The Acquisition of Lexical Knowledge from Combined Machine-Readable Dictionary Sources

Antonio Sanfilippo*
Computer Laboratory, University of Cambridge
New Museum Site, Pembroke Street
Cambridge CB2 3QG, UK
Antonio.Sanfilippo@cl.cam.ac.uk

Victor Poznański
IRIDIA, Université Libre de Bruxelles
50 Avenue F. Roosevelt, CP 194/6
B-1050 Bruxelles, Belgique
vicpoz@is1.vub.ac.be

Abstract

This paper is concerned with the question of how to extract lexical knowledge from Machine-Readable Dictionaries (MRDs) within a lexical database which integrates a lexicon development environment. Our long term objective is the creation of a large lexical knowledge base using semiautomatic techniques to recover syntactic and semantic information from MRDs. In doing so, one finds that reliance on a single MRD source induces inadequacies which could be efficiently redressed through access to combined MRD sources. In the general case, the integration of information from distinct MRDs remains a problem hard, perhaps impossible, to solve without the aid of a complete, linguistically motivated database which provides a reference point for comparison. Nevertheless, advances can be made by attempting to correlate dictionaries which are not too dissimilar. In keeping with these observations, we describe a software package for correlating MRDs based on sense merging techniques and show how such a tool can be employed in augmenting a lexical knowledge base built from a conventional MRD with thesaurus information.

1 Introduction

Over the last few years, the utilization of machine readable dictionaries (MRDs) in compiling lexical components for Natural Language Processing (NLP) systems has awakened the interest of an increasing number of researchers. This trend has largely arisen from the recognition that access to a large scale Lexical Knowledge Base (LKB) is a *sine qua non* for real world applications of NLP systems. Computer-aided acquisition of lexical knowledge from MRDs can be made to satisfy this prerequisite in a manner which is both time- and cost-effective. However, to maximize the utility of MRDs in NLP applications, it is necessary to overcome inconsistencies, omissions and the occasional errors which are commonly found in MRDs (Atkins *et al.*, 1986; Atkins, 1989; Akkerman, 1989; Boguraev & Briscoe, 1989). This goal can be partly achieved by developing tools which make it possible to correct errors and inconsistencies contained in the information structures automatically derived from the source MRD before these are stored in the LKB or target lexicon (Carroll & Grover, 1989). This technique is nevertheless of little avail in redressing inadequacies which arise from lack of information. In this case, manual supply of the missing information would be too time- and labour-intensive to be desirable. Moreover, the information which is missing can be usually obtained from other MRD sources. Consider, for example, a situation in which we wanted to augment lexical representations available from a conventional dictionary with thesaurus information. There can be little doubt that the integration of information from distinct MRD sources would be far more convenient and appropriate than reliance on manual encoding.

In the general case, the integration of information from distinct MRD sources for use within a lexicon development environment is probably going to remain an unsolved problem for quite some time. This is simply because dictionaries seldom describe the same word using the same sense distinctions. Consequently, the integration of information from distinct MRD sources through simple word-sense matches is likely to fail in a significant number of instances (Calzolari & Picchi, 1986; Atkins 1987; Klavans 1988; Boguraev & Pustejovsky 1990). Indeed, Atkins & Levin (1990) have suggested that the task of mapping MRDs onto each other is so complex that the creation of a complete 'ideal' database which provides a reference point for the MRD sources to be integrated may well be considered as an essential prerequisite. However, when dealing with MRD sources which use entry definitions which are not too dissimilar, a correlation technique based on word sense merging can be made to yield useful results, given the appropriate tools. Although sense matching across dictionaries in this case too is prone to errors, there are several reasons why the effort is worthwhile. First, the number of correct sense matches across MRD sources in this case is guaranteed to be high. Second, there are many instances in which an incorrect sense-to-sense match does not affect the final

*The research reported in this paper was carried out at the Computer Laboratory (University of Cambridge) as part of an ongoing study on the acquisition of lexical knowledge from machine-readable dictionaries within the ACQUILEX project, ESPRIT BRA 3030.

result since the information with respect to which a sense correlation is being sought may generalize across closely related word senses. Third, a close inspection of infelicitous matches provides a better understanding of specific difficulties involved in the task and may help develop solutions. Finally, the construction of a linguistically motivated database aimed to facilitate the interrogation of combined MRD sources would be highly enhanced by the availability of new tools for lexicology. Partial as it may be, the possibility of mapping MRDs onto each other through sense-merging techniques should thus be seen as a step forward in the right direction.

The goal of this paper is to describe a software package for correlating word senses across dictionaries which can be straightforwardly tailored to an individual user's needs, and has convenient facilities for interactive sense matching. We also show how such a tool can be employed in augmenting a lexical knowledge base built from a conventional MRD with thesaurus information.

2 Background

Our point of departure is the Lexical Data Base (LDB) and LKB tools developed at the Computer Laboratory in Cambridge within the context of the ACQUILEX project. The LDB (Carroll, 1990) gives flexible access to MRDs and is endowed with a graphic interface which provides a user-friendly environment for query formation and information retrieval. It allows several dictionaries to be loaded and queried in parallel.

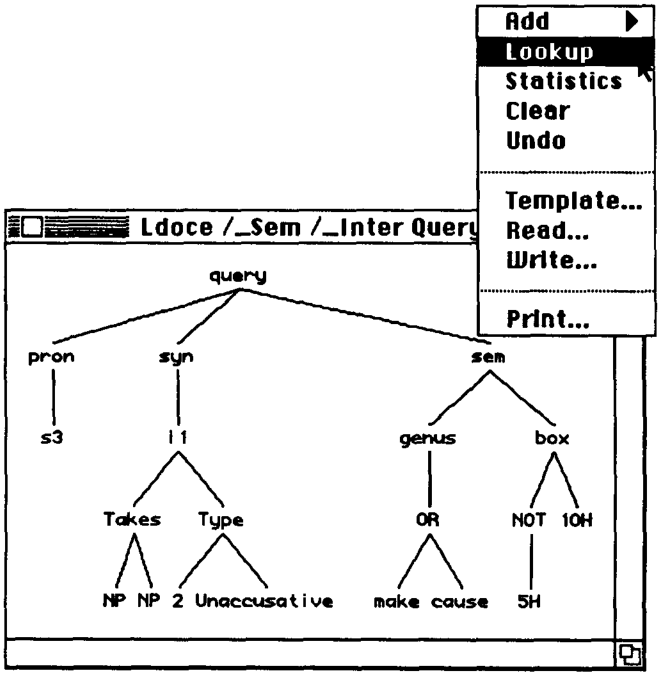


Figure 1: LDB query from combined MRD sources (a main MRD and two derived ones)

Until recently, this facility has been used to extract information from combined MRD sources which included a main dictionary and a number of dictionaries derived from it. For example, the information needed to build LKB representations for English verbs (see below) was partly obtained by running LDB queries which combined information from the Longman Dictionary of Contemporary English (LDOCE) and two other dictionaries derived from LDOCE: LDOCE_Inter and LDOCE_Sem. LDOCE_Inter was derived by a translation program which mapped the grammar codes of LDOCE entries into theoretically neutral intermediate representations (Boguraev & Briscoe, 1989; Carroll & Grover, 1989). LDOCE_Sem was derived by extracting genus terms from dictionary definitions in LDOCE (Alshawi, 1989; Vossen, 1990). Figure 1 provides an illustrative example of an LDB query which combines these MRD sources. We used these LDB facilities for running queries from combined MRD sources which included more than one MRD — i.e. LDOCE and the Longman Lexicon of Contemporary English (LLOCE), a thesaurus closely related to LDOCE.

The LKB provides a lexicon development environment which uses a typed graph-based unification formalism as representation language. A detailed description of the LKB's representation language is given in papers by Copestake, de Paiva and Sanfilippo in Briscoe *et al.* (forthcoming); various properties of the system are also discussed in Briscoe (1991) and Copestake (1992). The LKB allows the user to define an inheritance network of types plus restrictions associated with them, and to create lexicons where such types are assigned to lexical templates extracted through LDB queries which give word-sense specific information. Consider, for example, the lexical template relative to the first LDOCE sense of the verb *delight* in (1) where sense specific information is integrated with a reference to the LKB type STRICT-TRANS-SIGN which provides a general syntactic and semantic characterization of strict transitive verbs.[1]

```
(1) delight L_2_1
    STRICT-TRANS-SIGN
    <cat:result:result:m-feats:diathesis>=INDEF-OBJ
    <cat:result:result:m-feats:reg-morph>=TRUE
    <cat:active:sem:arg2> = E-HUMAN
    <sense-id:dictionary> = "LDOCE"
    <sense-id:ldb-entry-no> = "9335"
    <sense-id:sense-no> = "1".
```

When loaded into the LKB, the lexical template above will expand into a full syntactic and semantic representation as shown in Figure 2; this representation arises from integrating sense-specific information with the information structure associated with the type STRICT-TRANS-SIGN.[2]

[1] The type specification INDEF-OBJ in (1) corresponds to the LDB value Unaccusative (see Figure 1) and marks transitive verbs which are amenable to the *indefinite object* alternation, e.g. *a book which is certain to delight them* vs. *a book which is certain to delight.* Information concerning diathesis alternations is also derived from LDOCE_Inter. The value TRUE for the attribute reg-morph indicates that *delight* has regular morphology. OBJ and E-HUMAN are sorted variables for individual objects.

[2] According to the verb representation adopted in the LKB (Sanfilippo, forthcoming), verbs are treated as predicates of eventualities and thematic roles as relations between eventualities and individuals (Parsons, 1990). The semantic content

```
[strict-trans-sign
 ORTH:delight
 CAT: RESULt: [strict-intrans-cat
                RESULT: [sent-cat
                          CAT-TYPE: sent
                          M-FEATS: [sent-m-feats
                                     DIATHESIS: indef-obj
                                      REG-MORPH: true]]
                DIRECTION: forward
                ACTIVE: [np-sign
                          SEM: <1>]]
      DIRECTION: backward
     ACTIVE: [dir-obj-np-sign
               SEM:<2>]]
 SEM:[ strict-trans-sem
       IND:<0>  =eve
       PRED:and
       ARG1:[verb-formula
              IND:<0>
              PRED: delight_l_2_1
              ARG1:<0>]
       ARG2:[binary-formula
              IND: <0>
              PRED: and
              ARG1: <1> =[p-agt-formula
                           IND: <0>
                           PRED:p-agt
                           ARG1:<0>
                           ARG2:obj]
              ARG2: <2> =[p-pat-formula
                           IND: <0>
                           PRED:p-pat
                           ARG1:<0>
                           ARG2:e-human ]]]
 SENSE-ID:[sense-id ]]
```

Figure 2: LKB entry for sense 1 of the verb *delight* in LDOCE

Lexical templates such as the one in (1) are generated through a user definable conversion function — a facility included in the LKB — which makes it possible to establish correspondences between information derived through LDB queries and LKB types. For example, information relative to selectional restrictions for transitive verbs (e.g. `e-human` and `obj` in (1)) is encoded by establishing a correspondence between the value for the individual variables of the subject and object roles in LKB representations and the values retrieved from the relevant LDOCE entry for box codes 5 and 10 (see Figure 1). Similarly, the assignment of verb types (e.g. `STRICT-TRANS-SIGN`) to verb senses is carried out by relating LKB types for English verbs — about 30 in the current implementation (Sanfilippo, forthcoming) — to subcategorization patterns retrieved from LDOCE_Inter. For example, if a verb sense in LDOCE_Inter were associated with the information in (2), the conversion function would associate the lexical template being generated with the type `STRICT-TRANS-SIGN`.

(2) `((Cat V) (Takes NP NP) ...)`

Needless to say, the amount of information specified in LKB entries will be directly proportional to the amount of information which can be reliably extracted through LDB queries. With respect to verbs, there are several ways in which the representations derived from templates such as the one in (1) can be enriched. In the simplest case, additional information can be recovered from a single MRD source either directly or through translation programs which allow the creation of derived dictionaries where information which is somehow contained in the source MRD can be made more explicit. This technique may however be insufficient or inappropriate to recover certain kinds of information which are necessary in building an adequate verb lexicon. Consider the specification of verb class semantics. This is highly instrumental in establishing subcategorization and regimenting lexically governed grammatical processes (see Levin (1989), Jackendoff (1990) and references therein) and should be thus included within a lexicon which supplied adequate information about verbs. For example, a verb such as *delight* should be specified as a member of the class of verbs which express emotion, i.e. psychological verbs. As is well known (Levin, 1989; Jackendoff, 1990), verbs which belong to this semantic class can be classified according to the following parameters:

- affect is positive (*admire, delight*), neutral (*experience, interest*) or negative (*fear, scare*)
- stimulus argument is realized as object and experiencer as subject, e.g. *admire, experience, fear*
- stimulus argument is realized as subject and experiencer as object, e.g. *delight, interest, scare*

Psychological verbs with experiencer subjects are 'noncausative'; the stimulus of these verbs can be considerec to be a 'source' to which the experiencer 'reacts emotively'. By contrast, psychological verbs with stimulu subjects involve 'causation'; the stimulus argument ma be consided as a 'causative source' by which the experiencer participant is 'emotively affected'. Six subtypes o psychological verbs can thus be distinguished accordin to semantic properties of the stimulus and experience arguments as shown in (3) where the verb *delight* is specified as belonging to one of these subtypes.

(3)

STIMULUS	EXPERIENCER	EXAMPLE
non-causative source	neutral, reactive, emotive	*experience*
non-causative source	positive, reactive, emotive	*admire*
non-causative source	negative, reactive, emotive	*fear*
neutral, causative source	neutral, affected, emotive	*interest*
positive, causative source	positive, affected, emotive	*delight*
negative, causative source	negative, affected, emotive	*scare*

Correct classification of members of the six ver classes in (3) through LDB queries which used as sourc a standard dictionary (e.g. LDOCE) is a fairly hopeles pursuit. Standard dictionaries are simply not equippe to offer this kind of information with consistency an exhaustiveness. Furthermore, the technique of creatin derived dictionaries where the information contained i a main source MRD is made more explicit is unhelpful in this case. For example, one approach would b

of roles is computed in terms of entailments of verb meanings which determine the most (p-agt) and least (p-pat) agentive event participants for each choice of predicate; see Figures 4 and 5 for illustrative example. This approach reproduces the insights of Dowty's and Jackendoff's treatments of thematic information (Dowty, 1991; Jackendoff, 1990) within a neo-Davidsonian approach to verb semantics (Sanfilippo, 1990).

to derive a dictionary from LDOCE where verbs are organized into a network defined by IS-A links using the general approach to taxonomy formation described by Amsler (1981). Such an approach would involve the formation of chains through verb definitions determined by the genus term of each definition. Unfortunately, the genus of verb definitions is often not specific enough to supply a taxonomic characterization which allows for the identification of semantic verb classes with consistency and exhaustiveness. In LDOCE, for example, the genus of over 20% of verb senses (about 3,500) is one of 8 verbs: *cause, make, be, give, put, take, move, have*; many of the word senses which have the same genus belong to distinct semantic verb classes. This is not to say that verb taxonomies are of no value, and in the final section we will briefly discuss an important application of verb taxonomies with respect to the assignment of semantic classes to verb senses. Nevertheless, the achievement of adequate results requires techniques which reclassify entries in the same source MRD(s) rather than making explicit the classification 'implicit' in the lexicographer's choice of genus term. Thesauri provide an alternative semantically-motivated classification of lexical items which is most naturally suited to reshape or augment the taxonomic structure which can be inferred from the genus of dictionary definitions. The LLOCE is a thesaurus which was developed from LDOCE and there is substantial overlap (although not identity) between the definitions and entries of both MRDs. We decided to investigate the plausibility of semi-automatic sense correlations with LDOCE and LLOCE and to explore the utility of the thesaurus classification for the classification of verbs in a linguistically motivated way.

3 DCK: A Flexible Tool for Correlating Word Senses Across MRDs

Our immediate goal in developing an environment for correlating MRDs was thus to merge word senses, and in particular verb senses, from LDOCE and LLOCE. More generally, our aim was to provide a Dictionary Correlation Kit (DCK) containing a set of flexible tools that can be straightforwardly tailored to an individual user's needs, along with a facility for the interactive matching of dictionary entries. Our DCK is designed to correlate word senses across pairs of MRDs which have been mounted on the LDB (henceforth *source-dict* and *destination-dict*) using a list of comparison heuristics. Entries from the source-dict and destination-dict are compared to yield a set of *correlation structures* which describe matches between word senses in the two dictionaries. A function is provided that converts correlation structures into entries of a derived dictionary which can be mounted and queried on the LDB.

3.1 General Functionality of DCK

Entry fields in the source-dict and destination-dict are compared by means of *comparators*. These are functions which take as input *normalized field information* extracted from the entries under analysis, and return two values: a score indicating the degree to which the two fields correlate, along with an *advisory datum* which indicates what kind of action to take. The objective of each match is to produce a correlation structure consisting of a source-dict sense and a set of destination-dict sense/score pairs representing possible matches. Prior to converting correlation structures into derived dictionary entries, the best match is selected for each correlation structure on the basis of the comparator scores. When there is ambiguity as to the best match, a *correlation dialog* window pops up that allows the user to peruse the candidate matches and manually select the best match (see Figure 3).

3.2 Customising DCK

Two categories of information must be provided in order to correlate a pair of new LDB-mounted dictionaries:

- functions which normalize dictionary-dependent field values, and
- dictionary independent comparators which provide matching heuristics.

Field values describing the same information may be labeled differently across dictionaries. For example, pronouns may be tagged as *Pron* in the part-of-speech field of one dictionary and *Pronoun* in part-of-speech field of another dictionary. It is therefore necessary to provide normalizing functions which convert dictionary-specific field values into dictionary-independent ones which can be compared using generic comparators.

Comparators take as arguments pairs of normalized field values relative to the senses of the two MRDs under comparison, and return a score associated with an advisory datum which indicates the course of action to be followed. The score and advisory datum provide an index of the degree of overlap between the two senses.

3.3 Determining the Best Sense

A correlation structure contains a list of destination-dict sense/score pairs which indicate possible matches with the corresponding source-dict sense. The most appropriate match can be determined automatically using two user-provided parameters:

1. the *threshold*, which indicates the minimal acceptable score that a comparator list must achieve for automatic sense selection, and
2. the *tolerance*, which is the minimum difference between the top two scores that must be achieved if the top sense with the highest score is to be selected.

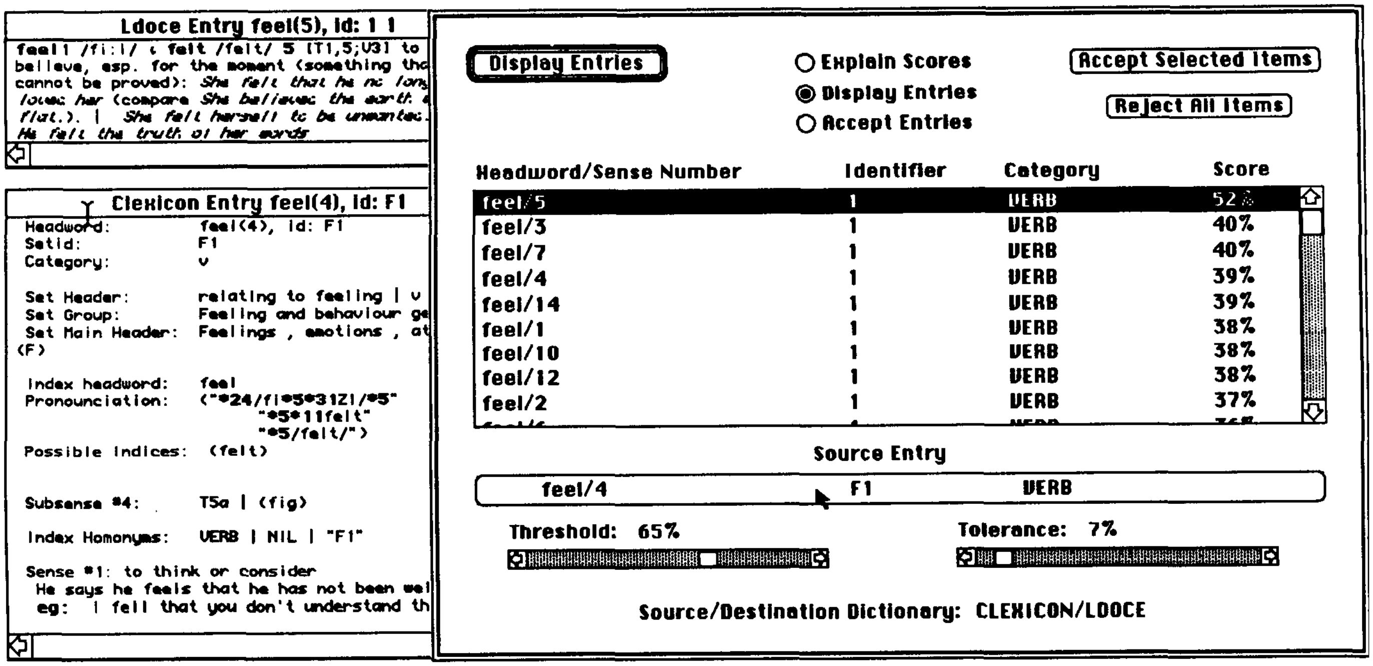
The sense/score pair with the highest score is automatically selected if:

A. the advisory datum provides no indication that the correlation should be queried,

B. the score relative to a single match exceeds the threshold, or

C. the score relative to two or more matches exceeds the threshold, and the difference between the top two scores exceeds the tolerance.

If either one of these conditions is not fulfilled, the correlation dialog is invoked to allow a manual choice to be made.



Ldoce Entry feel(5), Id: 1 1

feel1 /fi:l/ ι felt /felt/ 5 [T1,5;V3] to
believe, esp. for the moment (something tha
cannot be proved): *She felt that he no long*
loved her (compare *She believed the earth i*
flat.). | *She felt herself to be unwanted.*
He felt the truth of her words

Clexicon Entry feel(4), Id: F1

```
Headword:           feel(4), Id: F1
Setid:              F1
Category:           v

Set Header:         relating to feeling | v
Set Group:          Feeling and behaviour ge
Set Main Header:    Feelings , emotions , at
(F)

Index headword:     feel
Pronounciation:     ("*24/fi*5*31Z1/*5"
                          "*5*11felt"
                          "*5/felt/")
Possible Indices:   (felt)


Subsense *4:        T5a | (fig)

Index Homonyms:     VERB | NIL | "F1"

Sense *1: to think or consider
 He says he feels that he has not been wel
 eg:  I fell that you don't understand th
```

Headword/Sense Number	Identifier	Category	Score
feel/5	1	VERB	52%
feel/3	1	VERB	40%
feel/7	1	VERB	40%
feel/4	1	VERB	39%
feel/14	1	VERB	39%
feel/1	1	VERB	38%
feel/10	1	VERB	38%
feel/12	1	VERB	38%
feel/2	1	VERB	37%

Source Entry

feel/4	F1	VERB

Threshold: 65%

Tolerance: 7%

Source/Destination Dictionary: CLEXICON/LDOCE

Figure 3: Sample interaction with correlation dialog

3.4 The Correlation Dialog

The correlation dialog allows the user to examine correlation structures and select none, one or more destination-dict senses to be matched with the source-dict sense under analysis. A typical interaction can be seen in Figure 3. A scrollable window in the centre of the dialog box provides information about the destination-dict senses and their associated scores. Single clicking the mouse button on one or more rows makes them the current selection. The large button above the threshold and tolerance indicators summarizes source-dict sense information. Clicking on this button invokes an LDB query window which inspects the source-dict sense (cf. bottom left window in Figure 3).

The dialog can be in one of three modes:

- Explain Scores — the mode specific key pops up a window for each destination-dict sense in the current selection, explaining how each score was obtained from the comparators;
- Display Entries — the mode specific key invokes standard LDB browsers on the destination-dict senses in the current selection (cf. top-left window in Figure 3), and
- Accept Entries — the mode specific key terminates the dialog and accepts the current selection as the best match.

Two additional buttons on the top right of the dialog box allow the current selection to be accepted independent of the current mode, or all senses to be rejected (i.e. no match is found). At the bottom of the screen, two 'thermometers' allow the user to adjust the threshold and tolerance parameters dynamically.

4 Using DCK

We run DCK with LLOCE as source-dict and LDOCE as destination-dict to produce a derived dictionary, LDOCE_Link, which when loaded together with LDOCE would allow us to form LDOCE queries which integrated thesaurus information from LLOCE. The work was carried out with specific reference to verbs which express 'Feelings, Emotions, Attitudes, and Sensations' and 'Movement, Location, Travel, and Transport' (sets 'F' and 'M' in LLOCE). Correlation structures were derived for 1194 verb senses (over 1/5 of all verb senses in LLOCE) using as matching parameters degree of overlap in grammar codes, definitions and examples, as well as equality in headword and part-of-speech. After some trial runs, correlations appeared to yield best results when all parameters were assigned the same weight except the comparator for 'degree of overlap in examples' which was set to be twice as determinant than the others. Tolerance was set at 7% and threshold at 65%. The rate of interactions through the correlation dialog was about one for every 8-10 senses. It took about 10 hours running time on a Macintosh IIcx to complete the work, with less than three hours' worth of interactions.

A close examination of over 500 correlated entries disclosed an extremely low incidence of infelicitous matches (below 1%). In some cases, sense-matching inadequacies could be easily redressed without reassignment of correlation links. For example, DCK erroneously correlated the verb sense for *float* in LLOCE with the first verb sense of *float* in LDOCE. As shown in (4), the LLOCE sense refers only to the intransitive use of the verb, while the LDOCE sense refers to both transitive and intransitive uses of the verb (i.e. the LLOCE sense is subsumed by LDOCE sense).

(4) a LLOCE
float[I0] to stay on or very near the surface of a liquid, esp. water
b LDOCE
float^2 *v* **1** [I0;T1] to (cause to) stay at the top of a liquid or be held up in the air without sinking water

One way to redress this kind of inadequate match would be to augment DCK with a lexical rule module catering for diathesis alternations which made it possible to establish a clear relation between distinct syntactic realizations of the same verb. For example, the transitive and intransitive senses of *float* could be related to each other via the 'causative/inchoative' alternation. This augmentation would be easy to implement since information about amenability of verbs to diathesis alternations is recoverable from LDOCE_Inter, as shown below for *float* (`Ergative` is the term used in LDOCE_Inter to characterize verbs which like *float* are amenable to the causative/inchoative alternation).

```
(5) (float)
    (1 2)
    (2 1 <
       ((Cat V)(Takes NP NP)(Type 2 Ergative)))
    (2 2 ...
```

Notice, incidentally, that even though DCK yielded an incorrect sense correlation for the verb entry *float*, the information which was inherited by LDOCE from LLOCE through the correlation link was still valid. In LLOCE, *float* is classified as a verb whose set, group and main identifiers are: *floating-and-sinking*, *Shipping* and *Movement-location-travel-and-transport*. This information is useful in establishing the semantic class of both the transitive and intransitive uses of *float*. This is also true in those rare cases where DCK incorrectly preferred a sense match to another as shown below for the first LLOCE sense of *behave* which DCK linked to the third LDOCE sense rather than the first. Either sense of *behave* is adequately characterized by the set, group and main identifiers 'behaving', 'Feeling-and-behaviour-generally', and 'Feelings-emotions-attitudes-and-sensations' which LDOCE inherits from LLOCE through the incorrect sense correlation established by DCK.

(6) a LLOCE
behave 1 [L9] to do things, live, etc. usu in a stated way: *She behaved with great courage when her husband died* ...
b LDOCE
behave *v* **1** [L9] to act; bear oneself: *She behaved with great courage.* ... **3** [L9] (of things) to act in a particular way: *It can behave either as an acid or as a salt* ...
c DCK Correlation
LLOCE **behave 1** = LDOCE **behave 3**

5 LKB Encoding of Lexical Knowledge from Combined MRD Sources

LDOCE_Link was derived as a list of entries consisting of correlated LLOCE-LDOCE sense pairs plus an explicit reference to the corresponding set identifier in LLOCE, as shown in (7).

```
(7) ((amaze)
       (LL F237 < amaze < 0))
    ((desire) (2 1 < <)
       (SNO 1)(LL F6 < desire < 1)
       (SNO 2)(LL F6 < desire < 2))
```

Loading LDOCE with LDOCE_Link makes it possible to form LDOCE queries which include thesaurus information from LLOCE (i.e. the set identifiers). The integration of thesaurus information provides adequate means for developing a semantic classification of verbs. With respect to psychological verbs, for example, the set identifiers proved to be very helpful in identifying members of the six subtypes described in (3). The properties used in this classification could thus be used to define a hierarchy of thematic types in the LKB which gave a detailed characterization of argument roles. This is shown in the lattice fragment in Figure 4 where the underlined types correspond to the role types used to distinguish the six semantic varieties of psychological predicates.[3]

The correspondence between LLOCE set identifiers and the thematic role types shown in Figure 4 made it possible to create word-sense templates for psychological verbs from LDB queries which in addition to providing information about morphological paradigm, subcategorization patterns, diathesis alternations and selectional restrictions, supplied thematic restrictions on the stimulus and experiencer roles. Illustrative LKB entries relative to the six verb subtypes described in (3) are shown in Figure 5.

6 Final Remarks

Taking into consideration the size of the LLOCE fragment correlated to LDOCE (1/5 of LLOCE verb senses) and the results obtained, it seems reasonable to expect that this work should extend straightforwardly to other verbs as well as word senses of different category types.

As far as we were able to establish, the major limitation of the work carried out arises from the fact that the entries and senses per homonyn in the source dictionary were considerably fewer than those in the destination dictionary (e.g. 16,049 entries with 25,100 senses in LLOCE vs. 41,122 entries with 74,086 senses in LDOCE). Consequently, many senses of correlated verb entries as well as entire verb entries in LDOCE are bound to be left without a specification of thesaurus information. We are currently exploring the possibility of using verb taxonomies to extend the results of LLOCE-LDOCE correlations to those LDOCE entries and verb

[3] The labels 'p-agt' and 'p-pat' are abbreviations for 'proto-typical' agent and patient roles which subsume clusters of entailments of verb meanings which qualify the most and least agentive event participants for each choice of predicate (see footnote 2).

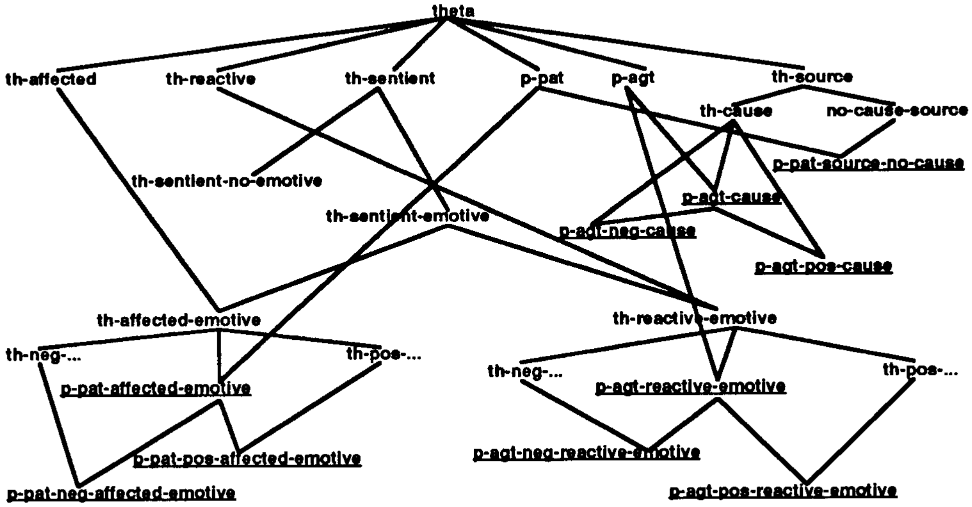


Figure 4: LKB types for thematic roles of psychological verbs

```
[strict-trans-sign
 ORTH:experience
 CAT strict-trans-cat
 SEM:[strict-trans-sem
      IND: <0> =eve
      PRED:and
      ARG1 verb-formula
      ARG2:[binary-formula
            IND: <0>
            PRED: and
            ARG1: <1> =[p-agt-formula
                        PRED: p-agt-reactive-emotive]
            ARG2: <2> = [p-pat-formula
                        PRED: p-pat-source-no-cause]]]]
```

```
[strict-trans-sign
  ORTH: admire
  CAT strict-trans-cat
  SEM: [strict-trans-sem
        IND: <0> =eve
        PRED: and
        ARG1: verb-formula
        ARG2: [binary-formula
               IND: <0>
               PRED: and
               ARG1: <1> =[p-agt-formula
                           PRED: p-agt-pos-reactive-emotive]
               ARG2: <2> = [p-pat-formula
                           PRED: p-pat-source-no-cause]]]]
```

```
[strict-trans-sign
 ORTH:fear
 CAT strict-trans-cat
 SEM:[strict-trans-sem
      IND: <0> =eve
      PRED:and
      ARG1 verb-formula
      ARG2:[binary-formula
            IND: <0>
            PRED: and
            ARG1: <1> =[p-agt-formula
                        PRED: p-agt-neg-reactive-emotive]
            ARG2: <2> = [p-pat-formula
                        PRED: p-pat-source-no-cause]]]]
```

```
[strict-trans-sign
 ORTH:interest
 CAT strict-trans-cat
 SEM: [strict-trans-sem
       IND: <0> =eve
       PRED:and
       ARG1 verb-formula
       ARG2:[binary-formula
             IND: <0>
             PRED: and
             ARG1: <1> =[p-agt-formula
                         PRED: p-agt-cause]
             ARG2: <2> = [p-pat-formula
                         PRED: p-pat-affected-emotive]]]]
```

```
[strict-trans-sign
 ORTH:delight
 CAT strict-trans-cat
 SEM: [strict-trans-sem
       IND: <0> =eve
       PRED:and
       ARG1 verb-formula
       ARG2:[binary-formula
             IND: <0>
             PRED: and
             ARG1: <1> =[p-agt-formula
                         PRED: p-agt-pos-cause]
             ARG2: <2> = [p-pat-formula
                         PRED: p-pat-pos-affected-emotive]]]]
```

```
[strict-trans-sign
 ORTH:scare
 CAT strict-trans-cat
 SEM: [strict-trans-sem
       IND: <0> =eve
       PRED:and
       ARG1 verb-formula
       ARG2:[binary-formula
             IND: <0>
             PRED: and
             ARG1: <1> =[p-agt-formula
                         PRED: p-agt-neg-cause]
             ARG2: <2> = [p-pat-formula
                         PRED: p-pat-neg-affected-emotive]]]]
```

Figure 5: Sample LKB entries for psychological verb subtypes

senses which were not assigned a link to a corresponding LLOCE entry/sense. The basic idea is to derive top-down taxonomies using as parent nodes verb entries for which at least a sense match was found — generalizing LLOCE set identifiers across verb senses where appropriate — and let the daughter nodes of these taxonomies inherit the thesaurus specifications associated with the parent nodes. We expect that this use of verb taxonomy should provide a significant solution for the lack of sense-to-sense correlations due to differences in size.

Acknowledgements

We are indebted to Ted Briscoe, John Carroll and Ann Copestake for helpful comments and technical advice. Many thanks also to Victor Lesk for mounting the LLOCE index and having a first go at correlating LLOCE and LDOCE.

References

Akkerman, E. An Independent Analysis of the LDOCE Grammar Coding System. In Boguraev, B. & T. Briscoe (eds.) *Computational Lexicography for Natural Language Processing.* Longman, London, 1989.

Amsler, R. A. A Taxonomy for English Nouns and Verbs. In *Proceedings of the 19th ACL*, Stanford, pp. 133-138, 1981.

Alshawi, H. Analysing the Dictionary Definitions. In Boguraev, B. & Briscoe, T. (eds.) *Computational Lexicography for Natural Language Processing.* Longman, London, 1989.

Atkins, B. Semantic ID Tags: Corpus Evidence for Dictionary Senses. In *Advances in Lexicology*, Proceedings of the Third Annual Conference of the Centre for the New OED, University of Waterloo, Waterloo, Ontario, 1987.

Atkins, B. Building a Lexicon: the Contribution of Lexicography. Unpublished ms., Oxford University Press, Oxford, 1989.

Atkins, B., J. Kegl & B. Levin. Explicit and Implicit Information in Dictionaries. In *Advances in Lexicology*, Proceedings of the Second Annual Conference of the Centre for the New OED, University of Waterloo, Waterloo, Ontario.

Atkins, B. & B. Levin. Admitting Impediments. In Zernik, U. (ed.) *Lexical Acquisition: Using On-Line Resources to Build a Lexicon.*, forthcoming.

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

Boguraev, B. & J. Pustejovsky. Lexical Ambiguity and the Role of Knowledge Representation in Lexicon Design. In *Proceedings of the COLIN*, Helsinki, Finland, 1990.

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

Briscoe, T., A. Copestake and V. de Paiva (eds.) *Default Inheritance within Unification-Based Approaches to the Lexicon.* Cambridge University Press, forthcoming.

Calzolari, N. & E. Picchi. A Project for Bilingual Lexical Database System. In *Advances in Lexicology*, Proceedings of the Second Annual Conference of the Centre for the New OED, University of Waterloo, Waterloo, Ontario, pp. 79-82, 1986.

Carroll, J. *Lexical Database System: User Manual*, ACQUILEX Deliverable 2.3.3(a), ESPRIT BRA-3030, 1990.

Carroll, J. & C. Grover. 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, 1989.

Copestake, A. (1992) The ACQUILEX LKB: Representation Issues in Semi-Automatic Acquisition of Large Lexicons. This volume.

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

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

Klavans, J. COMPLEX: A Computational Lexicon for Natural Language Systems. In *Proceeding of the COLIN*, Helsinki, Finland, 1988.

Levin, B. *Towards a Lexical Organization of English Verbs.* Ms., Dept. of Linguistics, Northwestern University, 1989.

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

Parsons, T. *Events in the Semantics of English: a Study in Subatomic Semantics.* MIT press, Cambridge, Mass, 1990.

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

Sanfilippo, A. *Grammatical Relations, Thematic Roles and Verb Semantics.* PhD thesis, Centre for Cognitive Science, University of Edinburgh, Scotland, 1990.

Sanfilippo, A. LKB Encoding of Lexical Knowledge from Machine-Readable Dictionaries. In Briscoe, T., A. Copestake and V. de Paiva (eds.) *Default Inheritance within Unification-Based Approaches to the Lexicon.* Cambridge University Press, forthcoming.

Vossen, P. *A Parser-Grammar for the Meaning Descriptions of LDOCE*, Links Project Technical Report 300-169-007, Amsterdam University, 1990.