be attributed to Lesk [14], who suggested how this might be done automatically, using the overlap of words between the definitions of the context words and the sense definitions of the word being disambiguated.

For example, to disambiguate "bank" in the sentence "We got a bank loan to buy a house," Lesk's

general method looks at all the sense definitions of "bank" in a standard dictionary. Each dictionary sense-definition of "bank" is then intersected with the union of all senses of the local context {we, got, a, loan, to, buy, house} of "bank," and the number of words in the overlap is counted. The sense definition with the largest overlap with the context is the winner.

Veronis and Ide in [7] describe a way to create very large neural networks automatically from MRDs to use for disambiguation. However, the experiments were actually very small and resulted in 23 disambiguated words in small hand-constructed contexts, relative to the *Collins English Dictionary*. Scalability of neural net approaches to NLP applications continues to be a very difficult proposition.

White in [11] describes a word sense disambiguation scheme that depends on a sort of "minimalist" rendering of the dictionary: The senses in the MRD are associated only with their taxonomic structure, as derived from the tangled hierarchy procedures mentioned, and collocationally with the words of their own definitions. With only these two structures, sense disambiguation in free text is accomplished by comparison to a collocational structure built for an unknown word, then employing a fairly simple metric for rating the candidates and choosing the best.

Several automatic methods for lexical disambiguation relative to a standard dictionary are under investigation at the Computing Research Laboratory (CRL) at NMSU. The goal of these methods is to disambiguate words in text relative to LDOCE. All three methods attempt to use the context of a word to determine its sense.

One class of techniques gathers information about word associations from text and then uses that information for disambiguation. This work uses the LDOCE definitions and example sentences themselves as a corpus of about one million words from which to gather information about word associations, but the methods are easily extensible to gathering information from larger corpora.

Words that occur frequently with a given word may be thought of as forming a "neighborhood" of that word. If we can determine which words (as spelling forms) cooccur frequently with each word sense, we can use these neighborhoods to disambiguate the word to its proper sense in a given text. Assume that we know only the two classic senses of the word "bank": A repository for money; and piled earth at the edge of a river.

We can expect the "money" sense of "bank" to cooccur frequently with such words as "money," "loan," and "robber," while the "river" sense would be more frequently associated with "river," "bridge," and "earth." In order to disambiguate "bank" in a text, we would produce neighborhoods for each sense and intersect them with the text, on the assumption that the neighborhood that shared more words with the text would determine the correct sense. Variations of this idea appear in [14], McDonald et al. in [18] and Veronis and Ide in [7].

The study of word cooccurrence in a text is based on the cliche (attributed to Firth) that "a word is known by the company it keeps." Guthrie et al. in [4] hold that it also makes a difference where that company is kept, since a word may occur with different sets of words in different contexts; with this in mind, they use the subject field code markings in LDOCE to construct word neighborhoods that depend on the subject of the text in question. These neighborhoods are then used to disambiguate words in much the same way as in the previous technique.

In this method, the senses are computed one word at a time, and the question arises of whether and how to incorporate the fact that a sense has been chosen for one word when attempting to disambiguate the next. Should this first choice be changed in light of how other word senses are selected? Although these problems were pointed out in Lesk's original paper, they have been addressed only in the method of Veronis and Ide in [7] and the following technique of Cowie, Guthrie and Guthrie in [8].

These authors describe the application to lexical disambiguation of the computational optimization technique of simulated annealing (where a probabilistic search gradually moves from "hot" nearly random transitions to "cool" nearby ones) using a basic method similar to that of Lesk [14], which also uses the subject area markings in LDOCE, but without making use of other features such as part-of-speech tagging. The simplicity of the technique makes it fully automatic, and it requires no hand-tagging of text or hand-crafting of neighborhoods. When this basic method operates under the guidance of the simulated annealing algorithm, sense selections are made concurrently for all ambiguous words in the sentence in a way designed to optimize their choice. The system's performance on a set of test sentences was encouraging and can be expected to improve when some refinements, such as part-of-speech tagging, are incorporated.

Creating Bilingual Lexicons for NLP Projects

Lexicons need to be bi- or even multilingual for many NLP projects, and the goal of automating the construction of bilingual NLP lexicons is being approached by several groups. Klavans and Tzoukermann in [7] have used a bilingual English-French, French-English dictionary (the Collins Robert) in conjunction with the Hansard corpus (the bilingual record of the Canadian Parliament) as the basis for the creation of a bilingual lexical database (BLDB). Their IBM BICORD system is designed to probe corpora for verb correspondences and to expand and rebuild the bilingual dictionary with collocation and frequency information. The end goal is to establish

lexical correspondences with associated translations and then attach frequencies to the translations for the purpose of mixed probabilistic and transfer-based machine translation.

Neff and McCord [15] use the Collins English-German (CEG) bilingual dictionary, in conjunction with other monolingual sources, for building frames automatically for IBM's lexical machine translation system. The frames are derived from CEG and augmented by frames taken by IBM's UDICT, a large encoded lexicon built over the years from several sources, including LDOCE. The transfer process includes sense disambiguation of the source terms (in English), selection of German target terms, and further representational augmentations taken from various lexical databases created from LDOCE, Webster's Seventh, and the Collins Synonym Dictionary.

Research is being conducted at the NMSU CRL toward automatically creating individual language lexicons for the ULTRA and PANGLOSS interlingual machine translation systems. To a great extent, they have already succeeded in automating the construction of one type of lexical entry in the system, those of the interlingual, or concept, lexicon, using an MRD. Work is now going on towards automating the creation of the other type of lexical entry in ULTRA, those for the individual language lexicons. To date, such lexicons for machine translation systems have had to be hand-crafted by fluent speakers of that language (e.g., Farwell, Guthrie, and Wilks in [8]).

The procedures for creating these can be thought of in the following way. Bilingual dictionaries (e.g., English-Spanish or vice versa) normally give you simply a list of Spanish equivalent words or phrases for a given English word, and an unskilled user of the dictionary does not usually know enough to select among them. The CRL approach employs an algorithm that, using LDOCE, makes that selection based on context, and does it so as to build up automatically a whole lexicon for processing Spanish or doing translation between English and Spanish.

The Relation of MRDs and Corpora

One of the arguments brought against using dictionaries for lexical knowledge acquisition is that they are static objects, representing, at best, a "frozen" snapshot of language. Given that language is a dynamic and evolving object, there is a degree to which, by the time a dictionary makes it through the compilation and publication process, it has grown out of date. To a certain extent, this argument is used to motivate more recent efforts for manual crafting of lexicons. More important, however, it brings into focus the value of text corpora as alternative (or complementary) sources of lexical information.

This is a comparatively recent development, brought on by several factors. The field has its own "canonical" corpora, which have been in existence for many years: the Brown corpus (compiled by Francis and Kucera and published by Houghton Mifflin in 1982), and the Lancaster-Oslo-Bergen (LOB) corpus (published by the Norwegian Computing Centre for the Humanities), are perhaps the best known examples for this category. Since these corpora are relatively old (and, by modern standards, relatively small—"merely" a million words each), it is safe to say they do not reflect the behavior of the entire language. However, recent technological developments in gathering, publishing, and distributing information have made it possible to have immediate access to very large volumes of text as it is being created. Through services like news wires, transcripts, and electronic publishing, text is available on a scale unmatched by any static language sample. Furthermore, developments in computer technology make it possible to handle such sizable samples without running into operational difficulties due to the limited processing power of earlier generations of computers.

One of the **arguments brought against using dictionaries for lexical knowledge acquisition is that they are static objects,** *representing, at best, a frozen snapshot of language.*

In spite of these developments, the increasing availability of text corpora does not in the least obviate MRDs. On the contrary, the two information sources can work together to provide more robust lexical resources if they are appropriately wed.

As mentioned earlier, the lexicons of real systems must be larger than those of the toy systems of yesterday. However, when processing real text, such as newswire text or newspaper articles, it is painfully obvious that the lexical structures derived from MRDs often do not reflect the behavior of the words in the corpus. There are two major ways in which the dictionary-derived lexical structure may differ from its actual corpus use: syntactic mismatches and selectional mismatches.

For example, in the domain of business joint ventures, there is a sense of *establish* as a three-argument

verb, where the second argument is a relational noun such as joint venture or consortium, as in "IBM will establish a joint venture with a local company."

Using statistical techniques ([17], and similar to those in Grishman and Sterling in [8]), the contexts and type restrictions for this use of the verb in the corpus can be automatically identified.

An example of a selectional mismatch comes with cases of *type coercion*, where the verb "expects" one type but an apparently inconsistent type appears, as in: "Mannheim Industries announced a joint venture with Maykoe Inc." Here, the verb *announce* is typed from LDOCE as taking a sentence object, but it appears in the corpus with a noun phrase. The acquisition technique presented in Pustejovsky shows how coercive contexts can be identified automatically from corpus analysis. This is an adaptation of the techniques of Wilks's Preference Semantics and the type coercion techniques described in [17], in which the arguments of verbs were changed to their preferred ones in certain grammatical and contextual circumstances. In general, the lexical structures described here can be thought of as providing for the shallowest possible semantic decomposition while still capturing significant generalizations about how words relate conceptually to one another.

Work by Church and Hanks [10] has been closely associated with the hope that a dictionary could be built directly from statistically determined associations in text, in a way that COBUILD merely envisioned but could not implement.

Similarly, it has been hoped by many researchers that computational lexicons, normally "seeded" by symbolic construction done by hand or from dictionaries, could be tuned or adapted to the surface associations or contexts implied by a particular corpus and then used in the processing of further, yet unseen, corpora, (e.g., for their lexical disambiguation by the kinds of methods described).

This view is argued strongly by Hindle in [3], who proposes a framework for using large corpora of naturally occurring text together with rule-based systems, in an attempt to build more effective linguistic processors. More recent work supports this view as well. Wilks et al. in [16] have applied statistical measures to a dictionary, treating it as a (highly representative) text sample. Boguraev in [25] instantiates aspects of the lexical semantics of nominals and verbs on the basis of fine-grained analysis of dictionary sources and argues for the need to complement the extraction results with corpus-derived data. Hindle acquires semantic data of a very similar nature, on the basis of studying distributional patterns over syntactic structures associated with the sentences in a large text corpus. Pustejovsky in [16] argues for the value of a theory of lexical semantics as a constraining agent for purely statistical collocational analysis, aimed at populating lexical semantic templates.

Conclusion

Almost no NLP project can function without a lexicon. In the end, even purely statistical projects come to add a lexical module to their systems, whatever their initial prejudices may be. Great strides have been made in automating the production and the tuning of these lexicons, derived from MRDs and from corpora, respectively, so as to express both the more permanent semantic aspects of the language as well as the collocational properties of particular text types. All this has enabled the construction of much more substantial and robust NLP systems for such general tasks as textual lexical disambiguation, in which as we saw, those same techniques have also been needed for the derivation of the lexicon itself (so as to disambiguate its own definitions)

Lexical disambiguation, vital though it is for such tasks as translation, can be understood only with respect to a lexicon itself. This fact, and the related bootstrapping processes, leaves some observers skeptical about the objective basis of sense distinctions. But there is no doubt that unaided coocurrence techniques applied to texts classify word occurrences into clusters that intuitively correspond to word senses, taken broadly—a fact that subsequent psychological experiment confirms.

Great strides have been made in automating the production and the tuning of lexicons *to express the more permanent semantic aspects of the language as well as the collocational properties of particular text types.*

The research effort toward creating NLP lexicons using MRDs has flowered since the late 1980s. The interested reader is directed to the several excellent edited collections that have emerged in that time [2, 9, 11, 12, 16, 22, 25] and to the single existing comprehensive survey [24]. The task for the future is the blending of statistical corpus techniques with others that extract symbolic content from MRDs and corpora to produce more objective, neutral, but still ultimately task-relative lexicons.

Acknowledgments

The authors wish to thank the many people who have contributed to this article through generous sharing of their time and insights to help us understand the rich MRD research environment. We regret that only a fraction of the best work on the subject could be cited here. A slightly longer version of this article, but with quadruple the literature citations, can be copied via anonymous ftp from: republic.ils.nwu.edu

The authors would like to dedicate this article to the memory of Robert Simmons, late of the University of Texas at Austin. Bob was one of the pioneers of MRD research, and contributed immeasurably to the fields of artificial intelligence and computational linguistics through a lifetime of research. He was a large character who stood out from the backdrop of life. He will be missed. ◘

References

1. Amsler, R. A., and White, J. Development of a computational methodology for deriving natural language semantic structures via analysis of machine-readable dictionaries. National Science Foundation, Tech. Rep. MCS77-01315, (1979).
2. Association for Computational Linguistics. Special Issue on Dictionary Processing. *Comput. Linguist., 13* (July-Dec. 1987) 3–4.
3. Association for Computational Linguistics. *Proceedings of the 28th Annual Meeting of the Association for Computational Linguistics (ACL-90)* (Pittsburgh, Pa.) 1990.
4. Association for Computational Linguistics *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics (ACL-91).* (Berkeley, Calif.) 1991.
5. Association for Computational Linguistics (International Committee). *Proceedings of the 8th International Conference on Computational Linguistics (COLING-80)* (Tokyo) 1980.
6. Association For Computational Linguistics (International Committee). *Proceedings of the 12th International Conference on Computational Linguistics (COLING-88)* (Budapest) 1988.
7. Association for Computational Linguistics (International Committee). *Proceedings of the 13th International Conference on Computational Linguistics (COLING-90)* (Helsinki) 1990.
8. Association for Computational Linguistics (International Committee). *Proceedings of the 14th International Conference on Computational Linguistics (COLING-92)* (Nantes, France) 1992.
9. Boguraev, B. K., and Briscoe, T., Eds. *Computational Lexicography for Natural Language Processing.* Longman Group, Harlow, London, 1989.
10. Church, K. W., and Hanks, P. Word association norms, mutual information, and lexicography. *Comput. Linguist. 16,* 1 (1990), 22–29.
11. Evens, M., Ed. *Relational Models of the Lexicon.* Cambridge University Press, Cambridge, England, 1988.
12. Guo, C. M., Ed. *Machine Tractable Dictionaries: Design and Construction.* Ablex, Norwood, N.J. 1992.
13. Helmreich, S., Guthrie, L., and Wilks, Y. A. The use of machine readable dictionaries in the PANGLOSS project. In *Proceedings of the AAAI Spring Symposium on Building Lexicons for Machine Translation* (Stanford Univ.) 1993.
14. Lesk, M. Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. In *Proceedings of the ACM SIGDOC Conference* (Toronto) 1986, 24–26.
15. Neff, M. S., and McCord, M. C. Acquiring lexical data from machine-readable dictionary resources for machine translation. In *Proceedings of the 3rd International Conference on Theoretical and Methodological Issues in MT.* (Austin, Tex.) 1990, 85–91.
16. Pustejovsky, J., Ed. *Semantics and the Lexicon.* MIT Press, Cambridge, Mass., 1993.
17. Pustejovsky, J. The generative lexicon. *Comput. Linguist.,17,* 4 (Dec. 1991), 409–441.
18. Schvaneveldt, R. W., Ed. *Pathfinder Networks: Theory and Applications.* Ablex, Norwood, N.J. 1990.
19. Slator, B. M. Extracting lexical knowledge from dictionary text. *Knowl. Acquis.: An Int. J. 1,* 1. Academic Press. (June, 1989) 89–112.
20. Slator, B. M., and Wilks, Y. A. PREMO: Parsing by conspicuous lexical consumption. In *Current Issues in Parsing Technologies,* M. Tomita, Ed. Kluwer Academic, Dordrecht, Germany, 1991, 85-102.
21. Sparck Jones, K. *Synonymy and Semantic Classification.* Edinburgh Information Technology Series, S. Michaelson and Y. A. Wilks, Eds., (and Studies in Computer-Aided Lexicology, for Sture Allen, Sprakdata, Univ. of Gothenburg, Sweden).
22. Walker, D., Zampolli, A., and Calzolari, N., Eds. *Automating the Lexicon.* Oxford University Press, Oxford, England, 1993.
23. Wilks, Y. A. On keeping logic in its place. In *Proceedings of the 3rd Workshop on Theoretical Issues in Natural Language Processing* (Las Cruces, N.M.) 1987, 110–114.
24. Wilks, Y. A., Slator, B. M., and Guthrie, L. *Electric Words: Dictionaries, Computers, and Meanings.* MIT Press, Cambridge, Mass., 1995.
25. Zernik, U., Ed. *Lexical Acquisition: Exploiting On-line Resources to Build a Lexicon.* Erlbaum, Hillsdale, N.J. 1991.

About the Authors:

LOUISE GUTHRIE joined the Lockheed Martin Text Exploitation Group as Lead Scientist in 1995. She is presently the program manager for their TIPSTER project, which is an ARPA-funded research project on text analysis and extraction. **Current Mailing Address:** Lockheed Martin, Building 10, Rm. 1527, P.O. Box 8048, Philadelphia, PA 19101; email: guthrie@mdso.ve.ge.com.

JAMES PUSTEJOVSKY conducts research in the areas of computational linguistics, lexical semantics, inheritance, and information retrieval and extraction. **Current Mailing Address:** Department of Computer Science, Brandeis University, Waltham, MA 02254; email: jamesp@cs.brandeis.edu.

YORICK WILKS is a professor of computer science at the University of Sheffield. From 1985 to 1993 he was Director of the Computer Research Laboratory at New Mexico State University, a center for research in AI and its applications. **Current Mailing Address:** Department of Computer Science, University of Sheffield, Sheffield S1 4DP, England; email: yorick@dcs.sheffield.ac.uk.

BRIAN M. SLATOR is a research assistant professor at The Institute for the Learning Sciences at Northwestern University where he manages projects aimed at implementing case-based approaches to instruction and productivity. **Current Mailing Address**: The Institute for the Learning Sciences, Northwestern University, Evanston, IL 60201; email: slator@ils.nwu.edu.