Exploring large-scale Acquisition of Multilingual Semantic Models for Predicates *

Jordi Atserias, Mauro Castillo, Francis Real, Horacio Rodríquez
TALP Research Center
Universitat Politècnica de Catalunya
{batalla,castillo,fjreal,horacio}@lsi.upc.es

German Rigau
IXA Group
Euskalerriko Unibersitatea
Donostia.
{rigau}@si.ehu.es

Resumen: Investigamos la posibilidad de obtener patrones semánticos a gran escala para cualquier lengua usando solamente análisis superficial y generalizaciones semánticas básicas. Siendo este un experimento exploratorio sólo hemos realizado una evaluación cualitativa. Hemos comparado varios patrones semánticos de traducción de verbos equivalentes en distintas lenguas y dominios.
Palabras clave: Adquisición, Modelos Semánticos, wordnets, Multilingualidad

Abstract: We investigate the feasibility to obtain large-scale semantic patterns for any language based only on shallow parsing and some basic semantic generalizations. Being this a exploratory experiment we performed only a qualitative evaluation. We compared several semantic patterns coming from translation equivalent verbs selected from different languages and domains.
Keywords: knowledge Acquisition, Semantic Patterns, wordnets, Multilinguality

1 Introduction

Recently, obtaining large, explicit lexicons rich enough for NLP has proved difficult. Methods for automatic lexical acquisition have been developed for many topics and include collocations (Justeson and Katz, 1995), word senses (Lin and Pantel, 2002), prepositional phrase attachment ambiguity (Hindle and Rooth, 1993), selectional preferences (Li and Abe, 1998; McCarthy, 2001; Agirre and Martinez, 2002), subcategorization frames (SCFs) (Brent, 1993; Manning, 1993; Briscoe and Carroll, 1997; Korhonen, 2002) and diathesis alternations (Lapata, 2001; Walde, 2000; McCarthy, 2001). Many of these methods are still under development and need further research before they can successfully applied to large scale acquisition.

Being a multidimensional problem, predicate knowledge is one of the most complex types of information to acquire. Predicates (verbs and their corresponding nominalizations) are essential for the development of robust and accurate parsing technology capable of recovering predicate-argument relations and logical forms. Without it, resolving most structural ambiguities of sentences is difficult, and understanding impossible.

Moreover, predicate-argument knowledge have been shown to vary across corpus type (written vs. spoken), corpus genre (e.g. financial news vs. balanced text), and discourse type (single sentences vs. connected discourse) (Roland et al., 2000). (Roland and Jurafsky, 2002) have showed that much of this variation is caused by the effects of different corpus genres on verb sense and the effect of verb sense on predicate-argument associations.

Full account of predicate information requires specifying the number and type of arguments, predicate sense under consideration, semantic representation of the particular predicate-argument structure, mapping between the syntactic and semantic levels of representation, semantic selectional restrictions/preferences on participants, control of the omitted participants and possible diathesis alternations. Unfortunately, all these kinds of knowledge are interdependent.

Basically, the acquisition of predicate-argument associations has been merely syntax driven. Following a bottom-up approach, from syntax to semantics, if we identify specific associations between SCFs and predicates, we can gather information from corpus data about head lemmas which occur in argument slots of SCFs and use this information as input to selectional preference acquisition (McCarthy, 2001; Walde, 2000). Se-

* This research has been partially funded by the European Comission (MEANING IST-2001-34460), Generalitat de Catalunya (2002FI 00648) and the Universidad Tecnológica Metropolitana - Chile.

lectional preferences are an important part of predicate information, since they can be used to aid anaphora resolution (Ge, Hale, and Charniak, 1998), WSD (Resnik, 1997; McCarthy, Carroll, and Preiss, 2001) and automatic identification of diathesis alternations from corpus data(Walde, 2000; Stevenson and Merlo, 1998; McCarthy, 2001).

However, (Korhonen, 2002) showed that in terms of SCF distributions, individual verbs correlate more closely with syntactically similar verbs and clearly more closely with semantically similar verbs, than with all verbs in general. Moreover, her results show that verb semantic generalisations can successfully be used to guide and structure the acquisition of SCFs from corpus data.

Thus, it is possible to devise alternative acquisition schemes going top-down from semantics to syntax. If we identify specific associations between participants and predicates (selectional preferences), we can also gather information from corpus data about their particular syntactic behaviour in relation to a predicate, helping the acquisition of SCFs, diathesis alternations, etc. However, this new approach requires to work directly at a sense level, having predicates and associations to participants semantically disambiguated.

Furthermore, in a multilingual semantic scenario, it seems possible to devise ways to acquire from a particular language and using a bottom-up approach some predicate-argument knowledge, and then, following a top-down fashion, to acquire or validate some knowledge in other language.

Two different and complementary dimensions can help to minimise the WSD problem: multilingualism and domains. Although, working in parallel with comparable corpora in several languages will increase the complexity of the process, we believe that language translation discrepancies among word forms can help the selection of the correct word senses (Habash and Dorr, 2002). Moreover, further reduction of the search space among sense candidates can be obtained by processing domain corpora (Gale, Church, and Yarowsky, 1992).

This paper presents the first steps towards testing the validity of this new approach for the acquisition of predicate knowledge (SCFs, Selectional Restrictions, diathesis alternations, etc). The work here presented explores some basic issues in the acquisition of semantic models. First, how the current technology and the knowledge available can help large-scale acquisition tasks, mainly subcategorization frames (SCFs) and selectional restrictions or preferences (SPs) for Spanish. Second, the impact in the acquisition process when using several languages at the same time and third, when using domain corpus instead of a general corpus.

After this introduction, section 2 presents the resources used in this exploration. Section 3 describes the methodology used to acquire large–scale Semantic Models for Spanish predicates. Section 4 provides some qualitative views with about the domain and multilingual exploration and finally, in Section 5 we conclude with some prospects for future work.

2 Experimental Setting

Summarising, this paper presents new ways for restricting the search space when performing acquisition tasks, in order to obtain more accurate knowledge for some languages and balance the coverage of such knowledge across languages. Thus, this experiment can be also seen as a common framework to study productive paths to exploit appropriately:

- available semantic knowledge (wordnets, Semantic Files, MultiWordNet Domains(Magnini and Cavaglià, 2000), EuroWordNet Top Ontology (Vossen, 1998), etc.)
- cross language discrepancies/agreements through the EuroWordNet Interlingual Index
- available comparable domain corpora
- large-scale selectional preferences already acquired from SemCor (Agirre and Martinez, 2002) and British National Corpus (McCarthy, 2001)

Next, we will provide a short description of each of these resources.

2.1 Spanish and English wordnets

Table 1 compares the amounts of synsets of the wordnets used in this experiment with respect different Part–of–Speech categories: English WordNet1.6 and the current version of the Spanish wordnet[1]. At a synset level,

[1] `http://nipadio.lsi.upc.es/wei.html`

overlapping between both wordnets is quite high and homogeneous across POS categories, ranging from 45% for nouns to 62% for verbs and adjectives.

	en16	spwn	Overlapping
Nouns	66,025	31,241	29,502
Verbs	12,127	7,563	7,464
Adjectives	17,915	11,135	11,087
Total	96.067	49.934	48.053

Tabla 1: Spanish-English WN overlapping

2.2 MultiWordNet Domains

In this experiment we use MultiWordNet Domains (Magnini and Cavaglià, 2000) which were partially derived from the Dewey Decimal Classification [2]. WordNet Domains is a hierarchy of 165 Domain Labels associated to WordNet 1.6 synsets.

Information brought by Domain Labels is complementary to what is already in WordNet. First of all a Domain Labels may include synsets of different syntactic categories: for instance MEDICINE groups together senses from nouns, such as *doctor* and *hospital*, and from verbs such as *to operate*. Second, a Domain Label may also contain senses from different WordNet subhierarchies For example, the SPORT contains senses such as athlete, deriving from life form, game equipment, from physical object, sport from act, and playing field, from location.

2.3 Selectional Preferences acquired from SemCor

This large set of selectional preferences (SPs) were obtained from gramatical relations extracted from Semcor ((Agirre and Martinez, 2001) and (Agirre and Martinez, 2002)). Basically, these SPs were collected parsing SemCor with the Minipar parser (Lin, 1998). In that way, it was possible to obtain triple dependencies, of the form [noun-synset, relation, verb-synset], for all annotated sense examples in Semcor. Table 2 presents the amounts of the *object* and *subject* relations.

The acquisition method provided 69,840 weighted subject preferences between 2,490 different verbal synsets (an average of 20.40 relations per verbal synset) and 5,398 nominal synsets (an average of 10.02 relations per nominal synset).

[2] `http://www.oclc.org/dewey`

Regarding object preferences, this process acquired 110,102 weighted semantic relations between 3,423 different verbal synsets (an average of 32.17 relations per verbal synset) and 6,964 nominal synsets (an average of 15,81 relations per nominal synset).

2.4 Selectional Preferences acquired from BNC

In this case, the selectional preferences were obtained by means of probability distributions over the WordNet 1.6 noun hyponym hierarchy using the ninety million words of the written portion of the British National Corpus (BNC) (McCarthy, 2001). In this case, the SPs were obtained also automatically from parsed text using the RASP parsing toolkit (Carroll, Briscoe, and Sanfilippo, 1998).

The preference models are modifications of the Tree Cut Models (TCMs) originally proposed by Li and Abe (Li and Abe, 1998) These were acquired for grammatical relations (subject, direct object and adjective-noun) involving nouns and grammatically related adjectives or verbs.

In table 2 we summarize the number of weighted subject, object preferences acquired from BNC.

	#verbal synsets	#nominal synsets	#relations
Semcor SUBJ	2,490	5,398	69,840
Semcor DOBJ	3,423	6,964	110,102
BNC SUBJ	6,151	2,588	95,065
BNC DOBJ	6,125	4,185	115,542

Tabla 2: Selectional Preferences

In this case, two different kind of relations were acquired from BNC. We can consider as *different* relations those captured as class-based preferences (including hyponyms) and synset-based preferences (excluding descendants, being considered as leaf nodes). While class-based preferences can be inherited through the noun hierachy, synset-based preferences only holds for those synsets selected (these relations can not be inherited).

2.5 Domain Corpora

We use EFE news agency articles for January, February and March 2000 from FINANCE and SPORT domains. Table 3 provides some general figures of this corpus. These articles are also categorised using IPTC codes[3].

[3] see http://www.iptc.org

Using this corpus, it is easy to select only those articles belonging to only one major IPTC code such as: FINANCE or SPORT. We expect different verb behaviours with respect FINANCE, SPORT and the general corpus.

Total of News articles	291,997
Total of Sentences	2,811,782
Total of Words	95,341,184
Average of sentences per article	9.63
Average of words per article	326.51
Average of words per sentence	33.99
Sports News articles	70,778
Finances News articles	45,099

Tabla 3: Figures for Spanish EFE corpus

2.6 Word Selection

In order to perform multilingual and domain comparisons we manually select 7 verbs (and their corresponding English translations) from the 100 most relevant verbs in Spanish and English and having good coverage in both domains (if possible).

As we can easily notice in Table 4, verb distributions are biased to SPORT domain. Some of them (i.e., *empatar* and *entrenar* mainly occur only on SPORT domain).

The average of sentence length shown in Table 5 suggests that it could be difficult to obtain a correct full parser for detecting the *object* or *subject* functions).

Spanish verb distribution			
	Sport	**Finance**	**Other**
ganar	24047	2055	7804
perder	8463	1820	7670
subir	1490	3754	2620
bajar	1168	3336	2377
empatar	2787	0	83
jugar	25534	169	1891
entrenar	4152	15	392

Tabla 4: Figures for the Spanish verbs

Sentence length		
	Sports	**Finances**
subir	45.53	36.34
bajar	44.56	37.49
ganar	43.02	37.93
perder	43.11	38.78
jugar	42.67	41.41
empatar	41.26	0
entrenar	42.34	41.57

Tabla 5: Figures for the Spanish verbs

3 Monolingual Spanish Acquisition

3.1 Spanish SCFs Acquisition

Although other approaches are possible (for instance, starting from raw data (Brent, 1993) or parsed data (McCarthy, 2001; Korhonen, 2002)) in this experiment we analysed all this sentences using the Natural Language Tools for Spanish, performing POS tagging, Name Entity Recognition and Classification (NERC) and chunking.

Basically, chunks (Abney, 1991) are non-recursive cores of major phrases, e.g. NPs, PPs, verb groups and so forth. Essentially, chunking allows factoring sentence structure into pieces allowing posterior generalisations on slot heads and prepositions.

The purpose of this task is to obtain basic chunking of main phrases, process passive sentences and to identify prepositions, head nouns, etc. The output for each sentence should be a simple list of words and chunks (syntactic patterns) as:

Ex: [NP] ganar [NP]

Each chunk has its head word, usually the last verb form (for verb phrases) and the first noun form (for noun phrases). To obtain possible direct objects and subjects from the sequence of chunks we used a *naive* heuristic: the first noun phrase to the left of the verb supposed to be the subject and the first noun phrase to the right of the verb is supposed to be the object. Due to the complexity of the sentences a list of barriers have been defined. These barriers usually act as discourse markers changing the focus of the sentence. These barriers prevent the algorithm to pick up chunks beyond them.

Once the left and right noun phrases (**NP**) has been selected using our naive heuristic, we can consider the minimal subcategorization frame of the verb simply as the chunk sequence between those (**NP**). For this experiment we have only considered noun phrases (**NP**) and prepositional phrases (**PP**) as elements for the subcategorization frame.

For instance, table 6 shows the most frequent SCFs for the Spanish verb *ganar* in FINANCE domain. The star (*) marks the chunk where a verbal form of *ganar* is detected.

Being this a preliminary experiment, we considered only the highest frequency set of syntactic patterns per verb and domain.

432	[NP][*VP][NP]
97	[NP][Fc][*VP][NP]
79	[NP][*VP][Fc][NP]
66	[NP][Fc][relative][*VP][NP]
37	[NP][Fc][*PP(tras)][NP]
27	[NP][relative][*VP][NP]
26	[NP][PP(de)][Fc][*VP][NP]
26	[NP][PP(de)][*VP][NP]
23	[NP][*PP(tras)][NP]
22	[NP][*VP]

Tabla 6: Most frequent SCF for *ganar* FINANCE

3.2 Semantic generalization verb–slot

Then, we perform a very basic generalisation on a particular verb-domain-slot in two steps:

1. Collect for a syntactic position all possible fillers.

   Ex: ganar / perder in FINANCE domain corpus the first NP to the right: dinero, dólar, euro, ...

2. Collect their possible synsets and associated SemanticFile+WordNetDomain sorted by frequency.

   Ex: NOUN.POSSESSION+MONEY

For simplicity, we performed initially this task only for those verb-slots acting possibly as subjects and objects.

Table 7 shows the most frequent words detected as Subject and Object for *ganar* in the Financial Domain. Even thought there are obvious errors (PERSON and ORGANIZATION are not suitable as direct objects) it seems quite reasonable to think that a frequency-cut method will minimise the effect of parser errors.

Object	Count	Subject	Count
PERCENTAGE	413	ORGANIZATION	280
punto	302	PERSON	149
AMOUNT	203	NOSUBJECT	131
NOOBJECT	100	PERCENTAGE	121
elección	61	empresa	74
terreno	34	acción	64
centavo	33	título	44
ORGANIZATION	31	compañía	36
PERSON	29	AMOUNT	34

Tabla 7: Most frequent Spanish Subject and Object heads for *ganar*

3.3 Acquisition of Semantic Patterns

Consulting again the corpus for instance sentences (slot heads) and filtering out automatically impossible combinations, we can perform basic and coarse-grained generalization of semantic patterns using at the same time several syntactic positions (e.g. first NP to the left and right)[4]:

Ex: for "La empresa ganó mucho dinero" (*The company gained a lot of money*) we obtain: GROUP+ECONOMY ganar POSSESSION+MONEY

To show the potentiality of this approach, for this experiment we chose the combination of Wordnet Semantic Field and MultiWordNet Domains as the semantic representation for each synset. We also map the Named Entities types (PERSON, ORGANIZATION, AMOUNT, PERCENTAGE, DATE, etc.) to the same semantic representation (Domain and Wordnet Semantic File).

In table 8, NONE stands for words that doesn't appear in the Spanish WordNet. In this table also appears two new syntactically tags: Fc which stands for punctuation marks (such as quotes, comas, etc.) and NO_SUBJECT and [–] which represent sentences where the subject is not detected.

Mostly due to errors, omissions and inconsistencies, the most frequent semantic pattern has no semantics associated ([NONE] ganar [NONE]). However, using this simply approach we are able to obtain more useful pattens such as:

```
[PERSON] ganar [PERCENTAGE]          : gain
[PERSON] ganar [ACT+POLITICS]        : win
[PERSON] ganar [COGNITION+FACTOTUM] : increase
[ORGANIZATION] ganar [PERCENTAGE]    : gain
[ORGANIZATION] ganar [AMOUNT]        : gain
```

4 About Domains and Multilinguality

In the previous section, we shown the feasibility to obtain large-scale semantic patterns for Spanish based only on shallow parsing and some basic semantic generalizations.

Having all this semantic knowledge we are also able to compare results and data across languages and domains. As the semantic patterns obtained are difficult to evaluate directly (no gold standard seems available for

[4] Obviously, this and the previous step can be performed altogether.

Count	Subject		Object		Subcat. Frame
	Lex. File	Domain	Lex. File	Domain	
25	NONE	NONE	NONE	NONE	[NP] (ganar) [NP]
10	NO_SUBJECT	NO_SUBJECT	NONE	NONE	[–] (ganar) [NP]
7	PERSON	PERSON	PERCENTAGE	PERCENTAGE	[NP] (ganar) [NP]
5	NONE	NONE	PERCENTAGE	PERCENTAGE	[NP] (ganar) [NP]
4	PERSON	PERSON	act	politics	[NP][Fc] (ganar) [NP]
4	PERSON	PERSON	cognition	factotum	[NP][Fc] (ganar) [NP]
4	ORGANIZATION	ORGANIZATION	PERCENTAGE	PERCENTAGE	[NP] (ganar) [NP]
4	ORGANIZATION	ORGANIZATION	AMOUNT	AMOUNT	[NP] (ganar) [NP]
4	NO_SUBJECT	NO_SUBJECT	PERCENTAGE	PERCENTAGE	[–] (ganar) [NP]
4	NO_SUBJECT	NO_SUBJECT	AMOUNT	AMOUNT	[–] (ganar) [NP]

Tabla 8: Most frequent Semantic for *ganar* in FINANCE domain

Spanish), we decided to perform two indirect qualitative evaluations. While section 4.1 presents some interesting examples and results when analysing comparable corpora (the English version of EFE), section 4.2 focusses on the use of specific domain corpora rather than general corpora.

4.1 Crosslingual Comparison

Table 9 presents the most frequent FINANCE subjects for the Spanish verb *ganar* and their corresponding English subjects for those English equivalent verbs to *ganar*. Both lists are quite different. Mainly because some basic problems concerning the different capabilities of the NLP tools used for English and Spanish: the English parser is not performing NERC. However, we are also detecting several equivalent translations (for instance, company, enterprise or market).

English		Spanish	
index	279	ORGANIZATION	280
transaction	114	PERSON	149
it	107	NOSUBJECT	131
which	64	PERCENTAGE	121
bond	54	empresa (*enterprise*)	74
they	43	acción	64
agreement	38	título	44
company	31	compañía (*company*)	36
government	30	AMOUNT	34

Tabla 9: Most frequent FINANCE subjects for *ganar* and its English translations

4.2 Domain and General Corpus Acquisition

This section studies the use of domain specific corpus and general corpus for acquisition. In order to carry out this comparison we used the Selectional Preferences (SPs) described in section 2.

We analyze the special case of *empatar*. This word is monosemous in Spanish while its English translations *tie* and *draw* are highly ambiguous (9 and 33 senses respectively). Moreover, the low number of selectional preferences acquired from BNC and Semcor allow to make a detailed analysis.

Table 10 presents the Object SPs acquired from equivalent translations of *empatar*, while table 11 shows the Object SPs acquired from Semcor. Being SemCor a sense disambiguated corpora, the SPs acquired from it tends to be more specific. There are only two SPs that overlap: 00017297n <event> and 00013018n <abstraction>.

Synset	Top Ontology	SF	Domain
00017297n	event	03	factotum
00013018n	abstraction	03	factotum
00017487n	human_activity, human_action, act	03	factotum
00020461n	phenomenon	03	factotum
00018376n	possession	03	factotum
00012865n	psychological_feature	03	psychology
00017954n	grouping, group	03	factotum
00016185n	state	03	factotum
00001740n	something, entity	03	factotum

Tabla 10: Object preferences acquired from BNC for the translations of *empatar*

Thus, if we perform direct intersections between the different sources, we obtain the following results:

- **Spanish EFE and Semcor** 00017487 (*acción*, action, *acto*, act), 00291567 (*juego*, play), 09768132 (*resultado* result, *puntuación* score).
- **Spanish EFE and BNC** 00017487 (action), 00017954 (group).

Synset	Top Ontology	SF	Domain
00017487n	human_activity, human_action,act	03	factotum
00013018n	abstraction	03	factotum
00261466n	activity	04	factotum
00020056n	quantum, amount, quantity, measure	03	metrology
00272358n	recreation, diversion	04	free_time
09765658n	number	23	math.
09756361n	definite_quantity	23	metrology
00291567n	game	04	play
09768132n	score	23	sport

Tabla 11: Object preferences acquired from SemCor for translations of *empatar*

- **English EFE and Semcor** 00291567 (game), 09768132 (score)
- **English EFE and BNC** 00017487 (action), 00017954 (group).
- **Spanish and English EFE** Gives 29 common synsets.

None of these intersections seems to be satisfactory enough. The most interesting result of this comparison is that the intersection between **Spanish EFE-English EFE and Semcor** are two synsets 00291567 (game) y 09768132 (score), both of SPORT domain.

None of these acquired knowledge resources seems to be accurate enough by its own. Instead, it seems to be a more appropriate to devise collaborative and productive ways to filter out too general or erroneous SPs.

5 *Conclusions and Future Work*

We have shown the feasibility to obtain large-scale semantic patterns for any language based only on shalow parsing and some basic semantic generalizations on wordnets. However, being this a preliminary and exploratory experiment (with many, hard and biased simplifications) we performed only a qualitative evaluation. We compared several semantic patterns coming from translation equivalent verbs selected from different languages and domains.

It seems clear that none of the large–scale semantic resources used in this experiment seems to be accurate enough by its own. Instead, it seems to be a more appropriate to devise collaborative and productive ways to filter out too general or erroneous patterns.

In order to continue performing collaborative multilingual knowledge acquisition analysis, we also need to ensure consistency outputs of the different Linguistic Processors (LPs) already available. This means for instance, to provide comparable full NERC capabilities to all LPs, anaphora, etc.

Regarding the EuroWordNet Top Ontology, a more detailed analysis is also planned. WordNet Semantic Files (or Lexicographic Files) can be seen as a simplification of EuroWordNet Top Ontology. The results obtained in this experiment suggest that the EuroWordNet Top Ontology (or the Suggested Upper Merged Ontology (SUMO) (Niles and Pease, 2001)) could be a good reference for generalising conceptual patterns such as agent or patient roles. We also plan to map Named Entities to these ontologies. We also plan to use more complex generalization slot mechanism e.g. using Conceptual Distance formulas.

Finally, selectional preferences has been used without expansion. This means that no inheritance has been performed. As the selectional preferences have been acquired by means of some kind of generalizations, we also plan to perform a full expansion process through the nominal part of the hierarchy.

References

Abney, S. 1991. Parsing by chunks. In *In R. C. Berwick, S. P. Abney, and Carol Tenny, editors, Principle-Based Parsing: Computation and Psycholinguistics*, Boston, MA. Kluwer Academic Publishers.

Agirre, E. and D. Martinez. 2001. Learning class-to-class selectional preferences. In *Proceedings of CoNLL01*, Toulouse.

Agirre, E. and D. Martinez. 2002. Integrating selectional preferences in wordnet. In *Proceedings of the first International WordNet Conference in Mysore*, India.

Brent, M. 1993. From grammar to lexicon: unsupervised learning of lexical syntax. *Computational Linguistics*, 19(2):243 – 262.

Briscoe, T. and J. Carroll. 1997. Automatic extraction of subcategorization from corpora. In *Proceedings of 5th Conference on Applied Natural Language Processing*, pages 356 – 363, Washington DC, USA.

Carroll, John, Ted Briscoe, and Antonio Sanfilippo. 1998. Parser evaluation: a survey and a new proposal. In *In Proceedings of*

the International Conference on Language Resources and Evaluation, pages 447–454.

Gale, W., K. Church, and D. Yarowsky. 1992. One sense per discourse. In *Proceedings of of DARPA speech and Natural Language Workshop*, Harriman, NY.

Ge, N., J. Hale, and E. Charniak. 1998. A statistical approach to anaphora resolution. In *Proceedings of the Sixth ACL/SIGDAT Workshop on Very Large Corpora*, pages 161–171.

Habash, N. and B. Dorr. 2002. Handling translation divergences: Combining statistical and symbolic techniques in generation-heavy machine translation. In *Proceedings of AMTA*, Tiburon.

Hindle, D. and M. Rooth. 1993. Structural ambiguity and lexical relations. *Computational linguistics*, 19(2):103–120.

Justeson, J. and S. Katz. 1995. Principled disambiguation: discriminating adjective senses with modified nouns. *Computational Linguistics*, 21(1):1–28.

Korhonen, A. 2002. *Subcategorization acquisition.* Ph.D. thesis, University of Cambridge.

Lapata, M. 2001. *The Acquisition and Modeling of Lexical Knowledge: A Corpus-based Investigation of Systematic Polysemy.* Ph.D. thesis, University of Edinburgh.

Li, H. and N. Abe. 1998. Generalizing case frames using a thesaurus and the mdl principle. *Computational linguistics*, 24(2):217–244.

Lin, D. 1998. Extracting collocations from text corpora. In *Proceedigns of First Workshop on Computational Terminology*, Montreal, Canada.

Lin, D. and P. Pantel. 2002. Concept Discovery from Text. In *15th International COLING*, Taipei, Taiwan.

Magnini, B. and G. Cavaglià. 2000. Integrating subject field codes into wordnet. In *In Proceedings of the Second Internatgional Conference on Language Resources and Evaluation LREC'2000*, Athens.

Manning, D. 1993. Automatic acquisition of a large subcategorization dictionary from corpora. In *Proceedings of 31th annual meeting of the Association for Computational Linguistics, ACL'93*, Columbus.

McCarthy, D. 2001. *Lexical Acquisition at the Syntax-Semantics Interface: Diathesis Aternations, Subcategorization Frames and Selectional Preferences.* Ph.D. thesis, University of Sussex.

McCarthy, D., J. Carroll, and J. Preiss. 2001. Disambiguating noun and verb senses using automatically acquired selectional preferences. In *Proceedings of the SENSEVAL-2 Workshop at ACL/EACL'01*, Toulouse, France.

Niles, I. and A. Pease. 2001. Towards a standard upper ontology. In *Proceedings of the 2nd International Conference on Formal Ontology in Information Systems.*

Resnik, P. 1997. Selectional preference and sense disambiguation. In *Proceedings of ACL SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What, and How?*, Washington, D.C.

Roland, D. and D. Jurafsky. 2002. Verb sense and verb subcategorization probabilities. In *Stevenson S. and Merlo P. (eds.) The Lexical Basis of Sentence Processing: Formal, Computational, and Experimental Issues*, John Benjamins, Amsterdam.

Roland, D., D. Jurafsky, L. Menn, S. Gahl, E. Elder, and C. Riddoch. 2000. Verb subcategorization frequency differences between business-news and balanced corpora. In *Proceedings of ACL Workshop on Comparing Corpora.*

Stevenson, S. and P. Merlo. 1998. Automatic verb classification using distributions of grammatical features. In *In Proc. of the 9th Conference of the EACL*, Bergen.

Vossen, P., editor. 1998. *EuroWordNet: A Multilingual Database with Lexical Semantic Networks* . Kluwer Academic Publishers .

Walde, S. 2000. Clustering Verbs Semantically According to their Alternation Behaviour. In *Proceedings of the 18th International Conference on Computational Linguistics (COLING-00)*, pages 747–753, Saarbrücken, Germany, August.