Very Large Lexical Databases: An ACL Tutorial

James Pustejovsky
Computer Science Department, Volen Center for Complex Systems, Brandeis University, Waltham, MA 02254 USA
jamesp@cs.brandeis.edu

Patrick Hanks
LingoMotors Inc., 585 Mass. Ave., Cambridge, MA. 02139
patrick@lingomotors.com

July 8, 2001

The current generation of natural language systems for text processing being deployed on the web and in enterprise settings is uniquely different from everything that the natural language community has tried to deploy in real settings to date, for several reasons. Among the most significant changes are the sheer size of the lexical databases and the diversity of content and formats of the lexical data required by present day applications. While only 10 years ago lexicons for NLP systems were struggling to accommodate thousands of lexical forms, recent systems need to handle lexicons that are several orders of magnitude larger. In this tutorial, we will present recent developments in computational lexicography and lexicology, which make it possible to design, instantiate, populate, and maintain Very Large Lexical Databases (VLLDBs) and the methodology for their construction and maintenance. We will focus largely on methodological issues rooted in:

1. Availability of very large corpora;
2. Heavy deployment of statistical techniques for analyzing corpora and context;
3. Content-based corpus analysis techniques;
4. Application of efficient learning algorithms for bootstrapping lexical acquisition;
5. Better understanding of the parametric nature of lexical entries.

The tutorial will highlight a range of challenges facing the designers of VLLDBs, such as size, complexity, diversity, efficiency, sparseness, and the constant change and evolution of language; recent advances in theoretical and computational linguistics, knowledge representation, computer science, and information retrieval, however, offer opportunities to offset these challenges.

Outline of Tutorial Topics

1. **Challenging Orthodox Assumptions about Lexicons**
   (a) What can we do with the data we have from traditional mechanisms?
   (b) For dictionary construction, the dictionary is an end in itself.
   (c) What are the evaluation criteria for a good dictionary?
   (d) One thing you don't want to do is to store word meanings.
   (e) Conclusions:
      - There are no senses.
      - *Lexicon* is a relational term and must be FOR something.

2. **What is a Very Large Lexical Database?**
   (a) Corpora are not lexicons
   (b) Lexicons tell you what is relevant for the application;
   (c) Toy systems and their Lexicons are irrelevant
   (d) Things can be hung on words so let's hang them there. Although senses may not be defensible, words generally are.
   (e) What good is a lexicon?
   (f) What is the Relationship between Corpus Size and Database Size?

3. **Possible versus Probable Meanings in Real Applications**
   (a) Content of Lexicons for Real Applications
      i. proper names: humans, locations, institutions, products, etc.
      ii. open nouns
      iii. open verbs
      iv. open descriptors (adjectives)
      v. what do we do with "collocations", words with white spaces?
   (b) Storage and Retrieval of Lexical Items

4. **Acquisition of Lexical Items and Lexical Features**
   (a) Statistical Techniques for tuning VLLDBs
   (b) Learning Algorithms for Acquiring Lexical Features

5. **Case Study 1: Corpus Analysis and how the results contribute to Lexicon Design**

6. **Case Study 2: Brilliant Lexicon and Parser/Interpreter meets Dull Reality: Reality Bites back**

Associated Reading Selection

1. **What is a Lexicon?**
   James Pustejovsky

2. **Models of Lexical Meaning**
   James Pustejovsky

3. **Computational Lexicography**
   Patrick Hanks

4. **Corpus Analysis of Climb**
   Patrick Hanks

5. **The World Wide Web as a Resource for Example-Based Machine Translation Tasks**
   Gregory Grefenstette

Note: Full Course Slide Packet will be Available for Tutorial Attendees

PART ONE
WHAT IS A LEXICON?

James Pustejovsky

1 The Notion of Lexicon

The lexicon is standardly viewed as a listing of all the morphemes of a language, with information indicating how each morpheme behaves in the components of grammar involving phonology, syntax, and semantics. In no small part, the shape and character of a grammar is determined by what the lexicon contains for these other grammatical devices. Nevertheless, both historically and conventionally, the lexicon has been seen as the passive module in the system of grammar.

More recently, the model of the lexicon has undergone significant revision and maturation. In particular, two trends have driven the architectural concerns of lexical researchers: (1) a tighter integration of compositional operations of syntax and semantics with the lexical information structures that bear them; and (2) a serious concern with how lexical types reflect the underlying ontological commitments of the grammar. In the process, the field has moved towards addressing more encompassing problems in linguistic theory, such as those below:

(1) How can we explain the polymorphic nature of language?

(2) How can we capture the creative use of words in novel contexts?

(3) How can semantic types predictably map to syntactic representations?

(4) What are the "atoms" of lexical knowledge, if they exist at all?

In this entry, we first review the conventional view of the lexicon and then contrast this with the theories of lexical information that have recently emerged in the last few years.

By all accounts, the conventional model of the lexicon is that of a database of words, ready to act in the service of more dynamic components of the grammar. This view has its origins squarely in the generative tradition (Chomsky, 1955) and has been an increasingly integral part of the concept of the lexicon ever since. While the *Aspects* model of selectional features restricted the relation of selection to that between lexical items, work by Jackendoff (1972) and McCawley (1968), showed that selectional restrictions must be available to computations

at the level of derived semantic representation rather than at deep structure. But where did this view come from? In order to understand both the classical model of the lexicon as database and the current models of lexically-encoded grammatical information, it is necessary to appreciate the structuralist distinction between *syntagmatic processes* and *paradigmatic systems* in language. The lexicon has emerged as the focal point communicating between these two components, and can be seen as a hook, which links the information at these two levels. One can go further still and view the elements of the lexicon as not just the building blocks for the more active components of the grammar, but also as actively engaging the building principles themselves.

While syntagmatic processes refer to the influence of horizontal elements on a word or phrase, paradigmatic systems refer to vertical substitutions in a phrasal structure. Syntagmatics evolved into the theory of abstract syntax while paradigmatics was all but abandoned in generative linguistics. In an early discussion of syntagmatic dependencies, Hjelmslev (1943) uses the term *selection* explicitly in the modern sense and notes the importance of integrating paradigmatic systems with the syntagmatic processes they participate in. For Hjelmslev, there are two possible types of relations that can exist between elements in a syntagmatic process: *interdependence* and *determination*, the latter of which is related to the notion of selectional restriction as developed by Chomsky (1965), as Cruse (1986) notes.

One reason that selectional restrictions were not integrated into mechanisms of grammatical selection and description in the 1970s and 1980s is that, if they are imposed correctly, the grammar is forced to model two computations:

1. the entailment relations between selectional restrictions as features must be modelled formally, in order to contribute to the computation of a syntactic description;

2. the manner in which selectional features or constraints contribute to the determination of the meaning of expressions must be enriched in order to exploit these very features.

Recently, with the convergence of several areas in linguistics (lexical semantics, computational lexicons, type theories) several models for the determination of selection have emerged which actively integrate these central syntagmatic processes into the grammar, by making explicit reference to the paradigmatic systems which allow for grammatical constructions to be partially determined by selection. Examples of this approach are Generative Lexicon Theory (Pustejovsky, 1995, Bouillon and Busa, 2001), and to a certain extent, Construction Grammar (Goldberg, 1995), CCG (Steedman, 1997), and HPSG (Pollard and Sag, 1994).

These recent theoretical developments have lead to a new direction in lexical design. Rather than restricting the scope of the lexicon to that of a passive

database, current frameworks have re-architected the relationship between syntactic and semantic representations and the underlying recurrence relations that generate them. These developments have helped to characterize the approaches to lexical design in terms of a hierarchy of semantic expressiveness. There are at least three such classes of lexical description, as defined below (cf. Pustejovsky, 1995 for discussion).

(A) Sense Enumerative Lexicons: lexical items have a single type and meaning, and ambiguity is treated by multiple listings of words.

(B) Polymorphic Lexicons: lexical items are active objects, contributing to the determination of *meaning in context*, under well-defined constraints.

(C) Unrestricted Sense Lexicons: meanings of lexical items are determined mostly by context and conventional use. Few if any restrictions are imposed on how a word may refer.

Although there have been proponents for each class of lexical description defined above, the most promising direction seems to be a careful and formal elucidation of the polymorphic lexicons, and this will form the basis of our subsequent discussion of both the structure and content of lexical entries below.

2 The Structure of a Lexical Entry

As mentioned above, it is generally agreed that there are three components to a lexical item: phonological, syntactic, and semantic information. In this entry, we focus mainly on the manner in which syntactic and semantic representations are encoded in the lexical entry.

There are two types of syntactic knowledge associated with a lexical item: its *category* and its *subcategory*. The former includes traditional classifications of both the major categories, such as noun, verb, adjective, adverb, and preposition, as well as the minor categories, such as adverbs, conjunctions, quantifier elements, and determiners.

Knowledge of the subcategory of a lexical item is typically information that differentiates categories into distinct, distributional classes. This sort of information may be usefully separated into two types, *contextual features* and *inherent features*. The former are features that may be defined in terms of the contexts in which a given lexical entry may occur. Subcategorization information marks the local syntactic context for a word. It is this information that ensures that the verb *devour*, for example, is always transitive in English, requiring a direct object; the lexical entry encodes this requirement with a subcategorization feature specifying that an NP appear to its right. Another type of context encoding is collocational information, where patterns that are not fully productive in the grammar can be tagged. For example, the adjective *heavy*

as applied to *drinker* and *smoker* is collocational and not freely productive in the language (Mel'čuk, 1988). *Inherent features*, on the other hand, are properties of lexical entries that are not easily reduced to a contextual definition, but rather refer to the ontological typing of an entity. These include such features as count/mass (e.g., *pebble* vs. *water*), abstract, animate, human, phyiscal, and so on.

Semantic information can also be separated into two categories: *base semantic typing* and *selectional typing*. While the former identifies the semantic class that a lexical item belongs to (such as entity, event, property), the latter class specifies the semantic features of arguments and adjuncts to the lexical item.

2.1 Word Classes and Typing Information

There are two major approaches to classifying lexical items by their type: syntactic and semantic. (Another influential tradition for verb classification, which we will not discuss in detail here, is a more descriptive approach to word classes, where membership is defined on the basis of grammatical behavior and verbal valency alternation, such as that elaborated on and compiled in Levin (1993)).

One obvious way to organize lexical knowledge, be it syntactic or semantic, is by means of lexical inheritance mechanisms. In fact, much recent work has focused on how to provide shared data structures for syntactic and morphological knowledge (Flickinger, Pollard, and Wasow 1985). Evans and Gazdar (1990) provide a formal characterization of how to perform inferences in a language for multiple and default inheritance of linguistic knowledge. The language developed for that purpose, DATR, uses value-terminated attribute trees to encode lexical information. Briscoe, dePaiva, and Copestake (1993) describe a rich system of types for allowing default mechanisms into lexical type descriptions.

Type structures can express the inheritance of syntactic features (Sanfilippo, 1993) as well as the relationship between more conventional taxanomic information, such as that shown below (cf. Pustejovsky, 1995, Copestake and Briscoe, 1992, and Pustejovsky and Boguraev, 1993):
Given a semi-lattice of types such as this, the lexical items in the language can be associated with a much richer system of differentiated semantic classes. Verbs may also be structured in hierarchical relations, as done in HPSG, LFG, and other lexical frameworks (cf. Pollard and Sag, 1994, Alsina, 1992, Koenig and Davis, 1999).

2.2 Argument Structure

Once the base syntactic and semantic typing for a lexical item has been specified, its subcategorization and selectional information must be encoded in some form. There are two major techniques for representing this type of knowledge:

1. Associate "named roles" with the arguments to the lexical item;

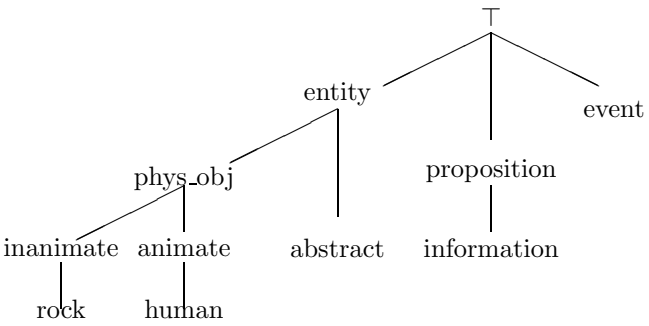

Figure 1: Fragment of a Type Hierarchy.

2. Associate a logical decomposition with the lexical item; meanings of arguments are determined by how the structural properties of the representation are interpreted (cf. Pustejovsky, 1995, Levin and Rappaport, 1995, Hale and Keyser, 1993).

One influential way of encoding selectional behavior is the theory of *thematic relations* (cf. Gruber, 1976, Jackendoff, 1972). Thematic relations are typically defined as partial semantic functions of the event being denoted by the verb or noun, and behave according to a pre-defined calculus of roles relations (e.g., Carlson, 1984, Dowty, 1989, Chierchia, 1989). For example, semantic roles such as agent, theme, and goal, can be used to partially determine the meaning of a predicate, when they are associated with the grammatical arguments to a verb.

The theory of argument structure as developed by Williams (1981), Grimshaw (1990), and others can be seen as a move towards a more minimalist description of semantic differentiation in the verb's list of parameters. The argument structure for a word can be seen as the simplest specification of its semantics, indicating the number and type of parameters associated with the lexical item as a predicate. For example, the verb *build* can be represented as a predicate taking two arguments, while the verb *give* takes three arguments.

(1) a. **build**(x,y)

b. **give**(x,y,z)

What originally began as the simple listing of the parameters or arguments associated with a predicate has developed into a sophisticated view of the way arguments are mapped onto syntactic expressions. Williams' (1981) distinction between *external* (the underlined arguments above) and *internal* arguments and Grimshaw's proposal for a hierarchically structured representation

(cf. Grimshaw, 1990) provide us with the basic syntax for one aspect of a word's meaning. Similar remarks hold for the argument list structure in HPSG (Pollard and Sag, 1994) and LFG (Bresnan, 1994).

The interaction of a structured argument list and a rich system of types, such as that presented above, provides a mechanism for semantic selection that overcomes the difficulties mentioned in the previous section. The most direct impact of semantic type systems on syntactic subcategorization can be seen with the analysis of a simple example.

(2) a. The man / the rock fell.

b. The man / *the rock died.

Returning to the example in (2), consider how the selectional distinction for the feature `[+/-animacy]` is modelled. For the purpose of illustration, the arguments of a verb will be represented in a list structure, where each argument is identified as being typed with a specific value.

(3) a. $\left[\begin{array}{l}\textbf{fall} \\ \text{ARGSTR} = \left[\text{ARG}_1 = \textbf{x:phys\_obj}\right]\end{array}\right]$

b. $\left[\begin{array}{l}\textbf{die} \\ \text{ARGSTR} = \left[\text{ARG}_1 = \textbf{x:animate}\right]\end{array}\right]$

In the sentences in (2), it is intuitively clear how rocks can't die and men can, but it is still not obvious how this inference is computed, given what we would assume are the types associated with the nouns *rock* and *man*, respectively. What accomplishes this computation is a rule of subtyping, Θ, that allows the type associated with the noun *man* (i.e., `human`) to also be accepted as the type `animate`, which is what the predicate *die* requires of its argument as stated in (3b).

(4) $\Theta[human \sqsubseteq animate] : human \rightarrow animate$

The rule Θ, applies since the concept `human` is subtyped under `animate` in the type hierarchy. Parallel considerations rule out the noun *rock* as a legitimate argument to *die* since it is not subtyped under `animate`. Hence, one of the concerns given above for how syntagmatic processes can systematically keep track of which "selectional features" are entailed and which are not is partially addressed by such lattice traversal rules as the one presented here.

Selection can also be employed to solve the problem of polymorphism, where the same lexical item can appear in multiple syntactic contexts, as illustrated in (5).

(5) a. Mary began to read the novel.
b. Mary began reading the novel.
c. Mary began the novel.

Generative Lexicon Theory handles such examples by the use of semantic selection (cf. Pustejovsky, 1995) and canonical syntactic mapping rules, specifying how semantic types correspond to syntactic expressions. Here, the verb *begin* is typed as taking an event description as its internal argument, **begin**(Ind,$\mathcal{E}$), which can be realized in one of the three syntactic forms shown above. This illustrates how a generative system of type operations can account for polymorphic behavior of selection in the syntax.

2.3 Decomposition and Event Structure

The second major approach to the specification of lexical knowledge is that taken by decompositional theories. Most of this research has focused on the lexical specification for verbs, their arguments, and their syntactic behavior. Some recent work, however, has been done on noun semantics as well (Busa, 1996). In this section, we examine the motivations for lexical decomposition in linguistic theory and the various proposals that have emerged for how to encode lexical knowledge as structured forms. We then relate this to the manner in which verbs refer to events, since this directly impacts the nature of the lexical decomposition structure.

Since Davidson (1967), events have played an increasingly important role in the determination of verb meaning. While early researchers on decompositional models (Lakoff, 1965, McCawley, Dowty, 1979) made no ontological commitments to events in the semantics for verbs, a new synthesis has emerged in recent years which attempts to model verb meanings as complex predicative structures with rich event structures (cf. Parsons, 1990, Pustejovsky, 1991, Tenny, 1994, Hale and Keyser, 1993). This research has developed the idea that the meaning of a verb can be analyzed into a structured representation of the event that the verb designates, and has furthermore contributed to the realization that verbs may have complex, internal event structures. Recent work has converged on the view that complex events are structured into an inner and an outer event, where the outer event is associated with causation and agency, and the inner event is associated with telicity (completion) and change of state. Under this view, a canonical accomplishment predicate as in "John sliced the bread" for example, can be represented as composed of an inner and an outer event. The inner event is the telic event in which the bread undergoes a change of state, and the outer event is the event in which John acts agentively (to do whatever is involved in the act of slicing). Since the outer event causes the inner one, it is associated lexically with causation.

Although there is a long tradition of analyzing causation as a relation between two events in the philosophical (cf. Davidson, 1967) and psychological literature (cf. Schank, 1973 and Miller and Johnson-Laird, 1976), in contemporary models of natural language semantics this idea has only recently been adopted. For example, Carter 1976, one of the earlier researchers in this area, represents the meaning of the verb *darken* as follows:

(6) x CAUSE ((y BE DARK) CHANGE))

The predicate CAUSE is represented as a relation between a causer argument x and an inner expression involving a change of state in the argument y. Although there is an intuition that the cause relation involves a causer and an event, Carter does not make this commitment explicitly. Levin and Rapoport 1988 follow a similar strategy, with a CAUSE predicate relating a causer argument and an inner expression involving a change of state in the argument y. The change of state is represented with the predicate BECOME:

(7) wipe the floor clean:
x CAUSE [y BECOME (AT) z] BY [x 'wipe' y]]
x CAUSE [floor BECOME (AT) clean] BY [x 'wipe' floor]

The work of Levin and Rappaport, building on Jackendoff's Lexical Conceptual Structures, has been influential in articulating the internal structure of verb meanings (see Levin and Rappaport 1995).

Jackendoff (1990) develops an extensive system of what he calls *Conceptual Representations*, which parallel the syntactic representations of sentences of natural language. These employ a set of canonical predicates including CAUSE, GO, TO, and ON, and canonical elements including Thing, Path and Event. These approaches represent verb meaning by decomposing the predicate into more basic predicates. This work owes obvious debt to the innovative work within generative semantics, as illustrated by McCawley's (1968) analysis of the verb *kill*. Recent versions of lexical representations inspired by generative semantics can be seen in the Lexical Relational Structures of Hale and Keyser 1993, where syntactic tree structures are employed to capture the same elements of causation and change of state as in the representations of Carter, Levin and Rapoport, Jackendoff, and Dowty.

Pustejovsky (1988,1991) extends the decompositional approach presented in Dowty (1979) by explicitly reifying the events and subevents in the predicative expressions. Unlike Dowty's treatment of lexical semantics, where the decompositional calculus builds on propositional or predicative units (as discussed above), a "syntax of event structure" makes explicit reference to quantified events as part of the word meaning. Pustejovsky further introduces a tree structure to represent the temporal ordering and dominance constraints on an event and its subevents. For example, a predicate such as *build* is associated with a complex event such as that shown below:

(8) [$_{\text{TRANSITION}}$ [e_1:PROCESS] [e_2:STATE]]

The process consists of the building activity itself, while the State represents the result of there being the object built. Grimshaw (1990) adopts this theory in her work on argument structure, where complex events such as *break* are given a similar representation. In such structures, the process consists of what x does

to cause the breaking, and the state is the resultant state of the broken item. The process corresponds to the outer causing event as discussed above, and the state corresponds in part to the inner change of state event. Both Pustejovsky and Grimshaw differ from the authors above in assuming a specific level of representation for event structure, distinct from the representation of other lexical properties. Furthermore, they follow Higginbotham (1986) in adopting an explicit reference to the event place in the verbal semantics.

2.4 Qualia Structure

Thus far, we have focused on the lexical information associated with verb entries. All of the major categories, however, are encoded with syntactic and semantic feature structures that determine their constructional behavior and subsequent meaning at logical form. How this is accomplished, of course, varies from theory to theory.

In Generative Lexicon Theory, it is assumed that word meaning is structured on the basis of four generative factors, or *qualia roles*, that capture how humans understand objects and relations in the world and provide the minimal explanation for the linguistic behavior of lexical items (these are inspired in large part by Moravcsik's (1975, 1990) interpretation of Aristotelian *aitia*).

> FORMAL: the basic category that distinguishes the object within a larger domain;
>
> CONSTITUTIVE: the relation between an object and its constituent parts;
>
> TELIC: its purpose and function;
>
> AGENTIVE: factors involved in the object's origin or "coming into being".

Qualia structure is at the core of the generative properties of the lexicon, since it provides a general strategy for creating new types. For example, consider the properties of nouns such as *rock* and *chair*. These nouns can be distinguished on the basis of semantic criteria which classify them in terms of general categories such as `natural_kind`, `artifact_object`. Although very useful, this is not sufficient to discriminate semantic types in a way that also accounts for their grammatical behavior. A crucial distinction between *rock* and *chair* concerns the properties which differentiate `natural_kinds` from *artifacts*: functionality plays a crucial role in the process of individuation of artifacts, but not of natural kinds. This is reflected in grammatical behavior, whereby "a good chair", or "enjoy the chair" are well-formed expressions reflecting the specific purpose for which an artifact is designed, but "good rock" or "enjoy a rock" are semantically ill-formed since for *rock* the functionality (i.e., TELIC) is undefined. Exceptions exist when new concepts are referred to, such as when the object is construed relative to a specific activity, such as in "The climber enjoyed that rock"; *rock*

itself takes on a new meaning, by virtue of having telicity associated with it, and this is accomplished by integration with the semantics of the subject NP. Although *chair* and *rock* are both `physical_object`, they differ in their mode of coming into being (i.e., AGENTIVE): artifacts are man-made, *rocks* develop in nature. Similarly, a concept such as *food* or *cookie* has a physical manifestation or denotation, but also a functional grounding, pertaining to the relation of "eating." These apparently contradictory aspects of a category are orthogonally represented by the qualia structure for that concept, which provides a coherent structuring for different dimensions of meaning.

For relations, the qualia act in a similar capacity to thematic relations, but where the individual qualia are possibly associated with entire event descriptions, and not just individuals. For discussion, see Pustejovsky (1995).

3 Consequence of Lexical Design

Considering the potential range of information that can be represented lexically results in a reconceptualization of what a lexicon is, so that the very design of the grammar is significantly impacted. Furthermore, our current understanding of psychological and computational properties of language processing suggests that the resources available for lexical storage and access are considerably higher than originally imagined by early grammatical theorists. The consequence is that, what had originally been accomplished in syntax, because of the combinatoric properties inherent in production rules, can be handled by the lexicon itself. The various approaches to lexical encoding can be analyzed in terms of two parameters:

1. Pre-compiling the information into lexical items/forms;
2. Computing or generating new forms or senses during the compositional process.

Typically, idioms are presented as examples of pre-compiled lexical entries. But some theories have adopted this idea as fundamental for the entire compositional operation. The best example of this is Combinatory Categorial Grammar (CCG) (Steedman, 1997). CCG has recently been articulated in enough detail to handle most of the major linguistics phenomena using a library of pre-compiled lexical types, together with the combinatoric rules of categorial syntax. If the grammar utilizes representations with such non-local dependencies, then there must be additional mechanisms for unifying these representations; these are provided in the form of function composition rules and lexical rules (see also Lexicalized TAGs, Schabes et al, 1988).

Lexical rules have been invoked in HPSG, as well, to explain the relationship between the various senses for lexical items, from grinding and packaging operations (such as that relating the animal and food senses of *chicken* and *lamb*),

to the relation between logically polysemous items, such as *book* (information and physical object), and *lecture* (information and event) (cf. Copestake and Briscoe, 1992). In Generative Lexicon such relations are represented explicitly in the type itself, by means of a typing called *dot objects*, and are disambiguated in context (cf. Johnston, 1995). It is very likely, however, that language makes use of both types of devices, namely complex types such as dot objects as well as the application of lexical rules. Regardless, both types of devices must be seriously constrained by the grammar in order not to overgenerate unwanted forms and interpreations.

Considered independently of the issue of pre-compiled vs. generated forms and senses, there is no question that the mental lexicon is large, containing arguably up to 400,000 lexical entries. This is based on fairly conservative estimates of speaker competence with active and passive vocabularies. For example, an average speaker lexicon might contain at least 5,000 distinct verbs, 30,000 distinct nominal forms, and over 5,000 adjectives. Combine this with an additional 10,000 compound forms and at least 300,000 distinct proper names. Obviously, the psychological (and hence computational) demands of these classes are quite distinct. There are two dimensions that can help us distinguish these classes: (1) the degree of combinatoric (functional) complexity of the lexical item; and (2) whether the lexical item is part of active or passive lexical knowledge. Most closed class items, for example, are functionally complex, as are many open class verbs and relational nouns. The majority of the open class items will also involve a fair amount of information regarding combinatoric possibilities. The class of names, however, is unique in that, although it is by far the largest class of lexical items, it is the least demanding in terms of computational resources.

In conclusion, we see that the lexicon is neither a mere listing of morphemes in the language, nor a database of items passively waiting in the service of grammatical processes. Rather, the lexicon is a dynamic and active system of grammar, incorporating as well as dictating essential components of syntactic and semantic composition and interpretation.

See also LEXICAL SEMANTICS, COMPOSITION, SYNTAX-SEMANTICS INTERFACE.

References

Alsina, A. (1992) "On the Argument Structure of Causatives". *Linguistic Inquiry* 23(4):517-555.

Briscoe, T., V. de Paiva, and A. Copestake, Eds. (1993). Inheri-tance, Defaults, and the Lexicon. Cambridge: Cambridge University Press.

Chomsky, N. 1955. *The Logical Structure of Linguistic Theory*, University of Chicago Press, first published 1975.

Chomsky, N. 1965. *Aspects of the Theory of Syntax*, MIT Press, Cambridge.

Davis, Anthony R ; Koenig, Jean-Pierre (2000) "Linking as constraints on word classes in a hierarchical lexicon" *Language* 76, no. 1.

Dowty, D. R. 1979. *Word Meaning and Montague Grammar*, D. Reidel, Dordrecht, Holland.

Grimshaw, J. (1990). Argument Structure. Cambridge, MA: MIT Press.

Gruber, J. S. 1976. *Lexical Structures in Syntax and Semantics*, North-Holland, Amsterdam.

Guthrie, L., J. Pustejovsky, Y. Wilks, and B. Slator. (1996). The role of lexicons in natural language processing. Communications of the ACM 39:1.

Hale, K., and J. Keyser. (1993). On argument structure and the lexical expression of syntactic relations. In K. Hale and J. Keyser, Eds., The View from Building 20. Cambridge, MA: MIT Press.

Halle, M., J. Bresnan, and G. Miller, Eds. (1978). Linguistic Theory and Psychological Reality. Cambridge, MA: MIT Press.

Hjelmslev, Louis. *Prolegomena to a Theory of Language*, translated by F. Whitfield, Madison, University of Wisconsin Press, first published in 1943, 1961.

Jackendoff, R. (1972). Semantic Interpretation in Generative Grammar. Cambridge, MA: MIT Press.

Jackendoff, R. (1990). Semantic Structures. Cambridge, MA: MIT Press.

Johnston, M. 1995. "Semantic Underspecification and Lexical Types: Capturing Polysemy without Lexical Rules," in *Proceedings of ACQUILEX Workshop on Lexical Rules*, August 9-11, 1995, Cambridgeshire.

Levin, B., and M. Rappaport Hovav (1995). Unaccusativity: at the Syntax-Semantics Interface. Cambridge, MA: MIT Press.

Mel'čuk, I. A. (1988) "Semantic Description of Lexical Units in an Explanatory Combinatorial Dictionary: Basic Principles and Heuristic Criteria," *International Journal of Lexicography* 1:165-188.

McCawley, James. 1968. Lexical Insertion in a Transformational Grammar without Deep Structure. Proceedings of the Chicago Linguistic Society 4.

Pinker, S. (1989). Learnability and Cognition: The Acquisition of Argument Structure. Cambridge, MA: MIT Press.

Pollard, C. and I. Sag. 1994. *Head-Driven Phrase Structure Grammar*, University of Chicago Press and Stanford CSLI, Chicago.

Pustejovsky, J. 1991b. "The Syntax of Event Structure," *Cognition* 41:47-81.

Pustejovsky, J. (1995). The Generative Lexicon. Cambridge, MA: MIT Press.

Pustejovsky, J., and P. Boguraev. (1993). Lexical knowledge representation and natural language processing. Artificial Intelligence 63:193-223.

Sanfilippo, A. 1993. "LKB Encoding of Lexical Knowledge," in T. Briscoe, V. de Paiva, and A. Copestake (eds.), *Inheritance, Defaults, and the Lexicon*, Cambridge University Press, Cambridge.

Talmy, L. (1985). Lexicalization patterns: semantic structure in lexical forms. In T. Shopen, Ed., Language Typology and Syntactic Description 3:

Grammatical Categories and the Lexicon. Cambridge: Cambridge University Press, pp. 57-149.

Williams, E. (1981). Argument structure and morphology. Linguistic Review 1:81-114.

Further Readings

Baker, M. (1988). Incorporation: A Theory of Grammatical Function Changing. Chicago: University of Chicago Press.

Bresnan, J. (1994) "Locative Inversion and the architecture of universal grammar", *Language* 70(1):2-31.

Boguraev, B., and E. Briscoe. (1989). Computational Lexicography for Natural Language Processing. Longman, Harlow and London.

Boguraev, B., and J. Pustejovsky. (1996). Corpus Processing for Lexical Acquisition. Cambridge, MA: Bradford Books/MIT Press.

Copestake, A., and E. Briscoe. (1992). Lexical operations in a unification - based framework. In J. Pustejovsky and S. Bergler, Eds., Lexical Semantics and Knowledge Representation. New York: Springer Verlag.

Dowty, D. (1991). Thematic proto-roles and argument selection. Language 67:547-619.

Goldberg, A. E. (1995). Constructions: A Construction Grammar Approach to Argument Structure. Chicago: University of Chicago Press.

Grimshaw, J. (1979). Complement selection and the lexicon. Linguistic Inquiry 10:279-326.

R Ingria, (1986) "Lexical Information for Parsing Systems : Points of Convergence and Divergence" *Automating the Lexicon* Marina di Grosseto, Italy.

Ingria, R., B. Boguraev, and J. Pustejovsky. (1992). Dictionary/Lexicon. In Stuart Shapiro, Ed., Encyclopedia of Artificial Intelligence. 2nd ed. New York: Wiley.

Levin, B. (1993). Towards a Lexical Organization of English Verbs. Chicago: University of Chicago Press.

Lyons, John. *Introduction to Theoretical Linguistics*, Cambridge, Cambridge University Press, 1968.

Miller, G. WordNet: an on-line lexical database. International Journal of Lexicography 3:235-312.

Miller, G. (1991). The Science of Words. Scientific American Library.

Pustejovsky, J. (1992). Lexical semantics. In Stuart Shapiro, Ed., Encyclopedia of Artificial Intelligence. 2nd ed. New York: Wiley.

Schabes, Y, An. Abeille, and A. Joshi, "Parsing Strategies with leixcalized grammars" in *Proceedings of the 12th international Conference on Computational linguistics*, Budapest.

Steedman, M. (1997) *Surface Structure Interpretation*, MIT Press.

Weinreich, U. (1972). *Explorations in Semantic Theory.* The Hague: Mouton.

PART TWO

MODELS OF LEXICAL MEANING

James Pustejovsky

You shall know the meaning of a word by the company it keeps.

−Firth

Indeed, but what kind of company might that be? Linguists, philosophers, and psychologists have debated this question for over a century, with separate and sometimes uncompromising disciplines emerging from the debate. At issue, is what, if anything, is required to understand language beyond the ability to analyze the structural context in which the words appear (e.g., the sentence), and the social context in which they are spoken. Most structural linguists from the 1940's and 1950's subscribed to a fairly standard form of behaviorism, and assumed that information theory would eventually explain the complexities of the linguistic signal. Generative linguists, on the other hand, have generally focused on the innate ability to speak, independent of other cognitive abilities. More recently, many psychologists and Artificial Intelligence researchers have stressed the role of general mechanisms of learning and behavior, which would subsume any specific linguistic mechanisms of mind.

Many of the goals of computational linguistics are the same as those of linguistics in general; to provide useful, testable, and hopefully explanatory theories of the nature of language and its relation to human cognition as a whole. Computational linguistics contributes to the study of language in a number of significant ways, most notably of which are the tools provided for the task. These tools are of two sorts: creating new classes of data, provided by machine−readable dictionaries and texts; and secondly, theories of knowledge representation and the analysis of algorithms operating over these structures.

Approaching the problem both as a computational and a theoretical linguist, my work has aimed at applying the formal techniques of computational models of intelligence to the study of human linguistic capacity. The results of our investigations point to the following model. The human language capacity is a reflection of our ability to categorize and represent the world in a particular way. What is uniquely human is not language per se so much as the generative ability to construct the world as it is revealed through language. Language is the natural manifestation of this faculty for generative categorization and compositional thought. In particular, the ability to categorize **cocompositionally** seems to be characteristic of human behavior uniquely. This is the ability to take a category and refine or redefine its use in a novel context. The continuous refinement and redefinition of what role an object plays in our environment, and how we conceptualize that object as having different properties in different contexts is the process of **cocomposition**.

For the past ten years our research has focused on how word meaning in natural language might be characterized both formally and computationally, in order to account for the ``creative" use of words and concepts in novel contexts. More specifically, our interest is in how words and their meanings combine to form meaningful texts. What makes this task so difficult is the problem of lexical ambiguity. All words are ambiguous to some extent. Even words that appear to have one fixed sense can exhibit multiple meanings in different contexts. 'Room', for example, can mean the physical object (e.g., "John painted the room") or the spatial enclosure defined by this object (e.g., "Smoke filled the room"). The space is just as much a part of the concept of 'room' as is the physical object. The conceptual relation between these two senses is referred to as **logical polysemy**, and this is what partly characterizes language as a ``semi−polymorphic" system of concepts, namely one where sense extensions are constrained in specific ways. Polysemous behavior is also illustrated by the verb 'last', which requires that its subject be an event with some duration; e.g. ``The party lasted all evening". Notice, however, that although the noun 'record' −−−i.e. vinyl object−−− is not an event, in the sentence ``This record lasts an hour", it refers not to the physical artifact itself but to the duration of the record playing. Similarly, the verb 'begin' presupposes that some activity is about to commence; e.g. "John began to swim". The noun 'book' −−−i.e. bound pages−−− is not an activity, yet in the sentence "Mary began the book", the noun refers not to the object itself but to an activity of reading or writing it. What these examples indicate is that the meaning of a word is not fixed throughout all the contexts in which it can appear. From a psychological perspective, data such as these illustrate the polymorphic behavior of our language and the different denotations they reflect.

It has been difficult to link psychologically inspired models of word meaning to the traditional semantic approaches to language involving logical analysis, mostly because these logics assume well−defined and somewhat conservative rules of composition (i.e, ``meanings are composed of the meanings of their parts"). The psychologist says that every word connects to every other word, while the semanticist says that words denote individuals, sets, or relations. Given this chasm, there would appear to be no way to reconcile the two traditions in order to come up with neurally and psychologically inspired logics for language. For example, although a psychologist would want to say that 'book' connects to everything we know about books, there is no formal way to do this in traditional logics, where nouns simply refer to properties and not relations. How then can we give a word such as 'book' the richer relational meaning that it seems to deserve? A representation called **qualia structure** can be seen as providing a minimal explanation for what words mean. For example, the meaning of 'book' encompasses something like the Aristotilean modes of explanation (viz. causes). That is, I need to know what its function is as well as its basic category; where it came from and what it is made of. Hence, when I enjoy a book, I am generally referring to the reading of the book, whereas when I refer to purchasing a book, I refer to the physical object itself. The qualia encode this information for our concept 'book' directly as its denotation.

What our research provides is a procedural method of lexical decomposition, incorporating a rich, recursive theory of semantic composition, the notion of ``semantic well−formedness", and a notion of how these representations are integrated into a larger knowledge representation language, through inheritance. Because there has been so little attention paid to other lexical categories besides verbs, our efforts have been centered on defining the minimal semantic representations for nouns and adjectives. Not until all major parts of speech been studied, can we hope to arrive at a balanced understanding of the conceptual lexicon and the methods of composition.

PART TWO

In studying the nature of conceptual and lexical ambiguity, there are some important issues to address regarding compositionality in general (i.e., the way concepts combine to make larger concepts). By viewing the process of categorization as governed by rules of a generative nature, we begin to address the issue of logical polysemy and the phenomenon of the creative use of words (i.e., the means by which words take on new senses in novel contexts).

This research suggests that lexical and conceptual decomposition is possible if it is performed generatively. Rather than assuming a fixed set of primitives, we assume a fixed number of generative devices that can be seen as constructing semantic expressions. Just as a formal language is described more in terms of the productions in the grammar than its accompanying vocabulary, a semantic language is defined by the rules generating the structures for expressions rather than the vocabulary of primitives itself. For this reason, we can think of a dictionary of concepts as a **generative lexicon**.

Similarly, from psychological considerations, a cognitive dictionary cannot be simply a listing of concepts without also a concern for space and time factors within the system (and the algorithms therein). The semantic system we have been developing is able to capture the variable space of possible sense extensions, while maintaining a constant number of lexical senses.

A grammar for lexical semantics is computationally interesting and useful only if the individual lexical representations can be tested over large samples of data. To empirically test this view of lexical semantics, we have been conducting research to apply such ``semantic intensive'' techniques to information retrieval tasks. This work is in effect a large−scale empirical test of many of the tenets of the semantic theory we have been working on for the past six years. We are currently utilizing machine−readable resources (e.g. dictionaries and large corpora) to extract subtle semantic relations between lexical items and phrases in texts. These relations are then stored with the words in the language to improve the performance of an information retrieval system during queries and data extraction.

PART TWO

PART THREE

COMPUTATIONAL LEXICOGRAPHY

Patrick Hanks

1. Introduction

An inventory of words is an essential component of programs for a wide variety of natural language processing applications, including information retrieval, machine translation, speech recognition, speech synthesis, and message understanding. Some of these inventories contain information about syntactic patterns and complementations associated with individual lexical items (see Chapter 3); some index the inflected forms of a lemma to the base form (see Chapter 2); some include definitions; some provide semantic links and hierarchies between the various lexical items (see Chapter 14). Some are derived from existing human−user dictionaries, as discussed below. None are completely comprehensive; none are perfect. Even where a machine−readable lexicon is available, a lot of computational effort may need to go into 'tuning' the lexicon for particular applications. Sometimes, an off−the−peg lexicon is deemed to be more trouble than it is worth, and a required lexicon may be constructed automatically by induction from texts.

At the same time, the craft of lexicography has been revolutionized by the introduction of computer technology. On the one hand, new techniques are being used for compiling dictionaries and word lists of various kinds; on the other, new insights are obtained by computational analysis of language in use.

In this chapter, two meanings of the term 'computational lexicography' are distinguished:

1. Restructuring and exploiting human dictionaries for computational purposes.

2. Using computational techniques to compile new dictionaries.

The focus is on computational lexicography in English. A comprehensive survey of computational lexicography in all the languages of the world is beyond the scope of this chapter. Lexicography in many of the world's neglected languages is now being undertaken in many research centres; the work is often computer−assisted and associated with a machine−readable product. Useful web sites in this connection are Robert L. Beard's index of on−line dictionaries and multilingual resources (http://www.yourdictionary.com) and the Omnilex site (http://www.omnilex.com).

For European languages, the TELRI association (Trans−European Language Resources Infrastructure; http://www.telri.de/) is a valuable resource, with objectives that go far beyond lexicography:

· to strengthen the pan−European infrastructure for the multilingual language research and development community;

· to collect, promote, and make available monolingual and multilingual language resources and tools for the extraction of language data and linguistic knowledge;

· to offer a customized comprehensive service to academic and industrial users;

· to prepare and organize research and development projects focusing on translation aids, multilingual authoring systems, and information retrieval;

· to provide a forum where experts from academia and industry share and assess tools and resources, assess software, evaluate new trends, investigate alternative approaches, and engage in joint activities;

· to make available the expertise of its partner institutions to the research community, to the public, and to language industry.

2 Historical background

Until recently, the only reason anyone ever had for compiling a dictionary was to create an artefact for other human beings to use. Up to the Renaissance, dictionaries were either bilingual tools for use by translators, interpreters, and travellers, or Latin and Greek word lists for students and scholars. As living languages and cultures became more complex, vocabularies expanded and people began to compile dictionaries of "hard words" in their own language—learned words which ordinary people might not understand. The earliest example in English is Robert Cawdrey's *Table Alphabeticall... of Hard Usual Words ... for the Benefit of Gentlewomen and other Unskillful Persons* (1604). It was not until the 18th century that lexicographers set themselves the objective of collecting and defining *all* the words in a language. For English, this culminated in Samuel Johnson's *Dictionary of the English Language* (1755), containing not only definitions but also illustrative citations from "the best authors".

Johnson's was the standard dictionary of English until the end of the 19th century, but already in 1857 Richard Chenevix Trench presented a paper to the Philological Society in London, *On some Deficiencies in our English Dictionaries*, in which he described lexicographers as “the inventory clerks of the language”. This paper played a large part in motivating the Philological Society's *New English Dictionary on Historical Principles*, alias *The Oxford English Dictionary* (1878–1928).

Many of the deficiencies that characterized 19th−century dictionaries still beset lexicography today, though sometimes in new forms, and they are of computational relevance. They arise from problems of both practice and principle. Chief among them are:

It is literally impossible to compile an exhaustive inventory of the vocabulary of a living language. Trench noted many omissions and oversights in the dictionaries of his day, but the creative nature of the lexicon means that every day new words are created ad hoc and, in most but not all cases, immediately discarded. It is impossible for the inventorist to know which neologisms are going to catch on and which not. Murray deliberately omitted the neologism **appendicitis** from the first edition of OED. An American dictionary of the 1950s deliberately omitted the slang term **brainwash**. The first edition of Collins English Dictionary (1979) omitted **ayatollah**. In their day, each of these terms was considered too obscure, informal, or jargonistic to merit inclusion, though hindsight proved the judgement to be an error. That said, almost all today's machine−readable dictionaries offer a very high degree of coverage of the vocabulary of ordinary non−specialist texts well over 99.9% of the words (as opposed to the names). Lexical creativity is peripheral, not central, in ordinary discourse.

Coverage of names is a perennial problem. Some dictionaries, on principle, do not include any entries for names; for example, they contain an entry for **English** (because it is classified as a word, not a name), but not for **England**. Other dictionaries contain a selection of names that are judged to be culturally relevant, such as **Shakespeare**, **New York**, **Muhammad Ali**, and **China**. Very few brand names and business names are found in dictionaries: **Hoover** and **Thermos flask** are judged to have become part of the common vocabulary, but no dictionary includes brand names such as **Malteser** or **Pepsi**, whatever their cultural relevance. No dictionary makes any attempt to include all the names found in a daily newspaper. However, names can be just as important as words in decoding text meaning. Hanks (1997), discussing the role of immediate−context analysis in activating different meanings, cites an example from the British National Corpus: in the sentence “Auchinleck checked Rommel” selection of the meaning ‘cause to pause’ for **check** depends crucially on the military status of the subject and object. If Auchinleck had been Rommel's batman, or a customs inspector, or a doctor, a different sense of **check** would have been activated.

Ghost words and ghost senses constantly creep in, evading the vigilance of lexicographers despite their best efforts. Crystal (1997: 111) mentions *commemorable* and *liquescency* as examples of words which have probably never been used outside the dictionaries in which they appear. He goes on to cite *Dord*, glossed as ‘density’, a ghost word which originated in the 1930s as a misreading of the abbreviation *D or d* (i.e. capital or lower−case d), which does indeed mean ‘density’.

No generally agreed criteria exist for what counts as a sense, or for how to distinguish one sense from another. In most large dictionaries, it might be said that minor contextual variations are erected into major sense distinctions. In an influential paper, Fillmore (1975) argued against "checklist theories of meaning", and proposed that words have meaning by virtue of resemblance to a prototype. The same paper also proposed the existence of 'frames' as systems of linguistic choices, drawing on the work of Marvin Minsky (1974) among others. These two proposals have been enormously influential. Wierzbicka (1993) argues that lexicographers should "seek the invariant", of which (she asserts) there is rarely more than one per word. This, so far, they have failed to do; nor is it certain that it could be done with useful practical results. Nevertheless Wierzbicka's exhortation is a useful antidote to the tendency towards the endless multiplication of entities (or, to put it more kindly, drawing of superfine sense distinctions) that is characteristic of much currently available lexicography.

In the emergent United States, the indefatigable Noah Webster published his *American Dictionary of the English Language* (1828), a work which paid particular attention to the American English, which was already beginning to differ from standard British English, although its definitions owe more to Johnson than its compiler liked to admit. Johnson, Murray, and Webster all compiled their dictionaries on 'historical principles'. That is, they trace the semantic development of words by putting the oldest meanings first. This is a practice still followed by many modern dictionaries. It is of great value for cultural and literary historians, but at best an unnecessary distraction and at worse a potential source of confusion in most computational applications. For purposes of computational linguistics, if word meaning is in question at all, it is more important to have an inventory that says that a *camera* is a device for taking photographs than to know that, before the invention of photography, the word denoted "a small room" and "the treasury of the papal curia".

The earliest comprehensive dictionary to make a serious attempt to put modern meaning first was Funk and Wagnall's (1898). Unfortunately, the great Funk and Wagnall's dictionaries of the early 20th century no longer exist in any recognizable form. Current American large dictionaries which claim to put modern meanings first are *The Random House Dictionary* (1964, 1996), the second edition of which is available on CD−ROM, and *The American Heritage Dictionary* (1969, 1992). A British counterpart is *Collins English Dictionary* (1979; fourth edition 1999).

Because they not only put modern meanings first, but also contain fuller syntactic information (including, in some cases, rudimentary hints about selectional preferences), dictionaries for foreign learners are popular among computational researchers and tool builders. The pioneering work in this class was A. S. Hornby's *Oxford Advanced Learner's Dictionary of Current English* (OALDCE; 1948). The sixth edition (2000) has been fully revised, taking account of corpus evidence from the British National Corpus.

Most such dictionaries are available in machine−readable form (MRDs: machine−readable dictionaries), and research rights can sometimes be negotiated with publishers. To overcome problems of commercial sensitivity, in some cases older editions are licensed. Probably the most widely cited dictionary in computational applications is the *Longman Dictionary of Contemporary English* (LDOCE; 1978; http://www.longman−elt.com/dictionaries). The latest edition of LDOCE is available on CD−ROM. Like

OALDCE, it has been revised using evidence from the British National Corpus. It also devotes considerable attention to spoken English. The electronic database of LDOCE, offered under specified conditions for NLP research, contains semantic domains and other information not present in the published text.

In 1987, with the publication of the COBUILD dictionary (an acronym for 'Collins Birmingham University International Language Database', 1987, 1995), a new development in lexicography emerged: the corpus−based dictionary. The word 'corpus' is a fashionable buzz word designating a wide variety of text collections (see Chapter 9). In the sense most relevant to lexicography, a corpus is a collection in machine−readable form of whole texts or large continuous extracts from texts. Such a collection provides a more statistically valid base for computational processing and study of contemporary English than a collection of citations or quotations. A corpus can be used to study words in use, but only indirectly to study word meanings. COBUILD is more intimately connected with its corpus than any other dictionary. It offers a highly interactive and informative web site (http://titania.cobuild.collins.co.uk). Unlike the British National Corpus, which maintains its balance by being static, the so−called 'Bank of English' is dynamic: a so−called 'monitor corpus', constantly growing. At the time of writing it consists of over 330 million words of running text. This provides Collins lexicographers with a magnificent resource for studying new words and meanings.

A recent addition to the stock of major corpus−based dictionaries is the *Cambridge International Dictionary of English* (CIDE; 1995; http://dictionary.cambridge.org), which has a number of interesting features, including associated data modules for NLP such as lists of verb complementation patterns, semantic classifications of nouns, and semantic domain categories.

In 1998, Oxford University Press published *The New Oxford Dictionary of English* (NODE), a dictionary for native speakers of English (as opposed to foreign learners) which draws both on the citation files of the large historical *Oxford English Dictionary*, collected by traditional methods, and on new corpus resources, in particular the British National Corpus of 100 million words of text. Use of a corpus enables lexicographers to make more confident generalizations about common, everyday meanings, while citation files provide a wealth of quotations to support rare, interesting, new, and unusual words and uses.

The biggest wordlist in a one−volume English dictionary is to be found in *Chambers English Dictionary*. This magnificent ragbag of curiosities achieves its vaunted 215,000 references by including a great deal of archaic Scottish and other dialect vocabulary (e.g. "**giz** or **jiz** (*Scot*) a wig") and obsolete literary forms (e.g. "**graste** (*Spenser*) *pa p* of **grace**"), of more interest to Scrabble players than to serious computational linguists.

The foregoing paragraphs mention the main 'flagship' dictionaries likely to be of interest to computational linguists. Each of the flagship publications is associated with a family of other lexical reference works, for example thesauruses, dictionaries of idioms, dictionaries of phrasal verbs, dictionaries for business English, and so forth.

Section 5 of this chapter makes further reference to corpus−based lexicography in Britain. No dictionaries based on serious large−scale corpus research have yet been published in the United States, although the *American Heritage Dictionary* made some use of the pioneering Brown Corpus (1 million words; see Francis and Kucera 1982), and an American edition of NODE, under the working title *Oxford Dictionary of the English Language*, is in preparation at the time of writing.

3. Restructuring and exploiting human dictionaries for computational purposes

All humans — foreign learners, native speakers, translators, and technical specialists alike — share certain attributes which are not shared by computers. Typically, humans are very tolerant of minor variation, whereas a computer process may be thrown by it. For example, the first edition of the *Oxford English Dictionary* (OED) contains innumerable minor variations which the 19th century compilers were unaware of or considered unimportant. To take a simple example, “Shakes.”, “Shak.”, and “Shakesp.” are among the abbreviations used for “Shakespeare”. When OED was prepared for publication in machine−readable form, at first on CD−Rom, and now on line (http://www.oed.com/), the editors spent much time and effort standardizing the text in order to ensure that user searches would produce comprehensive results as well as being swift, efficient, and robust. Imposing standardization has been a major concern for making dictionaries machine−tractable. At the more complex end of the spectrum, it is clearly desirable to impose standardization in definition writing, so that, for example, the definitions for all edible marine fish would be retrievable by searching for a single defining phrase. This involves standarization of innumerable variations such as “eatable fish”, “strong−tasting fish”, “edible sea fish”, “edible flatfish”, “marine fish with oily flesh”, etc. Such tasks present a potentially infinite series of challenges for the standardizer. Attempts to devise short cuts or automatic procedures using resources such as a machine−readable thesaurus can lead to unfortunate consequences such as equating the meaning of ‘shaking hands’ with ‘shaking fists’.

Early work in creating machine−readable dictionaries (MRDs) generally involved converting typesetters' tapes into a database format. Unbelievably large quantities of typographical instructions had to be stripped out, leaving just a few that could be converted into logical field delimiters. Nowadays, new dictionaries are routinely set up from the outset as structured files or databases, from which typesetter's files are derived. However, the vast size and cost of dictionaries, their long gestation periods, and the great length of their marketing lives mean that there are still quite a few electronic dinosaurs lumbering about, containing valuable information in text but encrusted with typographic details.

The earliest MRD was the computerization at SDC (Systems Development Corporation), of *Webster's 7th New Collegiate Dictionary* (Olney 1967; Revard 1968), which was keyboarded from the printed text. The choice of text now seems surprising, in view of the historical principles which determine the order of definitions in this dictionary and the complete absence of any clues linking meanings to use, other than basic part−of−speech classes. However, the project leaders presumably took the view that one dictionary is as good as any other, or else that the market leader for human use (selling over a million copies a year) must be good for computer applications. Among other things, the SDC group explored word frequencies in definitions, postulating a privileged semantic status for certain frequent terms such as “substance, cause, thing,” and

"kind," akin to the semantic primitives of Wierzbicka and Wilks, or the "semantic parts of speech" of Jackendoff. Revard later wrote that, in an ideal world, lexicographic definers would "mark every … semantic relation wherever it occurs between senses defined in the dictionary." (Revard, 1973).

Among the most comprehensive analyses of a machine−readable dictionary for lexicographic purposes is the work carried out under the direction of Yorick Wilks at New Mexico State University, and subsequently the University of Sheffield, on LDOCE. The electronic database of LDOCE contains systematic information on semantic domain, in addition to the published text. This work is reported in Wilks, Slator, and Guthrie (1996), which also includes a comprehensive survey of other work on making dictionaries machine−tractable. The most important of the earlier survey volumes is Boguraev and Briscoe (1989), a collection of nine essays describing work in the 1980s to extract semantic and syntactic information from dictionaries, in particular LDOCE.

4 Dictionary structure

Dictionaries are more highly structured than almost any other type of text. Nowadays, the norm is to follow the TEI (text−encoding initiative; www.uic.edu/orgs/tei) for SGML− and HTML−compatible markup.

The basic tag set for an entry in the *New Oxford Dictionary of English*, which may be regarded as typical, includes the following tags, with nesting (embedding) as shown:

<se> standard entry, *or*

<ee> encyclopedic entry, *embedding:*

<hw> headword

<pr> pronunciation

<ps> part of speech

<s1> sense level 1 (part of speech)

<s2 num=n> sense level 2, with number attribute, *embedding*:

<df> definition

<ms>meaning extension

<ex> example of usage (taken from the British National Corpus or the *Oxford English Dictionary* citation files)

<et> etymology

<drv>derivative form, *embedding:*

<ps>part of speech

Additional tags are used for optional and occasional information, for example usage notes. This tag set is derived from the considerably more elaborate tag set designed in the 1980s for the OED. Tagged, consistently structured dictionary texts can be searched and processed by algorithms of the kind designed by Tompa (1992) and his colleagues at the University of Waterloo. This software was designed with the computerized OED in mind, but it has a much wider range of applicability, to machine−readable texts of all kinds. The two principal components of this software are PAT, a full text search system offering a powerful range of search options, and LECTOR, a text display facility. PAT allows users to construct combinations of results using Boolean expressions or proximity conditions. Depending on the text structure, search conditions can be specified within certain fields or regions, some of which are pre−defined, while others may be made up ad hoc by the user. For example, a user may wish to find all definitions containing the word “structure” in entries for words beginning with R. PAT enables rapid text searches and retrieval within specified fields of specified groups of entries.

5. Using computational techniques to compile new dictionaries

Lexicographers were quick to seize on the benefits of computers in compiling and typesetting new dictionaries. As long ago as 1964, the *Random House Dictionary of the English Language* was set up as an electronic database, so that different technical senses could be dealt with in sets, regardless of alphabetical order, by relevant experts, thus greatly improving the consistency of treatment. Clearly, consistency of treatment in a dictionary benefits from compilation of entries for domain−related and semantically related words together as sets, without regard to where in the alphabet they happen to fall. This is now standard practice in the compilation of all new dictionaries (as opposed to revised editions and derivative or shortened versions, which usually proceed alphabetically).

Corpus−based lexicography raised a whole new raft of issues, affecting the selection, arrangement, and definition of the lexical inventory. For example, there may be plentiful evidence for a verbal adjective, e.g. *extenuating*, while the base form (*extenuate*) is rare or non−existent. Should there be an entry for the base

form, the verbal adjective, or both?

The evidence of a large general corpus can help to identify the most common modern meaning of a word, but it must be treated with caution. Frequency alone is not enough. Corpus lexicographers also need to look at the distribution: does the word occur in many different texts, only in a particular domain, or only in a single author? For an idiosyncrasy, even if repeated, is still an idiosyncrasy.

Another trap is the failure−to−find fallacy. Failure to find a particular word or sense in a corpus does not mean that that sense does not exist. It may exist in a register or domain that is inadequately represented in the corpus. On the other hand, it might be argued that a word, phrase, or sense that does not occur in a balanced corpus of 100 million words (let alone 300 or 400 million words), containing a broad selection of text types, cannot be very important –or, rather, can only be of importance in a highly restricted domain.

Corpus lexicographers invoke criteria such as generalizability to identify the “core meaning” of a word. So, for example, the expression “to shake one's head” is far more common in the British National Corpus than “to shake a physical object”, but the latter sense is still identified as the core meaning and placed first because the range of possible direct objects is so much wider. Core meanings have wider ranges of normal phraseology than derivative, metaphoric, and idiomatic senses.

Identifying the ‘literal’ modern meaning is often far from straightforward. A sense whose status is that of a conventionalized metaphor may be more common than the so−called literal sense. Literal meanings are constantly on the move: today’s metaphor may be tomorrow’s literal meaning. Thus, *torrents of abuse* and *torrents of verbiage* are more common in a large corpus of modern English than *torrents* denoting violently rushing mountain streams, but most English speakers would agree that the latter is nevertheless the literal meaning.

In the 1990s, dictionary publishers, especially publishers of foreign learners' dictionaries, have invested substantially in revising their dictionaries to conform better with corpus evidence, both for the word list and for the meaning and use of words. Corpus−driven revision can involve wholesale rewriting and re−structuring of definitions, seeking levels of generalization that conform with the evidence. This in turn might affect the view of semantic hierarchies derived from analysis of machine−readable dictionaries, though to the best of my knowledge no systematic comparison has been carried out.

Many computer scientists take the view that such details are too fine−grained to be of much interest for computing, or that the work has been done once (albeit on flawed sources) and is not worth doing again. It is therefore surprising to hear these same computer scientists complaining about the amount of lexical tuning needed to make a lexicon suitable for a particular application. It would be interesting to find out whether a lexicon more firmly rooted in empirical methods needs significantly less tuning for certain applications.

A revolutionary development of the 1990s was WordNet (see Fellbaum 1998; http://www.cogsci.princeton.edu/~wn/), an on−line reference system combining the design of a dictionary and a thesaurus with the rich potential of an electronic database. Instead of being arranged in alphabetical order, words are stored in a database with hierarchical properties and links, such that *oak* and *ash* are subsumed under *tree.* Fourteen different senses of *hand* are distinguished, each with its own set of links WordNet's design was inspired by psycholinguistic theories of human lexical memory. English nouns, verbs, adjectives and adverbs are organized into synonym sets, each representing one underlying lexical concept. Different relations link the synonym sets.

It has to be said, however, that , while WordNet's design is new and ground−breaking, its lexicography is often disappointing, owing virtually nothing to corpus linguistics and far too much to traditional dictionaries on historical principles. So, for example, sense 1 of the verb **fan** is glossed as "strike out a batter, in baseball" and sense 4 is "separate from chaff; of grain". It cannot be claimed that either of these senses is central to contemporary usage. The gloss at sense 3, "agitate the air", is technically deficient, in that it fails to indicate that this is normally a transitive verb with a selectional preference for a direct object denoting a person (or a person's face or body). Such faults are by no means unusual. A systematic revised edition of WordNet, taking account of current advances in lexicographic methodology and resources, is urgently needed. The present situation, in which different groups of researchers make their own adjustments on a piecemeal basis, is far from satisfactory.

In 1996, a European initiative, EuroWordNet, was set up to build a semantic net linking Spanish, Italian, and Dutch to the original English WordNet. EuroWordNet aims to be a standard for the semantic tagging of texts and an interlingua for multilingual systems of information retrieval and machine translation. One can look up a term in Dutch and get synonyms in English, Spanish, or Italian. Currently, other languages are being added to the inventory. The importance of EuroWordNet cannot be understated: it could well turn out to be a strategically significant language tool in enabling everyday communication and commerce to take place in the diverse languages of Europe.

The single most important feature of the WordNet projects, like that of more traditional dictionaries, is coverage. Unlike most other institutionally funded research projects, WordNet says something about everything. And, unlike commercial projects, it is free.

6 Linking Meaning and Use

A serious problem for computer applications is that dictionaries compiled for human users focus on giving lists of meanings for each entry, without saying much about how one meaning may be distinguished from another in text. They assume a decoding application for the dictionary, in which ordinary human common sense can be invoked to pick out the relevant meaning from a list of competing choices. Computers, on the other hand, do not have common sense. Many computer applications need to know how words are used and, ideally, what textual clues distinguish one sense from another. On this subject, dictionaries are largely silent. Learners' dictionaries offer syntactic patterns, but these are at a clausal level, without any more delicate

distinction between different semantic classes of direct object.

Choueka and Luisgnan (1985) were among the first to describe the essentials of choosing an appropriate meaning by reference to the immediate co−text. This is a technique that has been employed since, but is still a subject on which further research is needed. Part of the problem is distinguishing signal from noise, while another is lexical variability. It is clear that there are statistically significant associations between words (see Church and Hanks 1989; Church and others 1994), but it is not easy to see how to establish that, for purposes of choosing the right sense of **shake**, *earthquake* and *explosion* may be equated, while *hand* and *fist* may not. Corpus lexicographers often cite the words of J. R. Firth (1957): "You shall know a word by the company it keeps". Much modern research is devoted to finding out exactly what company our words do keep. This work is still in its infancy. Establishing the norms and variations of phraseology and collocation in a language will continue to be important components of many lexicographic projects for years to come. Recently, A European Society for Phraseology (Europhras; http://www.ik.fh−hannover.de/person/rothkegel/EUROPHRAS/Startseite.html) has been founded, with the specific objective of promoting the study of phraseology, which may be expected to yield relevant results in this context..

COBUILD's innovative defining style expresses links between meaning and use by encoding the target word in its most typical phraseology (e.g. "when a horse **gallops**, it...") as the first part of the definition (see Hanks, 1987). COBUILD does this impressionistically and informally, in a way designed for human users (foreign learners), not computers, but in principle a project to express similar information in a formal, computer−tractable way, is not inconceivable. The editor−in−chief of COBUILD, John Sinclair, briefed his editorial team: "Every distinction in meaning is associated with a distinction in form." A great deal of research is still required to determine exactly what counts as a distinction in meaning, what counts as a distinction in form, and what is the nature of the association. The immediate local context of a word in a text is often but not always sufficient to determine which aspects of the word's meaning are active in that text. For further discussion, see Hanks 1996 and 2000.

The Japanese Electronic Dictionary Research Institute (http://www.iijnet.or.jp/edr/) has developed a series of eleven linked on−line dictionaries for advanced processing of natural language by computers. Sub−dictionaries include a concept dictionary, word dictionaries, and bilingual dictionaries (English –Japanese). The *EDR Electronic Dictionary* is aimed at establishing an infrastructure for knowledge information processing.

7 Exploring the future

Innovation has been very much the exception rather than the rule in lexicography. Lexicography aims at breadth, not depth, and most lexicographic projects are required, for commercial reasons, to reach a very wide popular audience. Unlike most researchers, teams of lexicographers are obliged by the nature of their undertaking to say something about everything, even if they have nothing to say. These and other constraints mean that the style and presentation of most dictionaries tends to be very conservative, reflecting

18th−century concepts of meaning and definition for example.

In recent years, a number of research projects have explored possible new approaches to explaining or defining word meaning and use. Such studies do not cover the entire vocabulary, but rather explore new methodologies and presentations based on recent research in philosophy of language, cognitive linguistics, and other fields, along with new resources, in particular corpora. They point the way towards more comprehensive future developments. The most important of these projects are mentioned in this section.

The European Community's Research and Development Service (www.cordis.lu/) provides information on research projects funded by the EC. Of particular relevance was the Information Technologies programme of 1994−98 (named *Esprit*; see http://www.cordis.lu/esprit/src/). This sought, with an emphasis on commercial relevance, to favour research in the languages of Central Europe, the Baltic States, the Mediterranean region, and the states of the former Soviet Union, designed to bring the information society to everyone, including speakers of minority languages.

A major theme in the EC's "Fifth Framework" (1998−2002) is the development of "Information Society technology" (IST; www.cordis.lu/ist/). There appears to be disappointingly little provision for lexicographic research in this framework. The most important such project is Defi at the University of Li{e`}ge, which explores how to use the immediate context of a word in a text to select the right translation.

In the "Fourth Framerwork" lexicographically relevant projects were funded such as

Delis, Compass, Sparkle, and Eagles, all of which are described on the Cordis web site.

Hector: In the Hector project (Atkins 1993; Hanks 1994), lexicographers grouped word uses (as found in a sample corpus of 18 million words) and linked them to different senses. 1400 lexical items were studied, research that led to a new approach to defining style in the *New Oxford Dictionary of English* (1998). Links that were set up between corpus and dictionary by Hector, and it was used as a benchmark for word−sense disambiguation in the Senseval exercise (Kilgarriff 1998). For a fuller discussion of the word−sense disambiguation problem, see Chapter 15 (Stevenson and Wilks).

VerbNet: (Hanks, Krishnamurthy, and Palmer, forthcoming) picked up where Hector left off, making a systematic set of links between the syntax and semantics of English verbs, designed to cover all 'normal' uses. Those uses not covered by the analysis are by definition classified as 'exploitations of norms'.

Framenet: Fillmore and Atkins (1992) describe a lexicon, Framenet (http://www.icsi.berkeley.edu/~framenet/), in which verbs with similar meanings (e.g. verbs of movement)

are distinguished by the different semantic case roles of their arguments. It is corpus−based and contrastive (e.g., it asks precisely what semantic features distinguish *creeping* from *crawling*).

Hector, VerbNet, and Framenet are all pilot projects with limited coverage. It is to be hoped that at least some of these approaches will be carried out comprehensively across the whole lexicon, with resultant tools linking word senses to textual phraseology in a robust enough way to reduce or even eliminate the amount of lexical tuning needed to make a lexicon suitable for a wide variety of NLP applications.

8 Further Reading and Relevant Resources

The most useful readings in computational lexicography are to be found in the proceedings of conferences and in specialist journals:

The Waterloo−OED conference: annually from 1984 to 1994, organized jointly by Oxford University Press and the University of Waterloo Centre for the New OED and Text Research, headed by Frank Tompa (Waterloo, Ontario, Canada N2L 3G1). The proceedings contain accounts of most major developments in computational lexicography in this period, when seminal developments were taking place.

Complex: annual conference organized by the Hungarian Research Institute for Linguistics, Budapest (http://www.nytud.hu/). Proceedings edited by Franz Kiefer, Gabor Kiss, and Julia Pajsz, with many relevant papers.

Euralex: Biennial conference of the European Association for Lexicography (www.ims.uni−stuttgart.de/euralex/). Proceedings contain occasional reports on significant computational developments.

International Journal of Lexicography (ed. R. Ilson (to 1997), A. Cowie (from 1998); Oxford University Press; www3.oup.co.uk/jnls/list/lexico/), quarterly. Occasional articles of computational relevance.

Dictionaries: the Journal of the Dictionary Society of North America (ed. William S. Chisholm (to 1999), M. Adams (from 2000); polyglot.lss.wisc.edu/dsna/); annual. Until recently, disappointingly few articles have been of computational relevance.

Other relevant collections of essays include those in Zernik (1991); and Atkins and Zampolli (1994).

The Oxford Text Archive (http://ota.ahds.ac.uk/) and the Linguistic Data Consortium at the University of Pennsylvania (http://www.ldc.upenn.edu/) both hold copies of a variety of machine−readable dictionaries, which are available for research use under specified conditions.

Most dictionary publishers are willing to make machine−readable versions of their dictionaries available for bona−fide academic research, though great tenacity and diplomatic skill may be required to achieve agreement and delivery. Publishers' sensitivity about protecting commercial rights in their colossal, high−risk investments, along with the fact that negotiating the free gift of their products is not always among their highest priorities, can be perceived, usually erroneously, as hostility to research.

The Oxford Text Archive (http://ota.ahds.ac.uk/) and the Linguistic Data Consortium at the University of Pennsylvania (http://www.ldc.upenn.edu/) both hold copies of a variety of machine−readable dictionaries, which are available for research use under specified conditions.

The Oxford English Dictionary is available on CD−ROM and has recently become available on line through certain sites (http://www.oed.com/).

References

Amsler, Robert A., and J. White (1979): 'Development of a computational methodology for deriving natural language semantic structures via analysis of machine−readable dictionaries'. Washington DC: NSF technical report.

Apresyan, Yuri, Igor Mel'çuk, and A. K. Zholkovsky (1969) 'Semantics and lexicography: towards a new type of unilingual dictionary', in F. Kiefer (ed.) *Studies in Syntax and Semantics*, Dordrecht: D. Reidel.

Atkins, Beryl T. (Sue), Judy Kegl, and Beth Levin (1988): 'Anatomy of a Verb Entry' in *International Journal of Lexicography* 1:2.

Atkins, Beryl T. (Sue), and Beth Levin (1991): 'Admitting impediments' in Zernik 1991.

Atkins, Beryl T. (Sue) and Antonio Zampolli (eds., 1994): *Computational Approaches to the Lexicon*. Oxford University Press.

Boguraev, Bran, and Ted Briscoe (1989): *Computational Lexicography for Natural Language Processing*. Harlow, Essex: Longman Group; New York: John Wiley and Sons, Inc.

Choueka, Yaacov, and S. Luisgnan (1985): 'Disambiguation by short context' in *Computers and the Humanities*, 19: 147−157

Church, Kenneth W., and Patrick Hanks (1989):'Word Association Norms, Mututal Information, and Lexicography' in *Computation Linguistics* 16, pp. 22−29.

Church, Kenneth W., William Gale, Patrick Hanks, Donald Hindle, and Rosamund Moon (1994): 'Lexical Substitutability' in Atkins and Zampolli (1994).

Crystal, David (1997): *The Cambridge Encyclopedia of Language*. Cambridge University Press.

Fellbaum, Christiane (ed. 1998): *WordNet: an Electronic Lexical Database*. Cambridge, Massachusetts: MIT Press.

Fillmore, Charles J. (1975): 'An alternative to checklist theories of meaning' in *Papers from the First Annual Meeting of the Berkeley Linguistics Society*, pp. 123−132.

Firth, J. R. (1957): 'A Synopsis of Linguistic Theory 1930−55' in *Studies in Linguistic Analysis*. Oxford: Philological Society.

Francis, W. Nelson, and Henry Ku{cv}era (1982): *Frequency Analysis of English Usage*. Boston MA: Houghton Mifflin.

Hanks, Patrick (1987): 'Definitions and Explanations' in Sinclair (1987).

Hanks, Patrick (1996): 'Contextual Dependency and Lexical Sets' in *International Journal of Corpus Linguistics*: 1:1, pp. 75−98.

Hanks, Patrick (2000): 'Do Word Meanings Exist?' in *Computing and the Humanities*.

Hayes, Brian (1999): "The Web of Words" in *American Scientist* 87, pp. 108–112

Jackendoff, Ray (1990): *Semantic Structures* Cambridge, Mass.: The MIT Press

Kilgarriff, Adam (1993): "Dictionary word sense distinctions: an inquiry into their nature" in *Computing and the Humanities* 26: pp. 356–387

Minsky, Marvin (1974): 'A framework for representing knowledge': Artificial Intelligence Memo No. 306, M.I.T. Artificial Intelligence Laboratory.

Olney, J. (1967): "Toward the development of computational aids for obtaining a formal semantic description of English". Santa Monica, California: SDC technical report

Pustejovsky, James (1995): *The Generative Lexicon*. Cambridge, Massachusetts: MIT Press.

Pustejovsky, James, and Bran Boguraev (1996): ***Corpus Processing for Lexical Acquisition.*** Cambridge University Press.

Resnick, Phil (1997): 'Selectional Preferences and Word Sense Disambiguation' in *Proceedings of the SIGLEX Workshop 'Tagging Text with Lexical Semantics: What, why, and how'*.

Revard, Carter (1968): ‘On the Computability of Certain Monsters in Noah's Ark’. Santa Monica, California: SDC technical report

Revard, Carter (1973): ‘Towards a NUDE (New Universal Dictionary of English)’ in Raven I. McDavid and Audrey R. Duckert (eds.): *Lexicography in English*. New York Academy of Sciences.

Riloff, E., and W. Lehnert (1993): ‘Automatic Dictionary Construction for Information Extraction from Text’ in *Proceedings of the Ninth IEEE Conference on Artificial Intelligence for Applications*, pp. 93−99.

Sinclair, John (ed., 1987): *Looking Up: An Account of the Cobuild Project in Lexical Computing*. London and Glasgow: Collins ELT.

Sinclair, John (1991): *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Tompa, Frank W. (1992): *An Overview of Waterloo's Database Software for the OED*. University of Waterloo Centre for the New Oxford Dictionary and Text Research. Also published in T. R. Wooldridge (ed.): *Historical Dictionary Databases*: University of Toronto Centre for Computing in the Humanities.

Wierzbicka, Anna (1985): *Lexicography and Conceptual Analysis.* Ann Arbor, Michigan: Karoma.

Wierzbicka, Anna (1987): *English Speech Act Verbs*. New York: Academic Press.

Wierzbicka, Anna (1993): ‘What are the Uses of Theoretical Lexicography?’ in *Dictionaries: the Journal of the Dictionary Society of North America*, 14, pp. 44−78.

Wilks, Yorick A., Brian M. Slator, and Louise M. Guthrie (1996): *Electric Words: Dictionaries, Computers, and Meanings*. Cambridge, Mass.: The MIT Press

Zernik, Uri (ed., 1991): *Lexical Acquisition: Using On−line Resources to Build a Lexicon*. Hillsdale NJ: Lawrence Erlbaum Associates.

DICTIONARIES

Note: dates cited are for first editions unless otherwise specified.

Cawdrey, Robert (1604): *A Table Alphabeticall ... of Hard Usuall Words ... for the Benefit of Gentlewomen and other Unskillful Persons*.

Gove, Philip (ed., 1963): *Webster's 7th New Collegiate Dictionary*. Springfield, Mass. Merriam Webster

Hanks, Patrick, and others (eds., 1979): *Collins English Dictionary*. London and Glasgow: Collins

Hanks, Patrick, Judy Pearsall, and others (1998): *The New Oxford Dictionary of English*. Oxford: Oxford University Press

Johnson, Samuel (1755): *Dictionary of the English Language*. Facsimile Edition (1979): London: Times Books.

Murray, James, Henry Bradley, Alexander Craigie, and C. T. Onions (1878–1928): *A New English Dictionary on Historical Principles*, subsequently entitled *The Oxford English Dictionary*. Oxford: Oxford University Press

Procter, Paul, and others (1978): *Longman Dictionary of Contemporary English*, 1st edition. Harlow, Essex: Longman.

Procter, Paul, and others (1995): *The Cambridge International Dictionary of English*. Cambridge: Cambridge University Press

Morris, William (1969): *The American Heritage Dictionary*. Boston, Mass.: Houghton Mifflin

Sinclair, John, Patrick Hanks, and others (1987): *Collins COBUILD English Language Dictionary*. London and Glasgow: Collins

Stein, Jess, Laurence Urdang, and others (eds., 1964): *The Random House Dictionary of the English Language*. New York: Random House

Webster, Noah (1828): *An American Dictionary of the English Language.*

PART FOUR

CORPUS ANALYSIS OF CLIMB

Patrick Hanks

CLIMB: a process verb, with grammatically metaphorical uses as a verb of state.

I. TRANSITIVE USES

[HUMAN] climb [HIGH THING]

1. Stalin died in 1953, and Hillary climbed Everest "because it was there". In
2. dge University Climbing Club, to climb Mont Blanc by the Goutier route befo
3. r walkers, and almost anyone can climb Triglav: the last refuge is only 400
4. oad range. When Charles Whitman climbed the university tower in Austin, Te
5. Wood Green School, Witney. They climbed a drainpipe to enter the school th
6. lete that the postman has had to climb a ladder to the front entrance to de
7. generously collusive. He could climb an oak and sit there alone for all o
8. ted it. Show her a tree and she climbed it. Not so Prince Charles. He was
9. arsing everything. If necessary climb the scaffolding yourself to get the
10. r climb. Young boys are forever climbing things. Beaming she swung the go

****

[examples of exploitations, metaphors, and uncertainties]

11. on. How good are the beetles at climbing cereral plants and locating aphid

12. med down into the troughs before climbing the next steep wave. Away from th

13. plotter in the Air Force before climbing the civil service ladder with a j

14. he answer is probably that he is climbing the ladder of a lucrative career

15. don't know whether to eat it or climb it! A five−minute drive up to road

****

[HUMAN] climb [OBSTACLE]

16. the end of the footpath and then climbed a stile. He believed he got home u

17. 1942 I should think, I remember climbing some railings at the back of Guil

18. refugees. Some of the refugees climbed the embassy wall. Others broke thr

19. conceived of the possibility of climbing the Abbey wall. No suddenly it s

****

[HUMAN] climb [STAIR]

20. we crawled, troglodytes all. We climbed a narrow and broken staircase tow

21. ago on a gentle Autumn evening I climbed some steep stairs in a converted h

22. iety. She chewed her lip as she climbed the remaining stairs to Nevil's do

23. ; a rectilinear spiral. She had climbed nearly 400 steps and

****

[HUMAN] climb [PATH]

When the subject is HUMAN and the object is a PATH, it is not always clear from the immediate context whether the HUMANS are on foot or in a vehicle. In this case, the condition `ON FOOT' may be taken as a default

24. d through the ford and began to climb the gradual slope beyond. Dogs barke
25. ce. It was still raining as we climbed the pass to the Spanish frontier,
26. er water seemed louder when she climbed the road by herself. Martha though
27. of hundred feet above as they climbed the slope, like a fortress behind
28. gaps in the teak boards as we climbed the gangplank. A plump old man sit
29. Rashidiyeh. But they had never climbed the hill. There are, of course, s

****

[VEHICLE] climb [PATH]

30. were bumper to bumper as they climbed Headington Hill, the Astra behind
31. very efficiently. A trolleybus climbing a hill was often aided by power f

[PATH] climb [PATH]

Expresses a state rather than a process

32. hamlet the smaller unpaved road climbed a shallow hill before disappearing

33. tray of refreshments. The lawn climbs a slope several yards in front of t

34. down to Boscombe Pier. It then climbs the inevitably steep hill back to t

---

II. NULL COMPLEMENT

[HUMAN] climb

35. gainst the rock, Harlin began to climb. Charsky stared up after him. Then

36. a mixed Italian and German team climbing not far away, heading for

****

[An ambiguous use]

37. outh overhead Dunster Castle we climbed through the cloud which had now f

****

[PLANE] climb

38. where it was grown. The plane climbed ponderously but the mountain slid

****

[VAPOUR] climb

39. the column of steam and ash climbing eleven kilometres high about the

40. g explosions and oily smoke was climbing from the burning truck to the

41. her than later. Thunder−clouds climbed steeply over Poitiers, and as Peli

****

sun climb

42. matched their joy; the sun was climbing into a cloudless sky and beginnin

43. But faces grew red as the sun climbed, the cicadas chanted and the tar b

_____

III: WITH PP COMPLEMENT

[HUMAN] climb [from SOURCE] [via PATH] [to GAOL]

44. to a halt in front of her Maggie climbed aboard and went upstairs. She ador

45. aded when approaching a house or climbing across a fence. It it hadn't been

46. the embassy railings even as she climbed across to safety. Only the interve

47. olice said the man was trying to climb from a tower block's seventh floor t

48. et. Angry workers glowered as I climbed from my car. A policeman waves me

49. Taylor said: 'We have a man who climbs in with the sharks to clean the tan

50. Charlie loaded up the van, then climbed in. 'Mr Lawler will be upset that

51. to Mum and Dad's room. There he climbs into bed and goes to sleep. Mum and

52. fice in Sanaya, west Beirut, and climbing into his armoured Mercedes, waving

53. imbing−frame. That it should be climbed on, into and through, compliments

54. The front door blocked, the men climbed onto the roof and then things got

55. slowly, Gower wandered back and climbed over the stile. He made wretchedly

And, metaphorically ...

56. for first−time buyers trying to climb on to the first rung of the housing

With 'down' ...

57. ion of running water, attempt to climb down the slippery cemented sides of

58. an ice axe he would be lucky to climb down fifty feet without falling. It

59. third floor but people there had climbed down from the balconies and were

[PATH or GROUND] climb [ADVERBIAL OF DIRECTION]

Expresses a state rather than a process

60. and verges. A precipitous road climbs from Batcome to the crest of the d

61. ry here is in perfect order. It climbs in tiered rows up a hard, bare hill

62. oot, banks thick with daffodils, climbing out of sight, 'She would enjoy t

63. e next mile is a wonderful walk, climbing out of the valley, with panoramic

64. rly planted beet the pine forest climbed over gently undulating hills. 'Yo

65. ked up at the dim stairway which climbed steeply out of the bare and musty

****

[PLAYER] climb above [PLAYER]

A cliche. Genre: British sports journalism

66. −in enabling Chris Fairclough to climb above defenders and head past Carte

67. om their second corner, Robinson climbed above static defenders to head Ga

68. r 38 minutes when Alan Kernaghan climbed high to Putney's corner and heade

---

IV. [ABSTRACT] climb ([AMOUNT] or by [AMOUNT]) to [AMOUNT]

69. ring wage costs will accordingly climb by 4 per cent in 1990 and wages in

70. he good: coal prices look set to climb by 80 per cent over the next 25 yea

71. week of losses ended as the MIB climbed 10 points to 1,088, boosted by fo

72. ed 6p to 227 p and Racal Telecom climbed 12p to 342p. STC was the subject

****

[HUMAN or THING] climb [AMOUNT] PP

73. e money for diabetic children by climbing 15,000 feet up Mount Kilimanjaro.

74. towards the rim of the valley, climbing 2000 feet in eight relentless miles

PART FIVE

The World Wide Web as a Resource for Example−Based Machine Translation Tasks

Gregory Grefenstette

Xerox Research Centre Europe, Grenoble, France

Abstract. The WWW is two orders of magnitude larger than the largest corpora. Although noisy, web text presents language as it is used, and statistics derived from the Web can have practical uses in many NLP applications. For this reason, the WWW should be seen and studied as any other computationally available linguistic resource. In this article, we illustrate this by showing that an Example−Based approach to lexical choice for machine translation can use the Web as an adequate and free resource.

Key Words: WWW, Example−Based, machine translation, corpus linguistics, very large lexicon

1. Introduction

The idea of using attested linguistic events to choose between theoretically possible events underlies Example−Based Natural Language Processing tasks. This approach has been used for Machine Translation (Sato and Nagao, 1990; Dagan et al, 1991; Sumita et al, 1993) and to improve Cross−Language Information Retrieval (Ballesteros and Croft, 1998). For these tasks, candidate multiword translations are generated using human−compiled electronic dictionaries or using equivalence lexicons derived from bilingual aligned corpora (Brown et al, 1990). The candidate translations are scored using statistics of the candidates' attested appearances in a reference corpus, and the highest scoring candidate are chosen as the translation term.

It is evident that the World Wide Web can be considered as an extremely large corpus of attested examples. Some linguists cringe at the idea of using this uncharacterized and dirty corpus to derive linguistic information, but we argue that the sheer size of the WWW as a corpus allows signal to overcome noise. There exist a few large corpora that have been collected and cleanly prepared, such as the British National Corpus[1] of 100 million words (90 million from written text, and 10 million from spoken text), but the quantity of text available through the Web swamps these collections. To get an idea of the size of the World Wide Web, we show, in Table 1, a list of counts of some random noun phrases in this large British National Corpus and their counts in an indexed Web browser, *AltaVista*[2], on a given day in late 1998.

These examples show that the number of attestable patterns is almost two orders of magnitude larger on the Web than the number to be found in the largest available corpora. Statistical techniques, such as Example−Based methods, rely on the presence of events of to perform well. Many Example−Based techniques suffer performance drop−offs when they try to make choices using rare events, since the distinction between signal and noise becomes blurred. The size of the Web, however, weakens [3] the effect of Zipf's law (Zipf, 1965), since intuitively likely events do become common enough for statistical techniques to work.

	BNC	*WWW*
sample phrases	100 M Words	
medical treatment	202	46064
prostate cancer	28	40772
deep breath	374	54550
acrylic paint	20	7208
perfect balance	28	9735
presidential election	74	23745
electromagnetic radiation	24	17297
powerful force	54	17391
concrete pipe	8	3360
upholstery fabric	5	3157
vital organ	30	7371

Table 1. Counts of some random noun phrases in the British National Corpus and as found on the World Wide Web by the AltaVista browser in late 1998.

As an anecdotal example of how the Web can be used as a resource in the Example−Based task of lexical choice in dictionary−based machine translation, consider the following example. Take the compositional French noun phrase *groupe de travail*. In the Oxford−Hachette French−English dictionary, the French word *groupe* can be translated by the English words *cluster, group, grouping, concern* and *collective*. The French word *travail* can be translated by the English words *work, labor* or *labour*. The naïve translator has five

(from *groupe*) times three (from *travail*) possible ways of translating *groupe de travail*. Now, the AltaVista search portal allows the Web browser user to search for adjacent phrases by placing their query in double−quotes. Combining the possible translations of *groupe de travail* into all twenty−one possible noun phrases creatable by simply re−ordering the nouns and concatenating them to form English phrases, and then submitting these phrases to this Web browser yeilds, in Table 2, the actual occurrence statistics in the web pages indexed by this browser. We see that the phrase *work group* is much more frequent than all the others, and is the most likely domain−independent translation in the group[4].

	WWW count		*WWW count*
labor grouping	4	labor cluster	7
labour concern	8	work grouping	27
labor concern	28	work cluster	112
labor collective	144	labour collective	158
work concern	170	work collective	242
labor group	844	labour group	1131
work group	67238		

Table 2. Web counts of some possible ways of translation the French expression *groupe de travail* using the possible translations of *groupe* and *travail* given in a bilingual French−English dictionary Some possibilities (eg labour cluster) did not appear at all.

Going from anecdote to experimentation, we test in the next section the use of the World Wide Web as a resource for Example−Based Machine Translation on a large−scale.

2. Experimentation

In order to perform an objective, large−scale experiment on the adequacy of the World Wide Web as a linguistic resource for an Example−Based Machine Translation task, we created a gold standard of compositional compounds from a publicly available electronic bilingual dictionary[5]. The standard was created by eliminating all phrases in the dictionary which were not transparent translations of their subparts. We tested two language directions: German−to−English and Spanish−to−English. To find compositional

noun phrases in this multilingual dictionary, we extracted two complete sets of all German compound nouns and all Spanish nominal phrases satisfying the four criteria:

i. [compound] the dictionary entry was decomposable into two other Spanish or German words found in the dictionary,

ii. [compositionality] the compound term was translated in the English part of the dictionary by two word phrases,

iii. [transparency] the words in the English translations of the smaller German or Spanish components permitted the construction of candidate translations that included the dictionary−given compound−word translation, and

iv. [ambiguity] there was more than one possible English translation candidate.

These sets of words, then, correspond to the entire list of German compounds and Spanish terms in this full−size dictionary such that, if they were not in the dictionary, their proper English translation could be constructed from the translation of the subparts of the German word or Spanish term using that same dictionary. Only such words which had ambiguous translations were retained. This strategy led to a set of 724 German words constituting our gold standard of potentially ambiguous compositional German compounds, and a set of 1140 compositional Spanish terms. With each German word or Spanish term, we also have their preferred[6] English translations.

For each German word and for each Spanish term, we then ignored the dictionary entry for the compound, and created the English candidate translations as if the non−English term were not included in the dictionary. This situation reproduces what human users must do for most novel German compounds or novel Spanish terms encountered. In each case, we created all the possible two word translations using the decomposed[7] German word and the individual words of the Spanish terms (ignoring prepositions) and recombining the English translations of these subparts from the German−to−English or Spanish−to−English sides of the same dictionary.

Since each of the 724 German compound words was ambiguously translatable (given the translations of their components in the reference dictionary), 3556 possible English translations were generated. For the 1140 ambiguous Spanish multiword terms, there were 6186 possible English translations built using this simple concatenation strategy. Each possible translation candidate was sent to AltaVista as a phrasal query, and the frequency[8] of occurrence of the phrase was noted. To use the WWW as a decision mechanism for choosing the proper translation, the most frequently occurring phrase was chosen as the best example for translating the ambiguous term. This choice was compared against the actual translation that the dictionary gave for them. The results of this experiment are shown in the Table 3, showing that 86−87% of the choices were correct.

Number of German nouns responding to 4 criteria	724

Number of candidate English translations	3556
Number of correct translations choosing most frequent phrase in AltaVista as best	631
Percent of correctly chosen translations	87%

Number of Spanish terms responding to 4 criteria	1140
Number of candidate English translations	6186
Number of correct translations choosing most frequent phrase in AltaVista as best	976
Percent of correctly chosen translations	86%

Table 3. The results of creating translation candidates from subparts of German compounds and Spanish multiword expressions, and then choosing the translation candidate that appears most often in a Web Browser.

Here are some example of the translation candidates and their AltaVista frequencies. In the following tables, we give some examples of the German compound words and the Spanish terms with the English candidate translations that were generated by translating the components. For each candidate, the number of times that AltaVista had found the phrase is given as *AltaVista count.* The next two columns show whether the frequency information is sufficient to pick a dictionary−given translation: if there is the abbreviation DICT in column 5 then the English candidate translation of the components corresponds to the gold standard dictionary translation of the German compound or the Spanish term. The word MAX in the last column shows which of the English candidates was most frequent on the Web indexed by Altavista[9]. 87% of the ambiguous German words and 86% of the ambiguous Spanish multiword terms tested had DICT in column 5 and the word MAX in column 6, meaning that the most frequent attested candidate on the Web was also a gold standard translation of the compound word. For example *Appartementhaus* generates 8 candidate translations: *apartment chop, apartment cut, apartment house,* … of which *apartment house* is the most common on the Web and the translation given for the compound. On the other hand, *Aktienkurs* generated 8 translations of which *stock price* was the most common but not given in the dictionary. This last example was counted among the 13% incorrect German cases. Notice in the tables that many candidates that are not the most frequent ones still have no−zero frequencies, for example *apple sap,* one of the candidate translations of *Apfelsaft* still appeared 25 times on the Web.

German compound	***English candidate***	***AltaVista count***	***gold standard***	***most frequent***

Angebotspreis	offer price	9767	DICT	MAX
Angebotspreis	offer prize	206	–	
Apfelkraut	apple herb	167	–	MAX
Apfelkraut	apple syrup	159	DICT	
Apfelsaft	apple juice	13841	DICT	MAX
Apfelsaft	apple sap	25	–	
Appartementhaus	apartment chop	0	–	
Appartementhaus	apartment cut	127	–	
Appartementhaus	apartment house	8356	DICT	MAX
Appartementhaus	apartment rampage	0	–	
Appartementhaus	flat chop	10	–	
Appartementhaus	flat cut	621	–	
Appartementhaus	flat house	882	–	
Appartementhaus	flat rampage	0	–	
Bogenbrücke	arch bridge	2304	DICT	MAX
Bogenbrücke	bow bridge	224	–	

An example from the Spanish data shows that this experiment only gives the most common translations (corresponding to those appearing in the bilingual gold standard dictionary) whereas in a specific domain, a rarer translation might be acceptable. For example, the experiment erroneously chooses *energy field* as the translation of *campo de fuerzas*, rather than the dictionary supplied *force field*, but the choice of one or the other may well depend on the domain or context of application. Here, we are simply saying that the WWW provides an idea of the most common way of saying something.

German compound	*English candidate*	*AltaVista count*	*gold standard*	*most frequent*
Aktienkurs	share course	246	–	
Aktienkurs	share cure	0	–	
Aktienkurs	share price	48221	DICT	
Aktienkurs	share rate	598	–	
Aktienkurs	stock course	60	–	
Aktienkurs	stock cure	5	–	
Aktienkurs	stock price	48394	–	MAX
Aktienkurs	stock rate	167	–	
Blutspender	bleed donor	0	–	
Blutspender	bleed giver	0	–	
Blutspender	blood donor	5432	DICT	MAX
Blutspender	blood giver	5	–	
Blutzelle	bleed cell	0	–	
Blutzelle	blood cell	25514	DICT	MAX
Braunkohle	brown cabbage	20	–	
Braunkohle	brown coal	2317	DICT	MAX
Briefwaage	letter balance	509	DICT	MAX
Briefwaage	letter Libra	2	–	
Briefwaage	letter scales	131	DICT	
Brotmesser	bread knife	1167	DICT	MAX
Brotmesser	bread meter	0	–	
Brotmesser	loaf knife	0	–	

Brotmesser	loaf meter	0	−	

Note that AltaVista does not index noun phrases but merely contiguous words. These AltaVista counts are a rough estimate of a given noun phrase. This experiment could also be made more subtle by generating more varied syntactic forms (such as *A of B*) or through a more intelligent use of morphological variants, without modifying the way that the available Web browser indexes its pages. Ideally, the Web browsers would perform a more intelligent indexing, extracting not only contiguous terms but dependency structures that can be derived through current robust, shallow parsing systems (Appelt et al, 1993; Ait−Moktar and Chanod, 1997; Grefenstette, 1997). But even in its simple state, this German and Spanish to English experiment shows that the WWW is a linguistic resource of the same nature and same (though possibly greater) utility as those corpora now used in Natural Language Processing tasks.

3. Conclusion and Perspectives

We have presented an experiment in Example−Based Natural Language Processing using the World Wide Web as the exemplar linguistic resource for decision making. Our experiment was on a much larger scale than previous efforts (Dagan et al, 1991; Rackow et al, 1992), limited to a few dozen words, since we included all the potentially ambiguous compounds in a large translation dictionary, and worked with a corpus (the entire WWW visited by AltaVista) that is orders of magnitude larger than any previously used corpus.

Spanish term	***English candidate***	***AltaVista count***	***gold standard***	***most frequent***
agregado de prensa	press attaché	403	DICT	MAX
agregado de prensa	squeezer attaché	0	−	
agua corriente	common water	2815	−	
agua corriente	current water	5213	−	
agua corriente	draft water	1438	−	
agua corriente	draught water	11	−	
agua corriente	flowing water	13264	−	
agua corriente	going water	343	−	

agua corriente	ordinary water	2040	−	
agua corriente	power water	12695	−	
agua corriente	running water	49358	DICT	MAX
agua corriente	stream water	9264	−	
agua corriente	usual water	1252	−	
agua mineral	mineral water	33058	DICT	MAX
agua mineral	ore water	178	−	
agua salada	pickle water	284	−	
agua salada	salt water	98690	DICT	MAX
àguila real	actual eagle	60	−	
àguila real	essential eagle	11	−	
àguila real	real eagle	176	−	
àguila real	royal eagle	431	DICT	MAX
ahorro de energía	decisiveness saving	0	−	
ahorro de energía	energy saving	140148	DICT	MAX

A human (or computer) deciding on the correct translation of compositional noun phrases would be faced with the same choice as that presented in this Example−Based Natural Language Processing experiment. An extremely simple exploitation of the WWW provides the linguistic resource, a relatively free resource one might add, to resolve this choice with 86−87% accuracy.

This experiment argues for a greater exploitation and study of the Web as a linguistic resource, and for applying techniques of shallow parsing to create more linguistically informed indexes than those available through current web portals.

References

Salah Ait−Mokhtar and Jean−Pierre Chanod. 1997. Incremental finite−state parsing.

In *ANLP'97*, pages 72−−79, Washington.

Douglas E. Appelt, Jerry R. Hobbs, John Bear, David Israel, and Mabry Tyson. 1993. FASTUS: A finite−state processor for information extraction from real−word text.

In *Proceedings IJCAI '93*, Chambery, France, August.

Lisa Ballesteros and W. Bruce Croft. 1998. Resolving ambiguity for cross−language retrieval. In W. Bruce Croft, Alistair Moffat, C.J. van Rijsbergen, Ross Wilkinson, and Justin Zobel, editors, *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 64−−71, Melbourne, Australia, August. ACM Press, New York.

Peter F. Brown, John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, Robert L. Mercer, and Paul S. Roossin. 1990.A statistical approach to language translation. *Computational Linguistics*, 16(2):79−−85.

I. Dagan, I. Atai, and U. Schwall. 1991. Two languages are better than one. In *Proceedings of the 29th Meeting of the ACL*, pages 130−−137, Berkeley.

Gregory Grefenstette. 1997. SQLET : Short query linguistic expansion techniques: Palliating one or two−word queries by providing intermediate structure to text. In *RIAO'97, Computer−Assisted Information Searching on the Internet*, Montreal, Canada.

U. Rackow, I. Dagan, and U. Schwall. 1992. Automatic translation of noun compounds. In *Proceedings of COLING'92*, pages 1249−−1253, Nantes, France, August 23−−28.

S. Sato and M. Nagao. 1990. Towards memory−based translation. In H. Karlgren, editor, Proceedings of *COLING'90*, pages 247−−252, Helsinki.

Anne Schiller. 1996. Deutsche flexions− und kompositionsmorphologie mit pc−kimmo. In Roland Hausser, editor, *Linguistische Verifikation: Documentation zur Ersten Morpholympics 1994*, number 34 in *Sonderuck aus Sparche und Information*. Max Niemeyer Verlag, Tubingen.

E. Sumita, K. Oi, O. Furuse, H. Iida, T. Higuchi, N. Takahashi, and H. Kitano. 1993. Example−Based machine translation on massively parallel processors. In *Proc. of the 13th IJCAI*, pages 1283−−1288, Chambery, France.

G. K. Zipf. 1965. *Human Behavior and the Principle of Least Effort*. Hafner, New York.

Spanish term	***English candidate***	***AltaVista count***	***gold standard***	***most frequent***
ala delta	delta nostril	0	−	
ala delta	delta wing	1525	DICT	MAX
ala delta	delta winger	1	−	
àlbum de sellos	seal album	56	−	
àlbum de sellos	stamp album	1805	DICT	MAX
alfombra oriental	easterly carpet	0	−	
alfombra oriental	eastern carpet	115	−	
alfombra oriental	oriental carpet	5985	DICT	MAX
alumbrado de emergencia	emergency lighting	17940	DICT	MAX
alumbrado de emergencia	emergency lit	5	−	
ambiente del trabajo	labor atmosphere	105	−	
ambiente del trabajo	labor cosiness	0	−	
ambiente del trabajo	labor coziness	0	−	

ambiente del trabajo	labor snugness	0	−	
ambiente del trabajo	labour atmosphere	4	−	
ambiente del trabajo	labour cosiness	0	−	
ambiente del trabajo	labour coziness	0	−	
ambiente del trabajo	labour snugness	0	−	
ambiente del trabajo	work atmosphere	3437	DICT	MAX
ambiente del trabajo	work cosiness	0	−	
ambiente del trabajo	work coziness	0	−	
ambiente del trabajo	work snugness	0	−	
campaña de propaganda	propaganda campaign	4337	DICT	MAX
campaña de propaganda	propaganda expedition	2	−	
campaña publicitaria	advertising campaign	70816	DICT	MAX
campaña publicitaria	advertising expedition	3	−	
campaña publicitaria	advertizing campaign	150	DICT	
campaña publicitaria	advertizing expedition	0	−	
campeón mundial	world champion	143343	DICT	MAX
campeón mundial	worldwide champion	868	−	
campeonato mundial	world championship	121676	DICT	MAX
campeonato mundial	worldwide championship	53	−	
campo de concentración	concentration camp	26532	DICT	MAX
campo de concentración	concentration country	19	−	
campo de concentración	concentration countryside	0	−	
campo de concentración	concentration field	575	−	

campo de concentración	concentration provinces	0	–	

Spanish term	*English candidate*	*AltaVista count*	*gold standard*	*most frequent*
campo de fuerzas	energy camp	769	–	
campo de fuerzas	energy country	451	–	
campo de fuerzas	energy countryside	6	–	
campo de fuerzas	energy field	20968	–	MAX
campo de fuerzas	energy provinces	8	–	
campo de fuerzas	force camp	920	–	
campo de fuerzas	force country	292	–	
campo de fuerzas	force countryside	3	–	
campo de fuerzas	force field	16390	DICT	
campo de fuerzas	force provinces	21	–	
campo de fuerzas	power camp	103	–	
campo de fuerzas	power country	501	–	
campo de fuerzas	power countryside	10	–	
campo de fuerzas	power field	3301	–	
campo de fuerzas	power provinces	83	–	

campo de fuerzas	strength camp	515	−	
campo de fuerzas	strength country	259	−	
campo de fuerzas	strength countryside	0	−	
campo de fuerzas	strength field	556	−	
campo de fuerzas	strength provinces	7	−	
campo de fuerzas	vigor camp	1279	−	
campo de fuerzas	vigor country	29	−	
campo de fuerzas	vigor countryside	2	−	
campo de fuerzas	vigor field	97	−	
campo de fuerzas	vigor provinces	0	−	
campo de fuerzas	vigour camp	73	−	
campo de fuerzas	vigour country	1	−	
campo de fuerzas	vigour countryside	0	−	
campo de fuerzas	vigour field	3	−	
campo de fuerzas	vigour provinces	0	−	
campo de fuerzas	violence camp	1259	−	
campo de fuerzas	violence country	369	−	
campo de fuerzas	violence countryside	0	−	
campo de fuerzas	violence field	179	−	
campo de fuerzas	violence provinces	4	−	

Spanish term	***English candidate***	***AltaVcount***	***Gold stand***	most freq
campo de fútbol	football camp	4899	–	
campo de fútbol	football country	199	–	
campo de fútbol	football countryside	4	–	
campo de fútbol	football field	27967	DICT	MAX
campo de fútbol	football provinces	0	–	
campo de fútbol	soccer camp	4437	–	
campo de fútbol	soccer country	114	–	
campo de fútbol	soccer countryside	1	–	
campo de fútbol	soccer field	13944	–	
coleccionista de monedas	coin collector	7165	DICT	MAX
coleccionista de monedas	currency collector	255	–	
coleccionista de sellos	seal collector	24	–	
coleccionista de sellos	stamp collector	8655	DICT	MAX
collar de perlas	pearl collar	94	–	
collar de perlas	pearl necklace	9234	DICT	MAX
color de camuflaje	camouflage color	236	–	
color de camuflaje	camouflage colour	272	DICT	
color de camuflaje	camouflage paint	617	–	MAX
columna conmemorativa	commemorative column	37	–	
columna conmemorativa	commemorative pillar	18	–	
columna conmemorativa	memorial column	128	DICT	MAX
columna conmemorativa	memorial pillar	74	–	

[1] http://info.ox.ac.uk/bnc

[2] http://www. AltaVista.com

[3]This not to say that noise does not exist, or that every linguistic utterance appearing on the Web is immediately validated by its simple presence. For example, the canonical

counter−example of "colorless green'" can be found 337 times via AltaVista. But now that *valid utterances* do occur thousands of times on the Web, the impact of such self−reference generated noise is diminished.

[4] Though the morphological variant *working group*, found 530124 times is the preferred (as well as the more frequently occurring) translation.

[5] We used the Basic Multilingual Lexicon http://www.icp.grenet.fr/ELRA/

cata/text_det.html#basmullex, available from the ELRA as our dictionary. This dictionary contains 37,600 senses translated across five languages: English, French, Spanish, Italian, and German. We used the German−English and Spanish−English parts.

[6] By preferred, we mean what our dictionary gives as a translation of the term. One might raise the question about whether the dictionary might be wrong in this sense, but to remain objective, we considered that the dictionary was always right.

[7] Decomposed using techniques described in (Schiller, 1996).

[8] The page frequency. AltaVista returns a count of the number of times that a word or expression (enclosed in quotes), has been seen on the pages that it indexes, and the number of WWW pages that contain the term. The counts given in this paper were calculated in the beginning of 1999, and correspond to the number of pages found.

[9] Recent tests from June 1999 estimate that AltaVista indexes about 15% of the static Web pages accessible on the Web.

2001. Very Large Lexical Databases: A tutorial. ACL Workshop, Toulouse, France. J. Pustejovsky, A. Rumshisky, and J. Castano. 2002. Rerendering Semantic Ontologies: Automatic Extensions to UMLS through Corpus Analytics.Â M. Rooth, S. Riezler, D. Prescher, G. Carroll, and F. Beil. 1999. Inducing a semantically annotated lexicon via EMbased clustering. In Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL99), Maryland. J. Sinclair, P. Hanks, and et al. 1987. The Collins Cobuild English Language Dictionary.