[*First International Conference on Language Resources & Evaluation*, Granada, Spain, 28-30 May 1998]

Extending a Core Lexicon Using On-line Language Resources with Savoir-Faire

Evelyne Viegas, Arnim Ruelas, Stephen Beale and Sergei Nirenburg
New Mexico State University
Computing Research Laboratory
Las Cruces, NM 88003
USA
viegas,aruelas,sb,sergei@crl.nmsu.edu

Abstract

In this paper, we describe computational semantic lexicons and discuss methodologies to extend them in a cost effective way using on-line language resources with savoir-faire. First, we briefly present the methodology and ecology of computational semantic lexicon acquisition. Second, we report on the way to acquire a large scale lexicon using derivational morpho-semantic rules. Finally, we describe an experiment to extend a semantic-based core lexicon with paradigmatic relations, and predict the syntactic behavior of verbs based on their semantics. These lexicons have been developed within Mikrokosmos, a semantics based machine translation (MT) system.

1 Introduction

In this paper, we present the process of computational semantic lexicon making, from the core lexicon making process to broad coverage processes. In particular, we focus on different methodologies, depending on the availability of on-line language resources, in order to extend existing core lexicons.

First, we briefly introduce the methodology and ecology of computational semantic lexicon acquisition, as reported in (Viegas and Nirenburg, 1996), to build semi-automatically a core lexicon using available on-line resources, graphical tools and the savoir-faire of trained acquirers.

Second, we briefly report on the way to acquire a large-scale high-quality lexicon using derivational morpho-semantic lexical rules (MSLRs), (Viegas *et al.,* 1996). We used MSLRs as a conceptual tool to extend our core Spanish lexicon (7,000 word senses) to 40,000 word senses entirely automatically.

Finally, we present an experiment to extend a core lexicon using off the shelf resources, such as the WordNet database of synsets (Miller, 1990), to propagate the core English lexicon with synonyms. As part of this experiment, we also show how a semantic based approach, such as the one developed in Mikrokosmos, can help predict the syntactic behavior of words. Note that the reverse (predicting semantics from syntax) is not true, as our analysis of Levin's work (1993) below shows. We took advantage of the database of subcategorizations and alternations for English verbs in (Levin, 1993) to encode syntactic information in an entry. However, in order to benefit from Levin's work, we had to "de-emphasize" her notion of class, as we found out that it brought more noise than really helped in assigning a (set of) subcategorization(s) to a verb.

2 The Ecology of Acquisition

In this section, we first briefly describe some trade-offs between lexicon and ontology and outline the main interactions an acquirer is faced with when acquiring a computational semantic lexicon; we, then discuss the corpus-based and mental-based approaches we used in acquisition. Finally, we report on the importance of training people to get better agreement on the assignment of senses to words.

2.1 Lexicon-Ontology Trade-offs

The Mikrokosmos lexicon is mainly based on an ontology of 6000 concepts (Mahesh and Nirenburg, 1995). Knowledge acquisition and meaning representation involve continual trade-offs between the ontology and the lexicon. From a purely ontological perspective, it is desirable to aim for parsimony in the number of concepts. A smaller ontology is not only cheaper to acquire, we can also guarantee better quality of concepts and inter-conceptual relations when the size is small. However, a smaller number of concepts necessitates a greater degree of decomposition in meanings in representing word senses in the lexicon. Not only does this explode the cost of train-

ing and lexical acquisition, but it also creates problems in analysis and generation. In Mikrokosmos, we have strived to achieve an intermediate grain size of meaning representation in both the lexicon and the ontology. Many word senses have direct mappings to concepts in the ontology; many others must be decomposed and mapped indirectly through composition and modification of ontological concepts. One useful rule of thumb is that language-independent elements of meaning are captured in the concepts while language-dependent ones are dealt with in the lexicon.

We keep the number of concepts well below the number of lexical items for a given language, so that the concept INGEST, for instance, can be lexicalized as *eat* or *drink* according to the constraints put on the theme: FOOD and LIQUID, respectively.

```
[key:   "eat",
gram:   [pos:  V,
         subc:  #n=NP, #p= V, #m=NP(opt)],
semRep: [name:INGEST, #a=AGENT:ANIMAL, #t=FOOD]
synSem:  <[gram:  #n,  semRep: #a]
          [gram:  #m,  semRep: #t]>,  ...]

[key:   "drink",
gram:   [pos:  V,
         subc: #n=NP, #p= V, #m=NP(opt)],
semRep: [name:INGEST, #a=AGENT:ANIMAL,#t=LIQUID]
synSem:  <[gram: #n, semRep: #a]
          [gram: #m, semRep: #t]>,  ...]
```

In Mikrokosmos, we have been acquiring computational semantic core lexicons for Spanish, Chinese and English (about 7,000 word senses), and have been working on extending these core lexicons using different methodologies we describe in this paper. The Spanish lexicon has been "virtually" extended to 40,000 word senses, and we expect to automatically extend the English lexicon to about 30,000 word senses, before the application of derivational morphological rules.

2.2 Lexicon Acquirers' Interactions

The lexicon acquirers interact with at least five other people, who perform different functions in the acquisition process:

1) *Acquirer trainer:* the trainer, acquirer-level 1 (master acquirer, top level), must have an excellent knowledge of the field of investigation (such as computational semantics) from a theoretical and NLP perspective, of the approach (such as the semantics-based approach, or the statistics-based approach for the acquisition of hand-tagged sets), and of the framework being used (such as Mikrokosmos, WordNet, etc). Near-fluency, along with native intuitions in the language being acquired, is highly recommended. We distinguish two other levels among the acquirer-analysts: level2 and level3; both must be fluent with near native intuitions in the language being acquired. The trainer must train both acquirer-analyst of level 2 and 3 up to the point where they can perform the tasks they are being given autonomously. Difficult decisions, such as the number of senses, must be filtered out during the pre-acquisition phase by acquirer-level2, eventually supervised by the master-acquirer.

2) *Ontology developer:* the ontology developer interacts daily with the acquirer-level2 during the phase of pre-acquisition, and with both master-acquirer and acquirer-level2 during the phase of testing.

3) *Analyzer builder:* the analyzer builder interacts with the master-acquirer in order to decide on the needs for pre-processing some data or to have them changed in the lexicon. This interaction does not involve errors in the lexicon (such as forgetting a subcategorization) but rather involves such serious modification, as changes in the representational language for lexemes. These changes can be very substantial at the lexicon during the pre-acquisition phase only, after that, when actual acquisition has started. changes must be kept minimal and transparent for acquirer-level3 analysts.

4) *Domain specialist:* the domain specialist interacts essentially with the ontology developer and eventually with the master-acquirer, who has the language competences.

5) *Testing checker:* the testing checker evaluates the output of the analyzer or the generator, and tries to identify the origin of errors (lexicon: wrong POS, missing subcategorization, etc.; ontology: wrong slot, unconstrained filler, etc.). The requests for the lexicon are dealt with by acquirer-level2 analysts.

2.3 Importance of Training

The task of training acquirers for acquiring computational semantic lexicons is of major importance in the process of acquisition. The experiment below is an interesting practical confirmation of this statement.

We asked a native speaker of Spanish, who had not taken part in the lexicon training process, to add some senses to entries in the Spanish lexicon. It was done mainly for the purpose of testing the analyzer, as only 23 out of the 167 words were ambiguous in the Spanish texts which were analyzed.

The list of added senses was reviewed by two computational linguists: a master acquirer, fluent in Spanish and possessing native intuitions in that lan-

guage, and an analyzer builder. A total of 111 new senses were added to 55 open-class words. Among these 55 words, 33 were already ambiguous in the Spanish lexicon.

After a closer look at the Spanish lexicon and at the senses retrieved by the semantic analyzer, and after doing an on-line corpora search, the two computational linguists accepted less than 20 new senses from the 111 suggested.

This "overgeneration" of senses had different origins from less-important to more important, in the task of acquisition:

- the semantic analyzer did not present all the senses from the Spanish lexicon to the native speaker. It only presented the ones that were accepted after the syntactic binding; thus, some senses already present in the Spanish lexicon were added (this constituted a minor category).
- the senses added were "equivalent" to the senses already in the Spanish lexicon, and not recognized by the untrained acquirer, as they were "unspecified" compared to the ones suggested.

- the untrained acquirer hard-coded non-literal meanings of the words.
- the addition of senses was Machine Readable Dictionary (MRD) driven. In other words, the untrained acquirer tried to acquire the list of meanings provided by the Spanish-English Larousse and Collins, adopting the full enumeration approach that we had been discouraging during the training.
- the addition of senses was not corpus-based. In other words, most of the new senses added fell out of the range of the domain under study and the analyzer should not even bother with them after we specify the domain preference in the lexicon.

This exercise showed that training is an essential part of the acquisition of a computational semantic lexicon as among the new senses we validated, only a small part actually pertained to our domain.

MikroKosmos is an approach committed to complete coverage of the material (Nirenburg and Raskin 1996). There are mainly two approaches to complete coverage: mental-based and corpus-based.

2.4 Mental Driven Approach

In the case of a mental-driven approach, the master acquirer produces a set, as exhaustive as possible, of types of semantic representations. The moment a type is created, it is applied to all the lexical items, in which it can be used.

This approach maximizes the use of each type of lexical entry and, by the same token, of the ontological material it is based upon (ontological concepts, facets, etc.) and thus contributes significantly to the parsimony of the ontology, an important concern. It also makes uses of synonymy, antonymy, and other paradigmatic relations among words to generate lists of adjectives that can be acquired using a given lexical entry template. Availability of thesauri and similar on-line resources facilitates this method of acquisition.

In Mikrokosmos, we adopted the mental-driven approach for the acquisition of the English lexicon to be used in generation.

2.5 Corpus Driven Approach

In the case of a corpus driven approach, the master acquirer is interested in capturing primarily the meanings which appear in the corpora of the domain (for instance, corpora on mergers and acquisition, joint ventures, cooking, ...), along with some other meanings beyond the scope of the domain, as the whole methodology in a corpus driven approach is geared at the scalability of the approach, and not at just the discovery of a sub-language vocabulary. There are mainly two paradigms within the corpus driven approach, as mentioned in (Kilgarriff, 1997):

- textual; lexicographers work through the text, token by token; the project SEMCOR (Fellbaum *et al,* 1998) adopted this method; the ratio of agreement in assigning a WordNet sense to a token between the Princeton team and Singapore team was low; this seemed to be linked, in part, to the fact that lexicographers had to read a set of different synsets per different token; little room is left to "systematicity" in the acquisition of senses here, as tokens are different and so are the definitions of synsets.
- lexical; lexicographers are asked to work lexeme by lexeme through the corpus; in other words, an acquirer concentrates on one lexeme at a time and searches for all its senses through the corpus; the Hector project (Atkins, 1993) adopted this method; the agreement in assigning senses was higher, as here lexicographers could reach some speed, as they concentrated on one set of senses for a particular lexeme; also the possibility to recognize recurrences of a same pattern with a particular sense, accelerated sense assignment.

In Mikrokosmos, we adopted a refined lexical approach, so that each acquirer would be assigned a set of lexemes, pertaining to a particular semantic class, produced by a master acquirer. For instance, some acquirers worked on nouns of type Event, whereas others worked on nouns of type Object.

3 Propagation of Lexicons

We have presented so far some methodologies in the acquisition of core lexicons. In this section, we focus on how to "propagate" them monolingually and multilingually, by:

1 Using derivational morphology for extending core lexicons from a monolingual perspective and multilingual perspective for family-related languages;

2 Using off the shelf resources to enhance and extend core lexicons.

3.1 Extending the Lexicon with Derivational Morphology

The methodology consists in submitting each open-class lexeme of the core lexicon to a morpho-semantic generator which produces all its morphological derivations and, based on a detailed set of tested heuristics, attaches to each form an appropriate semantic Lexical Rule (LR) label; for instance, the nominal form *buyer* will be among the ones generated from the verb *buy* and the semantic LR 'agent-of' is attached to it; also the Part-of-Speech and the subcategorization are produced automatically.

In this paper, we deal with the discovery and representation of MSLRs in the process of large-scale semi-automatic computational lexicon acquisition (Viegas *et al.,* 1996) for Spanish. The advantage of MSLRs is twofold: first, they can be considered as a means to minimize the need for costly lexicographic heuristics, to reduce the number of lexicon entry types, and generally to make the acquisition process faster and cheaper; second, they can enhance the results of analysis processing by creating new entries (including the syntactic-semantic linking) for unknown words from the lexicon, found in corpora.

3.1.1 The Morpho-Semantic Lexical Rules

The central idea of our approach - that there are systematic paradigmatic meaning relations between lexical items, such that, given an entry for one such item, other entries can be derived automatically - is certainly not novel and has been treated under different types of lexical rules (see (Onyshkevych, 1998) for a review on LRs). (Viegas *et al,* 1996) addresses the theoretical background on our approach to MSLRs and mentions three different types of lexical rules: 1) inflected forms (passivation - dative alternation); 2) word formation (derivational morphology) and 3) polysemy (sense extension - type coercion). The discussion on when to apply the rules (acquisition time - lexicon load time or run time) is fully discussed in (Viegas *et al.,* 1996) and (Onyshkevych, 1998).

We developed about 100 language independent LRs, which applied to 1056 Spanish verb citation forms, with 1263 senses among them, helped acquire an average of 25 candidate new entries per verb sense, thus producing a total of 31,680 candidate entries (Viegas *et al.,* 1996).

3.1.2 Automatic Generation of Lexicon Entries

Figure 1 below illustrates the overall process of generating new entries from a citation form by applying MSLRs.[1]

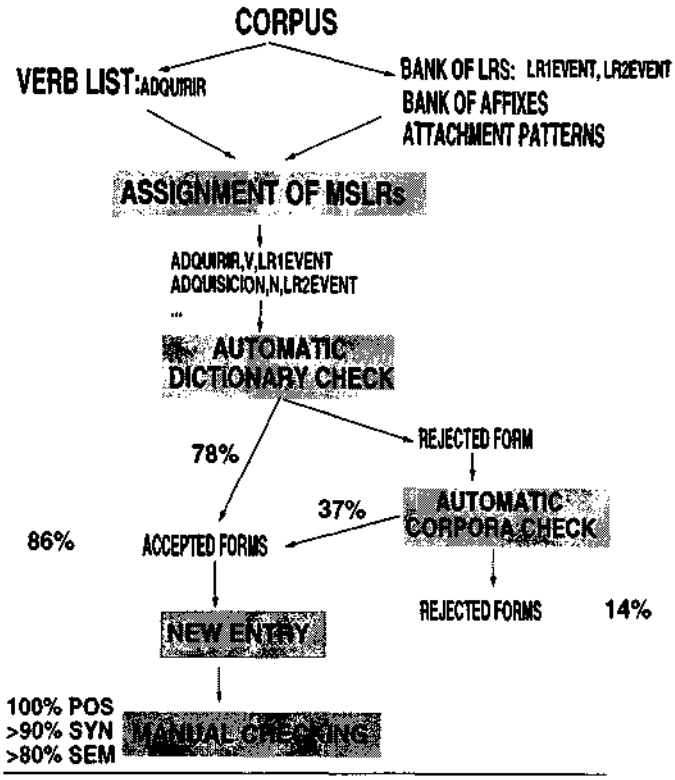

Figure 1: Automatic Generation of New Entries.

Generation of new entries usually starts with verbs. Each verb found in the corpora is submitted to the morpho-semantic generator, which pro-

[1] See (Viegas and Beale, 1996) for the details on the conceptual and technological tools used to check the quality of the lexicon.

duces all of its morphological derivations and, based on a detailed set of tested heuristics, attaches to each form an appropriate semantic LR label. For instance, the nominal form *comprador* (buyer) will be among the ones generated from the verb *comprar* (buy) and the semantic LR "agent-of" is attached to it.

The form list generated by the morpho-semantic generator is checked against MRDs and dictionaries and the forms found in them (accepted forms) are submitted to the acquisition process. However, forms not found in the dictionaries are not discarded outright because the MRDs cannot be assumed to be complete and some of these "rejected" forms can, in fact, be found in corpora or in the input text of an application system. This mechanism works because we rely on linguistic clues and therefore our system does not grossly overgenerate candidates.

The Lexical Rule Processor is an engine which produces a new entry from an existing one after applying a lexical rule, such as a new entry *compra* produced from the verb entry *comprar* after applying the LR2event rule.

The acquirer must check the definition and enter an example, but the rest of the information is simply retained. The LexRule zone in the entry specifies the morpho-semantic rule which was applied to produce this new entry and the verb it has been applied to.

3.1.3 Lexical Rules and Cost-Effectiveness

It is clear by now that LRs are most useful in large-scale acquisition. In the process of Spanish acquisition, 20% of all entries were created from scratch by master-acquirer and 80% were generated by LRs and checked by research associates. It should be made equally clear, however, that the use of LRs is not cost-free. Besides the effort of discovering and implementing them, there is also the significant time and effort expenditure on the procedure of semi-automatic checking of the results of the application of LRs to the basic entries, such as those for the verbs.

The shifts and modulations studied in the literature in connection with the LRs and generative lexicon have also been shown to be not problem-free: sometimes the generation processes are blocked-or preempted-for a variety of lexical, semantic and other reasons. In fact, the study of blocking processes, their view as systemic rather than just a bunch of exceptions, is by itself an interesting enterprise (see (Briscoe *et al*., 1995)).

Obviously, similar problems occur in real-life large-scale lexical rules as well. Even the most seemingly regular processes do not typically go through in 100% of all cases. This makes the LR-affected entries not generatable fully automatically and this is why each application of an LR to a qualifying phenomenon must be checked manually in the process of acquisition. However, the whole methodology can be applied to family-related languages at a lower cost, once the LRs have been discovered.

3.2 Adapting Off the Shelf Resources

In this section, we present an experiment to extend a core lexicon using off the shelf resources: 1) we used WordNet (Miller, 1990) to propagate the core English lexicon with synonyms; 2) we used Levin's database of subcategorizations and alternations for English verbs (Levin, 1993) to encode syntactic information in the verb entries.

We will show how a semantic based approach, such as the one developed in Mikrokosmos, can help predict the syntactic behavior of words. However, in order to benefit from Levin's work, we had to "de-emphasize" her notion of class, as we found out that it brought more noise than really helped in assigning a (set of) subcategorization(s) to a verb.

Using these resources, we can increase our verb lexicon automatically (involving some manual checking though, as described later) by a factor of 5. In this paper, we report on experiments we made on half of our core lexicon, showing promising results. However, in order to get to these results we had to work on filters and thresholds so that manual checking (with the help of GUIs) be kept to a minimum.

One of the major problems in using Levin's database was to be able to filter out homonyms, as classes in Levin's database are defined on the basis of the same subcategorization pattern and not on a semantic basis, as we detail below.[2]

The advantage of our approach is that it is semantic-based; this allows us to organize verbs into true (frame-based) semantic classes, to which are associated sets of subcategorizations. Therefore, we can predict that all verbs belonging to a particular semantic class will have the same syntactic behaviors. For instance, if one considers the semantic class of aspectual verbs which selects a theme of type Event, e.g. *begin, continue, finish,* then one can minimally associate to any verb belonging to this semantic class the following subcategorizations: (a) NP-V-NP in *John began his homework;* (b) NP-V-XCOMP *John began to work/ working.* Note that

[2]Many other experiments have yielded to the same observation (or close enough), as described in (Fellbaum, 1998), (Dorr *et al.,* 1997), (Dang *et al,* 1997).

the reverse is not necessarily true: verbs which accept (a) and (b) are not necessarily aspectuals, e.g. *forget* in *I forgot the key* or *I forgot to bring the key.*

Predicting adequately the subcategorizations for a semantic class depends on its grain size: the finer-grained, the better the prediction will be. However, in NLP applications, where one is constrained by time, only the semantics necessary for an application (such as for instance MT) is acquired, which means that in many cases the semantics is left at a coarser grained-size than the one required to predict the subcategorizations. In practice, we overgenerate some subcategorizations and need therefore to have them checked by humans.

We tested our methodology on 2892 English verbs mapped to 515 concepts; in other words we already had "near-synonyms" sharing the same concept or semantics in the input files to the expansion program. The aim of the experiment was to acquire, with minimal effort, synonyms for the lexemes in the core lexicon along with the subcategorizations for all English verbs. The basic process is shown in Figure 2.

The methodology consists in first 'pruning' the subcategorizations in Levin's database by recognizing homonymy or polysemy; for instance, the English verb *separate* belongs to two classes: 'separate' and 'dry'. The subcategorizations associated with different classes is not necessarily identical. For analysis, it is hardly a problem, as one does not expect an ungrammatical input, the problem is different for generation, where we do not want to produce ungrammatical structures for a verb.

Lexigen, represents our core lexicon for English. On the left hand-side of the diagram we show two main processes: 1) the filtering of Levin's database, where we got rid of subcategorizations not validated by humans and recognized homonymy as we describe further; and, 2) "translated" each subcategorization into our syntactic framework (LFG-like structures). In the centre of the diagram we present five main processes leading to the expansion of our English core lexicon:

- Retrieve the lexemes from Lexigen, consisted in getting for each concept in Lexigen its associated lexemes; (DIVIDE: separate, split...)

- Provide WordNet synonyms, consisted in retrieving from WordNet all the synsets which shared the same semantics as the concept from the ontology; this involved some filtering we describe further; (separate, split, divide, ...)

- Provide Levin's subcategorizations, involved associating a set of subcategorizations to verbs based on their semantics from the ontology; (separate →

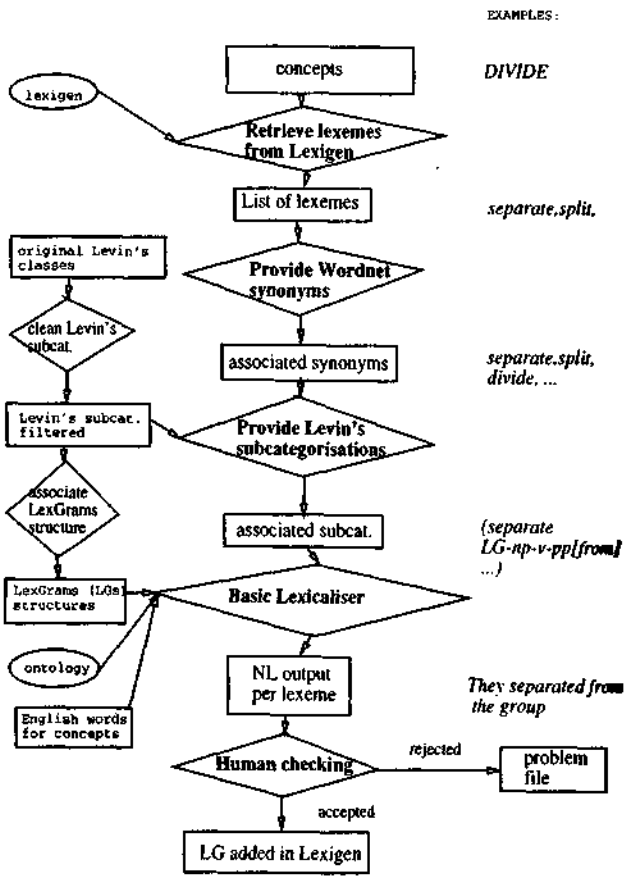

Figure 2: Filtering Off the Shelf Resources with Savoir-Faire

LG-np-v-pp [from])

- Basic Lexicaliser, generated sample sentences for each LG frame; (LG-np-v-pp [from] → *they separated from the group)*

- Human checking, consisted in validating or countervalidating the output of the basic lexicaliser in order to control the overgeneration of subcategorizations, result of coarse-grained semantic classes.

On the right side of the diagram we provide examples.

Briefly, to filter homonyms, we used a fuzzy string matching between the definition 'concept' attached to a particular verb and the Levin's prototype attached to that same verb. For instance, in the case of our Mikrokosmos 'concept' DIVIDE, to which is mapped the verb "separate" in our core lexicon, are attached two Levin's classes: (23,1 separate) and (45,4 dry). The Mikrokosmos concept definition for DIVIDE is: 'to separate (something) into parts.' In this case, only the prototype (separate) is kept for the English verb "separate" mapped to concept DI-

VIDE. In order to get more matches (as our aim is to expand the lexicon) we used from Levin all the verbs belonging to the same class as the prototype to do the fuzzy string matching against the ontology. The fuzzy string match was done against the Mikrokosmos concept definition, two levels of ISA and SUBCLASSES concepts from the ontology. If the Levin class had more than 50 verbs, we worked out a threshold of four matches in the Levin class, in order to filter out homonyms.

The associated synsets in WordNet are extracted with fuzzy string matches between the value of the concept, the ISA concept and the direct hypernyms and hyponyms for the lexeme in WordNet synsets. For instance, for the English verb *separate* we found in WordNet:

Sense 1: *separate, divide*

Sense 2: *separate, disunite, force apart, divide, part, take apart, pull apart → move, displace, make move*

Sense 3: *distinguish, separate, differentiate, secern, secernate, severalize, tell, tell apart → identify, place, recognize as being*

Sense *4: divide, split, split up, separate, dissever, carve up → change integrity*

Sense 5: *separate, part, split, move apart → move, change position*

Sense 6: *separate, divide into components → change integrity*

Sense 7: *classify, class, sort, assort, sort out, seperate → categorize, place into a category*

Sense 8: *separate, part, split up, split, break up*

Sense 9: *separate, divide*

Sense 10: *break, separate, split up, fall apart, come apart → change integrity*

Sense 11: *discriminate against, separate, single out → distinguish, separate, differentiate, secern, secernate, severalize, tell, tell apart*

Sense 12: *separate, divide, come apart, part → change*

Sense 13: *branch, ramify, fork, separate → diverge, move apart, draw apart*

Here, our algorithm will keep all the senses but 3, 7 and 13, as no match could be found with these synsets. We would therefore assign to all senses kept the prototype and class(es) associated with *separate,* as seen below in the examples, output of the program before manual checking.

The output of the process in Figure 2 is a file where each lexeme, associated to a particular concept in the ontology is assigned a set of subcategorizations which have to be ultimately checked by a human. In order to expand the lexicon as much as possible while still getting a reasonable output (so that overgeneration and/or undergeneration be kept minimal) we performed various tests which results are provided in Figure 3.

after:	number of entries	no-prototype in Levin's DB	prototype no-class matched
prototype assignment	3318	1258 37.91%	1153 34.74%
prototype enriching	5703	509 8.92%	1151 20.18%
WordNet enriching	13317	509 3.82%	3250 24.4%

Figure 3: Test Outputs.

We kept the last test, as it provided us with enough entries while keeping over-/under-generation reasonable. We provide below an example for the concept DIVIDE in the ontology, along with the synonyms from WordNet belonging to the same ontological class; the common subcategorizations are provided at the top.

```
(DIVIDE
 ([np,v,np,pp([from,around])]; Levin's 10.1 REMOVE
  [np,v,np,pp (from)]
  [np,v,np,pp ([from, under])]
  [np,v,adv(easily),pp(from)] ; Levin's 23.1 SEPARATE
  [np(and),v,adv(easily)]
  [np,v,np(and)]
  [np,v,np,pp(from)]
  [np,v,pp(from)]
  [np(and),v]
  [np,v,np(and).adv(apart)]
 (break)
 (break-up)
 (carve-up                ;  NO-CLASS DRY)
 (come-apart              ;  NO-CLASS DRY)
 (discriminate-against)
 (dissever                ;  NO-CLASS DRY)
 (divide                  ;  NO-CLASS DRY)
 (divide-into-components)
 (fall-apart)
 (move-apart)
 (part                      NO-CLASS DRY)
 (partition                 NO-CLASS UNSCREW)
 (separate                  NO-CLASS DRY)
 (set-apart)
 (single-out)
 (split                  ;  NO-CLASS DRY)
 (split-up               ;  NO-CLASS DRY)   ))
```

We compared the output file of this program to a test file, manually checked, results are as follows: 50% perfect match; 25% overgeneration; and, 25% undergeneration. In fact, if we compare the file checked manually against a corpus, the "extra" subcategorizations appear less frequently than the exact matches ones. This work is still in progress and further testing is required. However, we can already

foresee that starting the manual checking earlier on in the process, at the level of Levin's database should considerably improve the number of exact matches, because of the incompleteness of her database.

4. Perspectives

In this paper, we have shown 1) that a mental-driven approach is best suited for acquiring lexicons to be used in generation, whereas a lexical corpus-based approach better fits the acquisition of lexicons to be used to analyze on-line corpora; 2) ways of extending a core lexicon using a) language independent LRs which assigned to the affixes of a particular language help propagate entries containing syntactic and semantic information, and b) off the shelf resources to acquire subcategorizations and synonyms. Moreover, the methodology and resources can be easily adapted to family-related languages.

The methodologies described here are part of what is needed to build a multi-purpose multilingual lexical knowledge base while keeping the costs of acquisition as low as possible. Our lexicons are being coded with the formalism of Typed Feature Structures. They have been used for analysis, giving results of 97% accuracy in the task of Word Sense Disambiguation (Mahesh *et al.,* 1997). The English lexicon is now being tested in generation.

Acknowledgments

This work has been supported in part by DoD under contract number MDA-904-92-C-5189. We would like to thank the Mikrokosmos team and Bonnie Dorr and Doug Jones for letting us use their on-line database of English verbs subcategorizations.

References

Atkins, S. (1993) Tools for computer-aided lexicography: the Hector Project. In *Papers in Computational Lexicography,* COMPLEX'93, Budapest.

Briscoe, T., Copestake, A. & A. Lascarides (1995) Blocking. In Saint-Dizier & Viegas (eds.) *Computational Lexical Semantics.* Cambridge: CUP.

Dang, H., Rosenzweig, J. & M. Palmer (1997) Associating Semantic Components with Intersective Levin Classes. In Proceedings of the *AMTA/SIG-IL-1rst Workshop on Interlinguas,* MCCS-97-314.

Dorr, B., Olsen, M. & D. Clark (1997) Using WordNet to Posit Hierarchical Structure in Levin's Verb Classes. In Proceedings of the *AMTA/SIG-IL-1rst Workshop on Interlinguas,* MCCS-97-314.

Fellbaum, C., Grabowski, J., Landes, S. & A. Baumann (1998) Matching words to senses in WordNet: Naive *vs.* expert differentiation of senses. In C. Fellbaum (ed.) *WordNet: An Electronic Lexical Database and Some of its Applications.* MIT Press: Cambridge.

Fellbaum, C. (1998) Semantics via Conceptual and Lexical Relations. In Viegas, E. (ed.) *Breadth and Depth of Semantic Lexicons,* Kluwer Academic Press.

Mahesh, K., Nirenburg, S., Beale, S., Viegas, E., Raskin, V. & B. Onyshkevych (1997) Word sense disambiguation: why statistics when we have these numbers? In Proceedings of the *7th International Conference on TMI in MT,* Santa Fe, NM.

Kilgarriff, A. (1997) Sample the Lexicon. Technical report, ITRI-97-01.

Levin, B. (1993) Levin, Beth 1992. English Verb Classes and Alternations. A Preliminary Investigation. Chicago: University of Chicago Press.

Mahesh, K. & S. Nirenburg (1995) A situated ontology for practical NLP. In *Proceedings of A Workshop on Basic Ontological Issues in Knowledge Sharing at the IJCAI'95* Montreal.

Miller, G.A. (1990) Nouns in WordNet: A Lexical Inheritance System. In *International Journal of Lexicography 3.4,* 245-264.

Nirenburg, S. & V. Raskin (1996) Ten Choices for Lexical Semantics. Technical Report MCCS-96-304, Computing Research Laboratory, New Mexico State University.

Onyshkevych, B. (1998) Categorization of Types and Application of Lexical Rules. In Viegas, E. (ed.) *Breadth and Depth of Semantic Lexicons:* Kluwer Academic Press.

Viegas, E. & S. Beale (1996) Multilinguality and Reversibility in Computational Semantic Lexicons. In Proceedings of the 8th *International Natural Language Generation Workshop,* Sussex, England, 49-52.

Viegas, E. & S. Nirenburg (1996) The Ecology of Lexical Acquisition: Computational Lexicon Making Process. In Proceedings of Euralex96, Göteberg University, Sweden.

Viegas, E., Onyshkevych, B., Raskin, V. & S. Nirenburg (1996) *Submit* to *Submitted* via *Submission:* on Lexical Rules in Large-scale Lexicon Acquisition. In Proceedings of the *34th Association for Computational Linguists,* CA, 32-39.