# Word Knowledge Acquisition, Lexicon Construction and Dictionary Compilation

Antonio Sanfilippo*
SHARP Laboratories of Europe Ltd.
Oxford Science Park, Oxford OX4 4GA, UK
antonio@sharp.co.uk

## Abstract

We describe an approach to semiautomatic lexicon development from machine readable dictionaries with specific reference to verbal diatheses, envisaging ways in which the results obtained can be used to guide word classification in the construction of dictionary databases.

## 1 Introduction

The acquisition and representation of lexical knowledge from machine-readable dictionaries and text corpora have increasingly become major concerns in Computational Lexicography/Lexicology. While this trend was essentially set by the need to maximize cost-effectiveness in building large scale Lexical Knowledge Bases for NLP (LKBs), there is a clear sense in which the construction of such knowledge bases also caters to the demand for better dictionaries. Currently available dictionaries and thesauri provide an undoubtedly rich source of lexical information, but often omit or neglect to make explicit salient syntactic and semantic properties of word entries. For example, it is well known that the same verb sense can appear in a variety of subcategorization frames which can be related to one another through valency alternations (diatheses). Some dictionaries provide subcategorization information by means of *grammar codes*, as shown below for the "sail" sense of the verb *dock* in LDOCE — *Longman's Dictionary of Contemporary English* (Procter, 1978).

(1) **dock**[4] *v* [T1;I0: (*at*)], ...

The codes [T1;I0:(*at*)] indicate that the verb can be either transitive or intransitive with the possible addition of an oblique complement introduced by the preposition *at*:

(2) a. [T1 (*at*)]: Kim docked his ship (at Glasgow)
b. [I0 (*at*)]: The ship docked (at Glasgow)

Unfortunately, an indication of diatheses which relate the various occurrences of the verb to one another is rarely provided. Consequently, if we were to use the grammar code information found in LDOCE to create verb entries in an LKB by automatic conversion we would construct four seemingly unrelated entries for the verb *dock* (see §3). Inadequacies of this kind may be redressed through semiautomatic techniques which make it possible to supply information concerning amenability to diathesis alternations so as to avoid expanding distinct entries for related uses of the same verb. This practice would allow us to develop an LKB from dictionary databases which offers a more complete and linguistically refined repository of lexical information than the source databases. Such an LKB would be used to generate lexical components for NLP systems, and could also be integrated into a lexicographer's workstation to guide word classification.

*The research reported in this paper was carried out within the ACQUILEX project. I am indebted to Ted Briscoe, Ann Copestake and Pete Whitelock for helpful comments.

## 2 The ACQUILEX Lexicon Development Environment

Our points of departure are the tools for lexical acquisition and knowledge representation developed as part of the ACQUILEX project ('The Acquisition of Lexical Knowledge for NLP Systems').

The ACQUILEX Lexicon Development Environment uses typed graph unification with inheritance as its lexical representation language (for details, see Copestake (1992), Sanfilippo & Poznański (1992), and papers by Copestake, de Paiva and Sanfilippo in Briscoe *et al.* (1993)). It allows the user to define an inheritance hierarchy of types with associated restrictions expressed in terms of attribute-value pairs as shown in Fig 1, and to create lexicons where such types are used to create lexical templates which encode word-sense specific information extracted from MRDs such as the one in Fig 2. (Bold lowercase is used for types, caps for attributes, and boxes enclosing types indicate total omission of attribute-value pairs. Details concerning the encoding of verb syntax and semantics can be found in Sanfilippo (1993).)

Feature Structure (FS) descriptions of word senses such as the one in Fig 2 are created semiautomatically through a program which converts syntactic and

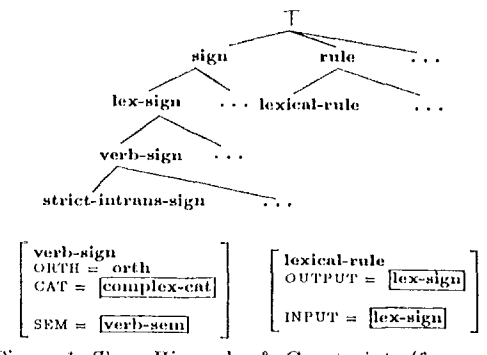



Figure 1: Type Hierarchy & Constraints (fragment).

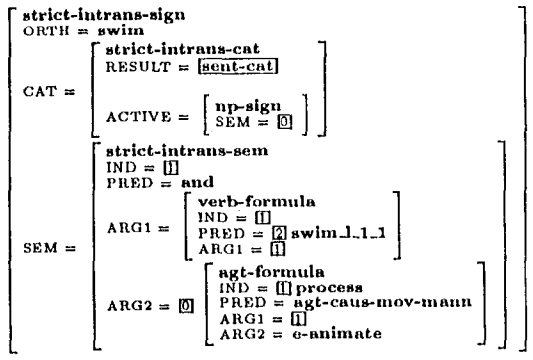



Figure 2: LKB Entry for *swim* (simplified).

semantic specifications encoded in MRDs into LKB types. For example, the choice of LKB types used in the characterization of the verb *swim* above was induced from the syntactic and semantic codes found in LDOCE and the *Longman Lexicon of Contemporary English* (LLOCE, McArthur 1980). In LDOCE, the first sense of the verb *swim* is marked as a strict intransitive verb ([I0]) whose subject is animate (`(box ----0)`); in LLOCE, the same verb sense is semantically classified as a movement verb with manner of motion specified (M19):

(3) **swim 1 (1)**

| | |
|---|---|
| LDOCE | [I0] `(box ----0)` ... |
| LLOCE | M19 – Particular ways of moving |

The MRD-to-LKB equivalences induced by the conversion algorithm are as shown in (4) where **agt-cause-move-manner** indicates that the subject participant relation implies self-induced movement with manner specified.

(4)

| | | |
|---|---|---|
| [I0] | → | **strict-intrans-sign** |
| `(box ----0)` | → | [CAT : ACTIVE : SEM : ARG2 = **e-animate**] |
| M19 | → | [CAT : ACTIVE : SEM : PRD = **agt-cause-move-manner**] |
| [I0], M19 | → | [SEM : IND = **process**] |

## 3 Verbal Diatheses and Lexical Acquisition

In the example discussed above, MRD-to-LKB conversion is relatively straightforward: a single LKB entry is created for *swim* since a single grammar code is found in the MRD sources used. Where a verb-sense entry gives more than one grammar code, however, the question arises whether or not each grammar code should be mapped into a distinct LKB entry. For example, the codes given in LDOCE for the verb *dock* (see (1)) could potentially be used to derive four LKB verb entries:

(5)

| | LKB Type | Example |
|---|---|---|
| a. | **strict-trans-sign** | *Kim docked the boat* |
| b. | **obl-trans-sign** | *Kim docked the boat at Southampton* |
| c. | **strict-intrans-sign** | *The boat docked* |
| d. | **obl-intrans-sign** | *The boat docked at Southampton* |

Notice, however, that in this case the creation of four distinct LKB entries is unnecessary insofar as the use of the verb exemplified in (5b) contains enough information to derive the remaining uses of the verb through lexical rules which progressively reduce the verb's valency by dropping the subject and/or prepositional argument(s). Such a step would be linguistically motivated in that it establishes a clear link between alternative uses of the same verb sense. Moreover, compact representation of verb use extensions is desirable from an engineering perspective as it reduces the size of the lexicon, allowing verb use expansion to be delayed till parsing time. This practice can be made to facilitate the resolution of lexical ambiguity by enforcing selective application of lexical rules (Copestake & Briscoe, 1994).

Compact representation of verb use extensions due to valency alternations requires that a note of all applicable lexical rules be made in each kernel entry. In choosing **obl-trans-sign** as the LKB type for *dock*, for example, specifications would be added saying that the verb is amenable to the *causative-inchoative* alternation relating agentive and agentless uses ((5a,b) vs. (5c,d)), and the *path* alternation pertaining to the omission of the prepositional argument ((5a,c) vs. (5b,d)). In addition, the *path* alternation would have to be specified as to whether it preserves amenability to a telic interpretation (accomplishment or achievement) of the event described by the verb or not. For example, the omission of the goal argument for a verb such as *drive, push* or *carry* induces an atelic (process) interpretation as indicated by incompatibility with a terminative adverbial:

(6) a. John drove his car to London in one hour
b. John drove his car (*in one hour)

Within a (partial) decompositional approach to verb semantics (Talmy, 1985; Jackendoff, 1990; Sanfilippo, 1993; Sanfilippo *et al.*, 1992)), this contrast can be explained with reference to the meaning component *path*. In (6a), the goal argument (*to London*) fixes a final bound for the path along which the driving event takes place. Assuming that the compositional meaning of the sentence involves establishing a homomorphism between the event described by the verb and the path along which such an event takes place (Dowty, 1991; Sanfilippo, 1991), it follows that with an unbounded path (e.g. (6b)) only a process interpretation is possible, whereas with a bounded path (e.g. (6a)) a telic interpretation is more likely. By contrast, the omission of the goal argument with verbs such as *deliver, bring, dock* and *send* does not inhibit amenability to a telic interpretation, e.g.

(7) We can deliver the goods (to your door) in one hour

Our aim, then, was to capture regularities across distinct uses of the same verb sense by relating the subcategorization frames relative to these uses via regular syntactic and semantic changes. To assess the feasibility of this approach, we augmented the MRD-to-LKB conversion code with facilities which make it possible to infer amenability to specific diathesis alternations from occurrence of multiple grammar codes and their associated semantic codes in the MRDs. To improve on the informational content of LDOCE grammar codes, we used an intermediate dictionary semiautomatically derived from LDOCE (LDOCE_Inter) where the subcategorization information inferrable from grammar codes and other orthographic conventions was made more explicit (Boguraev & Briscoe, 1989; Carroll & Grover, 1989). Semantic information about verb classes was obtained by mapping across LDOCE and LLOCE so as to augment LDOCE queries with thesaurus information, i.e. semantic codes (Sanfilippo & Poznański, 1992).

Syntactic and semantic information relative to verb senses was extracted through special functions which operate on pointers to dictionary entries. The extracted info was used to generate FS representations of word senses. The conversion process was carried out in such a way that whenever multiple subcategorization frames were found in association with a verb sense, only those which could not be derived via diathesis alternation were expanded into LKB entries. For example, the LDOCE_Inter entry for *dock* gives four subcategorization frames:

```
(dock)
 (((Cat V) (Takes NP) (Type 1))
  ((Cat V) (Takes NP PP) (Type 2) (PFORM at))
  ((Cat V) (Takes NP NP) (Type 2 Transitive))
  ((Cat V) (Takes NP NP PP) (Type 3) (PFORM at)))
```

In this case, the four uses of the verb can all be derived from the last one through application of the causative-inchoative and bounded-path alternations mentioned above; all that needs doing is to mark what diatheses are possible in the LKB entry derived, e.g.

(8)
$$\left[\begin{array}{l}\textbf{obl-trans-sign}\\ \text{ORTH} = \textbf{dock}\\ \\ \ldots\text{:M-FEATS:DIATHESES} = \left[\begin{array}{l}\textbf{trans-obl-diatheses}\\ \text{TRANS-ALT} = \textbf{caus-inch}\\ \text{OBL-ALT} = \textbf{b-path}\end{array}\right]\end{array}\right]$$

The algorithm which guides this process checks whether information regarding diathesis alternations can be inferred from dictionary entries in the MRD sources or must be manually supplied. In performing this check, subcategorization options relative to a given verb sense which can be inferred from a more informative subcategorization frame are ignored. This technique was successfully employed in semiautomatic derivation of lexicons for 360 movement verbs yielding over 500 additional possible expansions by application of lexical rules.

## 4 Verbal Diatheses and Knowledge Representation

To encode amenability to verbal diatheses, the feature DIATHESES was introduced as an extension of the morphological features associated with verbs (see (8)). This feature takes as value the type **alternations** which is in turn defined as having a variety of specialized types according to which diathesis alternations are admissible for each choice of verb type (e.g. intransitive, transitive, ditransitive), as shown in Figure 3 (see next page). The following table provides examples of the diatheses referred to in Fig 3.

| DIATHESIS | EXAMPLE |
|---|---|
| **caus-inch** | *Kim broke the glass* vs. |
| | *the glass broke* |
| **middle** | *Kim scares Sally* vs. |
| | *Sally scares easily* |
| **indef-obj** | *John ate a sandwich* vs. |
| | *John ate* |
| **def-obj** | *John did not notice the sign* vs. |
| | *John did not notice* |
| **recip** | *Kim met Bill* vs. |
| | *Kim and Bill met* |
| **pass** | *Bill read the Guardian* vs. |
| | *The Guardian was read by Bill* |
| **b-path** | *Kim returned the book to Sue* vs. |
| | *Kim returned the book* |
| | *Kim came away vs.* |
| | *Kim came* (particle alternation) |
| **u-path** | *Kim swam across the river* vs. |
| | *Kim swam* |
| | *Kim walked away* vs. |
| | *Kim walked* (particle alternation) |
| **to/fore** | *John brought a book to/for Sue* vs. |
| | *John brought Sue a book* |

Diathesis alternations are enforced by means of lexical rules which, on par with all other information structures in the LKB, are hierarchically arranged, as shown in Fig 4 with reference to the bound and unbound path alternations for intransitive verbs. Lexical rules

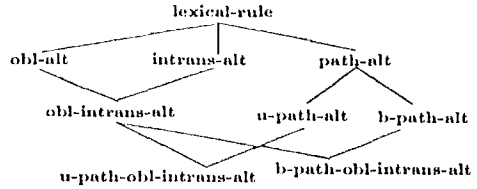
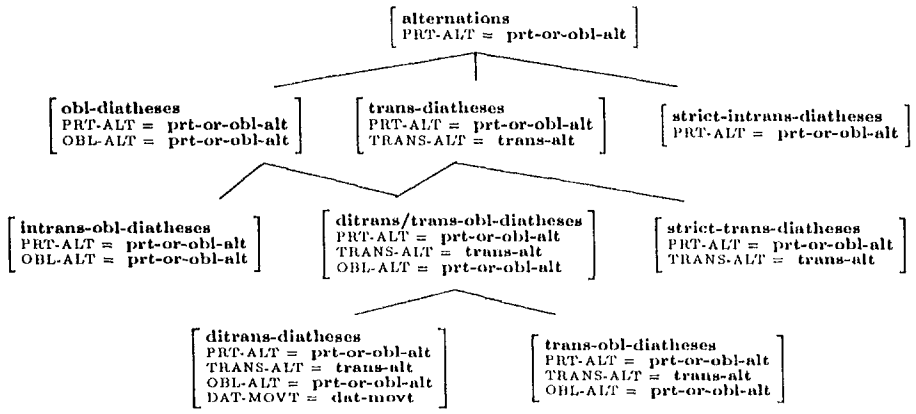



Figure 4: Lexical Rule Hierarchy (fragment).

enforcing diathesis alternations may involve a variety of syntactic, semantic and orthographic changes. For example, the **u-path-obl-intrans-alt** rule shown in Fig 5 below takes as input an FS of type **obl-intrans-sign** which represents a verb describing a non-stative eventuality (**dyn-eve**) whose subject participant (with semantics $\boxed{1}$) is implied as moving along a directed path (**th-move-dir**) the endpoint of which is specified by the oblique argument (**pp-sign**), e.g. the use of *swim* in *Kim swam across the river*. The output is an FS representing a strict intransitive verb (**strict-intrans-sign**) which describes a process and whose subject participant is like that of the input with the directed path specification removed (**th-move** instead of **th-move-dir**), e.g. *swim* in *Kim swam*).

## 5 Using the LKB to Guide Dictionary Compilation

There are at least two ways in which an LKB such as the one developed in ACQUILEX offers the means to




**trans-alt** $\sqsubseteq$ **caus-inch, middle, indef-obj, def-obj, recip, pass**
**prt-or-obl-alt** $\sqsubseteq$ **b-path, u-path**
**dat-movt** $\sqsubseteq$ **to, for**

Figure 3: Verbal Diatheses Hierarchy

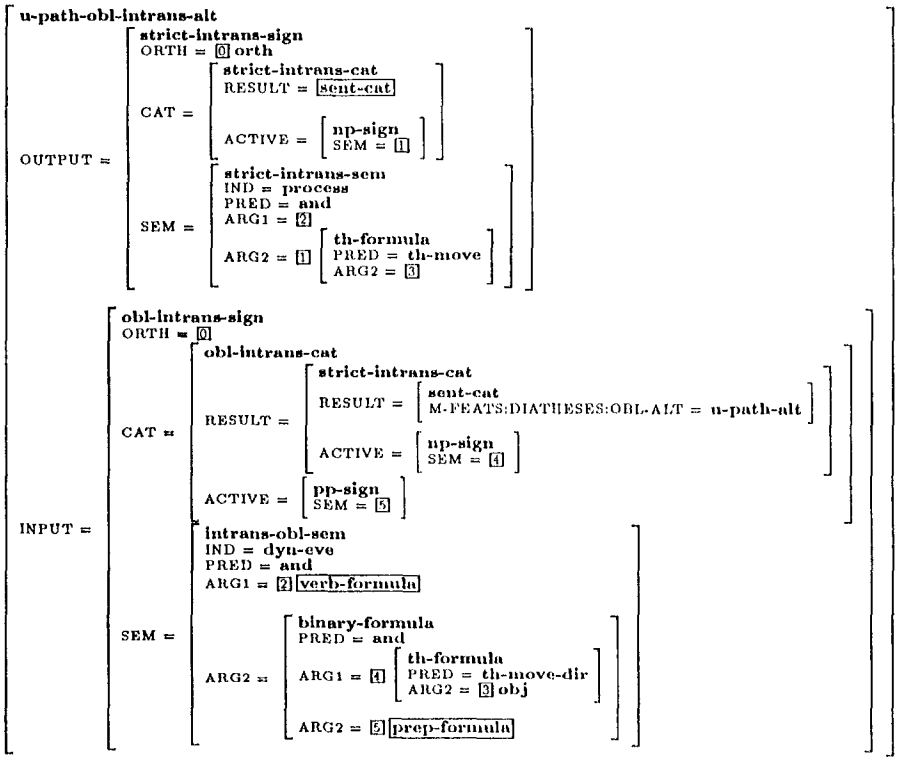



Figure 5: The "unbounded path" lexical rule for intransitives

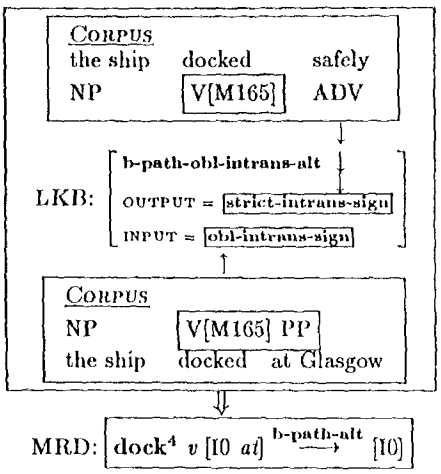



Figure 6: TBA

facilitate word classification in the compilation of new lexical databases.

First, the links between LKB types and dictionary entries established in the conversion stage can be used to run consistency checks on the MRD sources and to supply missing information or correct errors. This offers an efficient and cost-effective way of generating improved versions of the same dictionary.

Second, the types associated with specific word classes can be made to guide lexical aquisition from corpora when creating new dictionaries. It is now widely recognized that corpora are indispensable in the acquisition of lexical information relating to issues of usage such as the range and frequency of different patterns of syntactic realization. The availability of software tools for partial analysis of texts (e.g. morphological and semantic tagging, phrasal parsing etc.) has increased significantly the utility of corpora in lexical acquisition by providing ways to structure the information contained in them (see Briscoe (1991) and references therein). Further advances yet can be made by using LKB types to classify words in text corpora. Suppose, for example, we linked the input and output of lexical rules to semantically tagged subcategorization frames extracted from bracketed corpora (Poznański & Sanfilippo, 1993). As indicated in Fig 6, this would allow us to assess which alternations might be of interest in establishing regular verb sense/usage shifts. Such an assessment would provide an effective way to drive verb categorization from corpora in the domain of valency alternations.

## 6 Final Remarks

A key element in our approach to lexical acquisition and representation of verbal diathesis concerns the use of semantics constraints in formulating MRD queries and characterizing FS descriptions. This practice ensures that the results achieved in this work for motion verbs can be suitably extended to other semantic verb classes. For example, the class of verbs which undergo "extraposition" — e.g. *That Kim left early bothers Sue* vs. *It bothers Sue that Kim left early* — can be identified by using semantic constraints on MRD queries which identify psychological verbs with stimulus subject such as *bother, please,* etc. (Sanfilippo & Poznański, 1992). This approach provides an effective way of employing semiautomatic extraction of information from MRDs for lexicon construction, and it facilitates word classification from text corpora when compiling new dictionary databases.

## References


Briscoe, T. (1991) Lexical Issues in Natural Language Processing. In Klein, E. & F. Veltman (eds.). *Natural Language and Speech*, Springer-Verlag, 39-68.

Briscoe, T., A. Copestake and V. de Paiva (1993) *Default Inheritance within Unification-Based Approaches to the Lexicon*, CUP.

Boguraev, B. & T. Briscoe (1989) Utilising the LDOCE Grammar Codes. In Boguraev, B. & Briscoe, T. (eds.) *Computational Lexicography for Natural Language Processing*. Longman, London.

Carroll, J. & C. Grover (1989) The Derivation of a Large Computational Lexicon for English from LDOCE. In Boguraev, B. & Briscoe, T. (eds.) *Computational Lexicography for Natural Language Processing*. Longman, London.

Copestake, A. (1992) The ACQUILEX LKB: Representation Issues in Semi-Automatic Acquisition of Large Lexicons. In *Proceedings of the 3rd Conference on Applied Natural Language Processing*.

Copestake, A. and T. Briscoe (1994) Semi-productive Polysemy and Sense Extensions. Ms. Computer Laboratory University of Cambridge, Xerox Park (Menlo Park) and Rank Xerox Research Laboratory (Grenoble).

Dowty, D. (1991) Thematic Proto-Roles and Argument Selection. Language 67, pp. 547-619.

Jackendoff, R. (1990) *Semantic Structures*. MIT Press, Cambridge, Mass.

McArthur, T. (1981) *Longman Lexicon of Contemporary English*. Longman, London.

Poznański & Sanfilippo (1993) Detecting Dependencies between Semantic Verb Subclasses and Subcategorization Frames in Text Corpora. In B. Boguraev & J. Pustejovsky (eds) *Acquisition of Lexical Knowledge from Text*, Proceedings of a SIGLEX workshop, ACL-93, Ohio.

Procter, P. (1978) *Longman Dictionary of Contemporary English*. Longman, London.

Sanfilippo, A. (1991) Thematic and Aspectual Information in Verb Semantics. *Belgian Journal of Linguistics*, 6.

Sanfilippo, A. (1993) LKB Encoding of Lexical Knowledge. In Briscoe, T., A. Copestake and V. de Paiva (eds.).

Sanfilippo, A & V. Poznański (1992) The Acquisition of Lexical Knowledge from Combined Machine-Readable Dictionary Sources. In *Proceedings of the 3rd Conference on Applied Natural Language Processing*, Trento.

Sanfilippo, A., T. Briscoe, A. Copestake, M. Martí, M. Taulé and A. Alonge (1992) Translation Equivalence and Lexicalization in the ACQUILEX LKB. Proceedings of TMI-92, Montreal, Canada.

Talmy, L. (1985) Lexicalization Patterns: Semantic Structure in Lexical Form. In Shopen, T. (ed) *Language Typology and Syntactic Description 3. Grammatical Categoriesv and the Lexicon*, CUP.