Projecting Corpus-Based Semantic Links on a Thesaurus*

Emmanuel Morin
IRIN
2, chemin de la housinière - BP 92208
44322 NANTES Cedex 3, FRANCE
morin@irin.univ-nantes.fr

Christian Jacquemin
LIMSI-CNRS
BP 133
91403 ORSAY Cedex, FRANCE
jacquemin@limsi.fr

Abstract

Hypernym links acquired through an information extraction procedure are projected on multi-word terms through the recognition of semantic variations. The quality of the projected links resulting from corpus-based acquisition is compared with projected links extracted from a technical thesaurus.

1 Motivation

In the domain of corpus-based terminology, there are two main topics of research: term acquisition—the discovery of candidate terms—and automatic thesaurus construction—the addition of semantic links to a term bank. Several studies have focused on automatic acquisition of terms from corpora (Bourigault, 1993; Justeson and Katz, 1995; Daille, 1996). The output of these tools is a list of unstructured multi-word terms. On the other hand, contributions to automatic construction of thesauri provide classes or links between single words. Classes are produced by clustering techniques based on similar word contexts (Schütze, 1993) or similar distributional contexts (Grefenstette, 1994). Links result from automatic acquisition of relevant predicative or discursive patterns (Hearst, 1992; Basili et al., 1993; Riloff, 1993). Predicative patterns yield predicative relations such as *cause* or *effect* whereas discursive patterns yield non-predicative relations such as generic/specific or synonymy links.

The main contribution of this article is to bridge the gap between term acquisition and thesaurus construction by offering a framework for organizing multi-word candidate terms with the help of automatically acquired links between single-word terms. Through the extraction of semantic variants, the semantic links between single words are projected on multi-word candidate terms. As shown in Figure 1, the input to the system is a tagged corpus. A partial ontology between single word terms and a set of multi-word candidate terms are produced after the first step. In a second step, layered hierarchies of multi-word terms are constructed through corpus-based conflation of semantic variants. Even though we focus here on generic/specific relations, the method would apply similarly to any other type of semantic relation.

The study is organized as follows. First, the method for corpus-based acquisition of semantic links is presented. Then, the tool for semantic term normalization is described together with its application to semantic link projection. The last section analyzes the results on an agricultural corpus and evaluates the quality of the induced semantic links.

2 Iterative Acquisition of Hypernym Links

We first present the system for corpus-based information extraction that produces hypernym links between single words. This system is built on previous work on automatic extraction of hypernym links through shallow parsing (Hearst, 1992; Hearst, 1998). In addition, our system incorporates a technique for the automatic generalization of lexico-syntactic patterns.

As illustrated by Figure 2, the system has two functionalities:

* The experiments presented in this paper were performed on [AGRO], a 1.3-million word French corpus of scientific abstracts in the agricultural domain. The termer used for multi-word term acquisition is ACABIT (Daille, 1996). It has produced 15,875 multi-word terms composed of 4,194 single words. For expository purposes, some examples are taken from [MEDIC], a 1.56-million word English corpus of scientific abstracts in the medical domain.

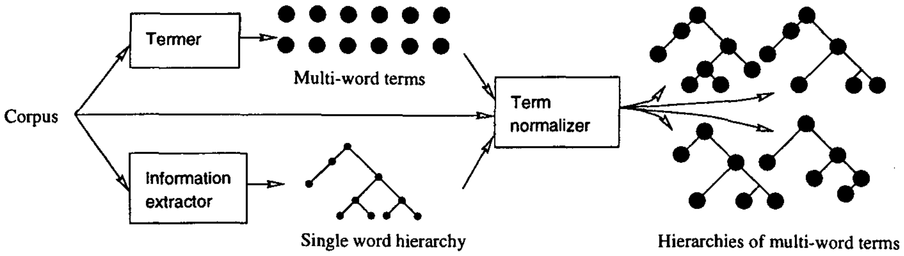

Figure 1: Overview of the system for hierarchy projection

1. The corpus-based acquisition of lexico-syntactic patterns with respect to a specific conceptual relation, here hypernym.
2. The extraction of pairs of conceptually related terms through a database of lexico-syntactic patterns.

Shallow Parser and Classifier

A shallow parser is complemented with a classifier for the purpose of discovering new patterns through corpus exploration. This procedure inspired by (Hearst, 1992; Hearst, 1998) is composed of 7 steps:

1. Select manually a representative conceptual relation, *e.g.* the hypernym relation.
2. Collect a list of pairs of terms linked by the previous relation. This list of pairs of terms can be extracted from a thesaurus, a knowledge base or manually specified. For instance, the hypernym relation *neocortex* IS-A *vulnerable area* is used.
3. Find sentences in which conceptually related terms occur. These sentences are lemmatized, and noun phrases are identified. They are represented as lexico-syntactic expressions. For instance, the previous relation HYPERNYM(*vulnerable area*, *neocortex*) is used to extract the sentence: *Neuronal damage were found in the selectively vulnerable areas such as neocortex, striatum, hippocampus and thalamus* from the corpus [MEDIC]. The sentence is then transformed into the following lexico-syntactic expression:[1]

$$\text{NP find in } \underline{\text{NP}} \text{ such as } \underline{\text{LIST}} \qquad (1)$$

4. Find a common environment that generalizes the lexico-syntactic expressions extracted at the third step. This environment is calculated with the help of a function of similarity and a procedure of generalization that produce candidate lexico-syntactic pattern. For instance, from the previous expression, and at least another similar one, the following candidate lexico-syntactic pattern is deduced:

$$\underline{\text{NP}} \text{ such as } \underline{\text{LIST}} \qquad (2)$$

5. Validate candidate lexico-syntactic patterns by an expert.
6. Use these validated patterns to extract additional candidate pairs of terms.
7. Validate candidate pairs of terms by an expert, and go to step 3.

Through this technique, eleven of the lexico-syntactic patterns extracted from [AGRO] are validated by an expert. These patterns are exploited by the information extractor that produces 774 different pairs of conceptually related terms. 82 of these pairs are manually selected for the subsequent steps our study because they are constructing significant pieces of ontology. They correspond to ten topics (trees, chemical elements, cereals, enzymes, fruits, vegetables, polyols, polysaccharides, proteins and sugars).

Automatic Classification of Lexico-syntactic Patterns

Let us detail the fourth step of the preceding algorithm that automatically acquires lexico-syntactic patterns by clustering similar patterns.

[1] NP stands for a noun phrase, and LIST for a succession of noun phrases.

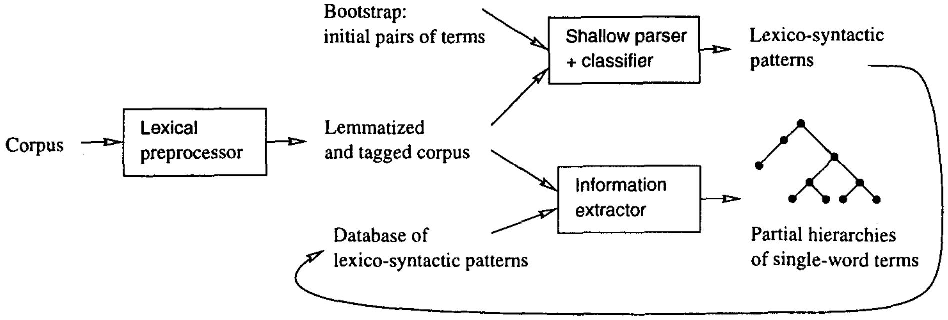

Figure 2: The information extraction system

As described in item 3. above, pattern (1) is acquired from the relation HYPERNYM(*vulnerable area,neocortex*). Similarly, from the relation HYPERNYM(*complication, infection*), the sentence: *Therapeutic complications such as infection, recurrence, and loss of support of the articular surface have continued to plague the treatment of giant cell tumor* is extracted through corpus exploration. A second lexico-syntactic expression is inferred:

$$\underline{\text{NP}} \text{ such as } \underline{\text{LIST}} \text{ continue to plague NP} \tag{3}$$

Lexico-syntactic expressions (1) and (3) can be abstracted as:[2]

$$\begin{aligned} &A = A_1 A_2 \cdots A_j \cdots A_k \cdots A_n \\ &HYPERNYM(A_j, A_k),\ k > j+1 \end{aligned} \tag{4}$$

and

$$\begin{aligned} &B = B_1 B_2 \cdots B_{j'} \cdots B_{k'} \cdots B_{n'} \\ &HYPERNYM(B_{j'}, B_{k'}),\ k' > j' + 1 \end{aligned} \tag{5}$$

Let $Sim(A, B)$ be a function measuring the similarity of lexico-syntactic expressions A and B that relies on the following hypothesis:

Hypothesis 2.1 (Syntactic isomorphy) *If two lexico-syntactic expressions A and B represent the same pattern then, the items A_j and $B_{j'}$, and the items A_k and $B_{k'}$ have the same syntactic function.*

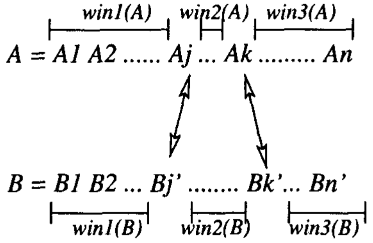
win1(A)
win2(A)
win3(A)
A = A1 A2 ...... Aj ... Ak .... An
B = B1 B2 ... Bj' .... Bk'... Bn'
win1(B)
win2(B)
win3(B)

Figure 3: Comparison of two expressions

Let $Win_1(A)$ be the window built from the first through *j-1* words, $Win_2(A)$ be the window built from words ranking from *j+1*th through *k-1*th words, and $Win_3(A)$ be the window built from *k+1*th through *n*th words (see Figure 3). The similarity function is defined as follows:

$$Sim(A, B) = \sum_{i=1}^{3} Sim(Win_i(A), Win_i(B)) \tag{6}$$

The function of similarity between lexico-syntactic patterns $Sim(Win_i(A), Win_i(B))$ is defined experimentally as a function of the longest common string.

After the evaluation of the similarity measure, similar expressions are clustered. Each cluster is associated with a candidate pattern. For instance, the sentences introduced earlier generate the unique candidate lexico-syntactic pattern:

$$\underline{\text{NP}} \text{ such as } \underline{\text{LIST}} \tag{7}$$

We now turn to the projection of automatically extracted semantic links on multi-word terms.[3]

[2] A_i is the ith item of the lexico-syntactic expression A, and n is the number of items in A. An item can be either a lemma, a punctuation mark, a symbol, or a tag (NP, LIST, etc.). The relation $k > j+1$ states that there is at least one item between A_j and A_k.

[3] For more information on the PROMÉTHÉE system, in

3 Semantic Term Normalization

The 774 hypernym links acquired through the iterative algorithm described in the preceding section are thus distributed: 24.5% between two multi-word terms, 23.6% between two single-word terms, and the remaining ones between a single-word term and a multi-word term. Since the terms produced by the termer are only multi-word terms, our purpose in this section is to design a technique for the expansion of links between single-word terms to links between multi-word terms. Given a link between *fruit* and *apple*, our purpose is to infer a similar link between *apple juice* and *fruit juice*, between any *apple* N and *fruit* N, or between *apple* N_1 and *fruit* N_2 with N_1 semantically related to N_2.

Semantic Variation

The extension of semantic links between single words to semantic links between multi-word terms is *semantic variation* and the process of grouping semantic variants is *semantic normalization*. The fact that two multi-word terms $w_1 w_2$ and $w'_1 w'_2$ contain two semantically-related word pairs (w_1,w'_1) and (w_2,w'_2) does not necessarily entail that $w_1 w_2$ and $w'_1 w'_2$ are semantically close. The three following requirements should be met:

Syntactic isomorphy The correlated words must occupy similar syntactic positions: both must be head words or both must be arguments with similar thematic roles. For example, *procédé d'élaboration* (process of elaboration) is not a variant *élaboration d'une méthode* (elaboration of a process) even though *procédé* and *méthode* are synonymous, because *procédé* is the head word of the first term while *méthode* is the argument in the second term.

Unitary semantic relationship The correlated words must have similar meanings in both terms. For example, *analyse du rayonnement* (analysis of the radiation) is not semantically related with *analyse de l'influence* (analysis of the influence) even though *rayonnement* and *influence* are semantically related. The loss of semantic relationship is due to the polysemy of *rayonnement* in French which means *influence* when it concerns a culture or a civilization and *radiation* in physics.

Holistic semantic relationship The third criterion verifies that the global meanings of the compounds are close. For example, the terms *inspection des aliments* (food inspection) and *contrôle alimentaire* (food control) are not synonymous. The first one is related to the quality of food and the second one to the respect of norms.

The three preceding constraints can be translated into a general scheme representing two semantically-related multi-word terms:

Definition 3.1 (Semantic variants) *Two multi-word terms $w_1 w_2$ and $w'_1 w'_2$ are semantic variants of each other if the three following constraints are satisfied:*[4]

1. *w_1 and w'_1 are head words and w_2 and w'_2 are arguments with similar thematic roles.*
2. *Some type of semantic relation S holds between w_1 and w'_1 and/or between w_2 and w'_2 (synonymy, hypernymy, etc.). The non semantically related words are either identical or morphologically related.*
3. *The compounds $w_1 w_2$ and $w'_1 w'_2$ are also linked by the semantic relation S.*

Corpus-based Semantic Normalization

The formulation of semantic variation given above is used for corpus-based acquisition of semantic links between multi-word terms. For each candidate term $w_1 w_2$ produced by the termer, the set of its semantic variants satisfying the constraints of Definition 3.1 is extracted from a corpus. In other words, a semantic normalization of the corpus is performed based on corpus-based semantic links between single words and variation patterns defined as all the

particular a complete description of the generalization patterns process, see the following related publication: (Morin, 1999).

[4] $w_1 w_2$ is an abbreviated notation for a phrase that contains the two content words w_1 and w_2 such that one of both is the head word and the other one an argument. For the sake of simplicity, only binary terms are considered, but our techniques would straightforwardly extend to n-ary terms with $n \geq 3$.

licensed combinations of morphological, syntactic and semantic links.

An exhaustive list of variation patterns is provided for the English language in (Jacquemin, 1999). Let us illustrate variant extraction on a sample variation:[5]

$$\begin{array}{l} \mathrm{N}_1 \text{ Prep } \mathrm{N}_2 \rightarrow \\ \underline{\mathcal{M}(\mathrm{N}_1,\mathrm{N})} \text{ Adv}^? \ \underline{\mathrm{A}^?} \ \underline{\text{Prep}} \ \underline{\text{Art}^?} \ \mathrm{A}^? \ \underline{\mathcal{S}(\mathrm{N}_2)} \end{array}$$

Through this pattern, a semantic variation is found between *composition du fruit* (fruit composition) and *composés chimiques de la graine* (chemical compounds of the seed). It relies on the morphological relation between the nouns *composé* (compound, $\mathcal{M}(\mathrm{N}_1,\mathrm{N})$) and *composition* (composition, N_1) and on the semantic relation (part/whole relation) between *graine* (seed, $\mathcal{S}(\mathrm{N}_2)$) and *fruit* (fruit, N_2). In addition to the morphological and semantic relations, the categories of the words in the semantic variant *composés*$_{\mathrm{N}}$ *chimiques*$_{\mathrm{A}}$ *de*$_{\mathrm{Prep}}$ *la*$_{\mathrm{Art}}$ *graine*$_{\mathrm{N}}$ satisfy the regular expression: the categories that are realized are underlined.

Related Work

Semantic normalization is presented as semantic variation in (Hamon et al., 1998) and consists in finding relations between multi-word terms based on semantic relations between single-word terms. Our approach differs from this preceding work in that we exploit domain specific corpus-based links instead of general purpose dictionary synonymy relationships. Another original contribution of our approach is that we exploit simultaneously morphological, syntactic, and semantic links in the detection of semantic variation in a single and cohesive framework. We thus cover a larger spectrum of linguistic phenomena: morpho-semantic variations such as *contenu en isotope* (isotopic content) a variant of *teneur isotopique* (isotopic composition), syntactico-semantic variants such as *contenu en isotope* a variant of *teneur en isotope* (isotopic content), and morpho-syntactico-semantic variants such as *dureté de la viande* (toughness of the meat) a variant of *résistance et la rigidité de la chair* (lit. resistance and stiffness of the flesh).

4 Projection of a Single Hierarchy on Multi-word Terms

Depending on the semantic data, two modes of representation are considered: a *link mode* in which each semantic relation between two words is expressed separately, and a *class mode* in which semantically related words are grouped into classes. The first mode corresponds to synonymy links in a dictionary or to generic/specific links in a thesaurus such as (AGROVOC, 1995). The second mode corresponds to the synsets in WordNet (Fellbaum, 1998) or to the semantic data provided by the information extractor. Each class is composed of hyponyms sharing a common hypernym—named *co-hyponyms*—and all their common hypernyms. The list of classes is given in Table 1.

Analysis of the Projection

Through the projection of single word hierarchies on multi-word terms, the semantic relation can be modified in two ways:

Transfer The links between concepts (such as fruits) are transferred to another conceptual domain (such as juices) located at a different place in the taxonomy. Thus the link between *fruit* and *apple* is transferred to a link between *fruit juice* and *apple juice*, two hyponyms of juice. This modification results from a semantic normalization of argument words.

Specialization The links between concepts (such as fruits) are specialized into parallel relations between more specific concepts located lower in the hierarchy (such as dried fruits). Thus the link between *fruit* and *apple* is specialized as a link between *dried fruits* and *dried apples.* This modification is obtained through semantic normalization of head words.

The Transfer or the Specialization of a given hierarchy between single words to a hierarchy between multi-word terms generally does not preserve the full set of links. In Figure 4, the initial hierarchy between *plant products* is only partially projected through Transfer on *juices* or *dryings of plant products* and through Specialization on *fresh* and *dried plant products.* Since multi-word terms are more specific than

[5] The symbols for part of speech categories are N (Noun), A (Adjective), Art (Article), Prep (Preposition), Punc (Punctuation), Adv (Adverb).

Table 1: The twelve semantic classes acquired from the [AGRO] corpus

Classes	Hypernyms and co-hyponyms
trees	arbre, bouleau, chêne, érable, hêtre, orme, peuplier, pin, poirier, pommier, sapin, épicéa
chemical elements	élément, calcium, potassium, magnésium, manganése, sodium, arsenic, chrome, mercure, sélénium, étain, aluminium, fer, cadium, cuivre
cereals	céréale, maïs, mil, sorgho, blé, orge, riz, avoine
enzymes	enzyme, aspartate, lipase, protéase
fruits	fruit, banane, cerise, citron, figue, fraise, kiwi, noix, olive, orange, poire, pomme, pêche, raisin
olives	fruit, olive, Amellau, Chemlali, Chétoui, Lucques, Picholine, Sevillana, Sigoise
apples	fruit, pomme, Cartland, Délicious, Empire, McIntoch, Spartan
vegetables	légume, asperge, carotte, concombre, haricot, pois, tomate
polyols	polyol, glycérol, sorbitol
polysaccharides	polysaccharide, amidon, cellulose, styrène, éthylbenzène
proteins	protéine, chitinase, glucanase, thaumatin-like, fibronectine, glucanase
sugars	sucre, lactose, maltose, raffinose, glucose, saccharose

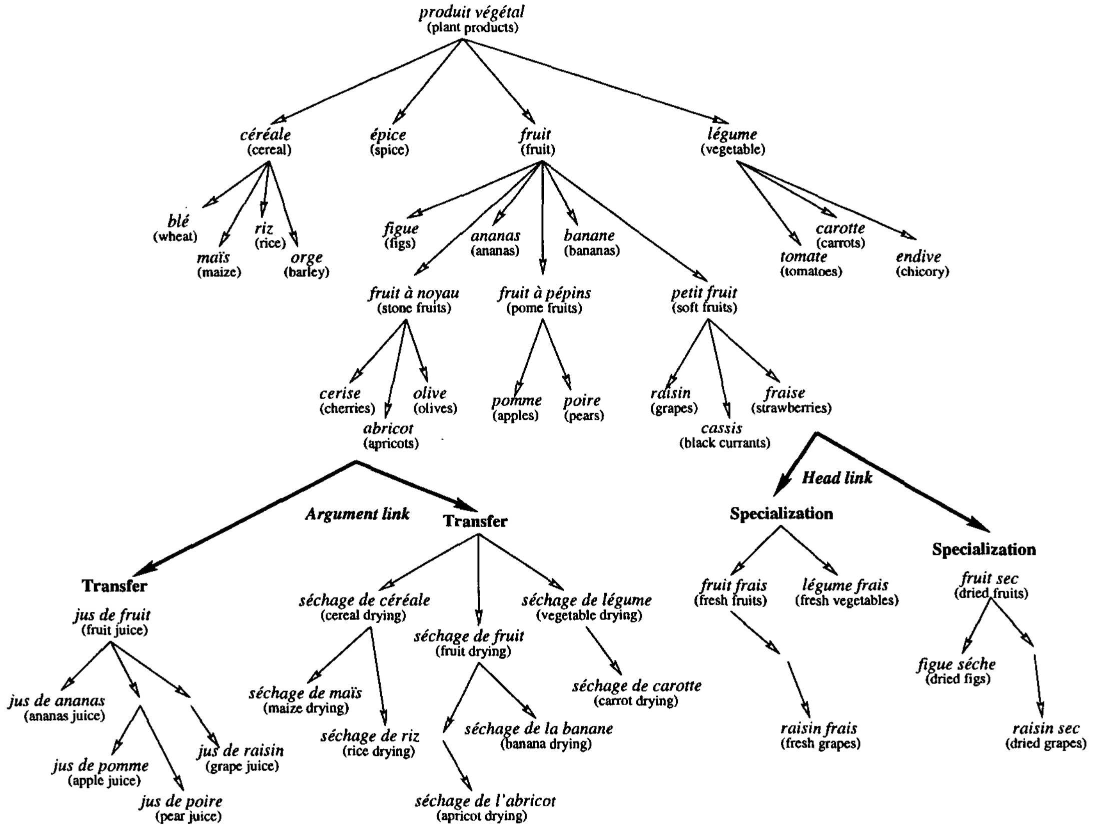

Figure 4: Projected links on multi-word terms (the hierarchy is extracted from (AGROVOC, 1995))

single-word terms, they tend to occur less frequently in a corpus. Thus only some of the possible projected links are observed through corpus exploration.

5 Evaluation

Projection of Corpus-based Links

Table 2 shows the results of the projection of corpus-based links. The first column indicates the semantic class from Table 1. The next three columns indicate the number of multi-word links projected through Specialization, the number of correct links and the corresponding value of precision. The same values are provided for Transfer projections in the following three columns.

Transfer projections are more frequent (507 links) than Specializations (77 links). Some classes, such as *chemical elements*, *cereals* and *fruits* are very productive because they are composed of generic terms. Other classes, such as *trees*, *vegetables*, *polyols* or *proteins*, yield few semantic variations. They tend to contain more specific or less frequent terms.

The average precision of Specializations is relatively low (58.4% on average) with a high standard deviation (between 16.7% and 100%). Conversely, the precision of Transfers is higher (83.8% on average) with a smaller standard deviation (between 69.0% and 100%). Since Transfers are almost ten times more numerous than Specializations, the overall precision of projections is high: 80.5%.

In addition to relations between multi-word terms, the projection of single-word hierarchies on multi-word terms yields new candidate terms: the variants of candidate terms produced at the first step. For instance, *séchage de la banane* (banana drying) is a semantic variant of *séchage de fruits* (fruit drying) which is not provided by the first step of the process. As in the case of links, the production of multi-word terms is more important with Transfers (72 multi-word terms) than Specializations (345 multi-word terms) (see Table 3). In all, 417 relevant multi-word terms are acquired through semantic variation.

Comparison with AGROVOC Links

In order to compare the projection of corpus-based links with the projection of links extracted from a thesaurus, a similar study was made using semantic links from the thesaurus (AGROVOC, 1995).[6]

The results of this second experiment are very similar to the first experiment. Here, the precision of Specializations is similar (57.8% for 45 links inferred), while the precision of Transfers is slightly lower (72.4% for 326 links inferred). Interestingly, these results show that links resulting from the projection of a thesaurus have a significantly lower precision (70.6%) than projected corpus-based links (80.5%).

A study of Table 3 shows that, while 197 projected links are produced from 94 corpus-based links (ratio 2.1), only 88 such projected links are obtained through the projection of 159 links from AGROVOC (ratio 0.6). Actually, the ratio of projected links is higher with corpus-based links than thesaurus links, because corpus-based links represent better the ontology embodied in the corpus and associate more easily with other single word to produce projected hierarchies.

6 Perspectives

Links between single words projected on multi-word terms can be used to assist terminologists during semi-automatic extension of thesauri. The methodology can be straightforwardly applied to other conceptual relations such as synonymy or meronymy.

Acknowledgement

We are grateful to Gaël de Chalendar (LIMSI), Thierry Hamon (LIPN), and Camelia Popescu (LIMSI & CNET) for their helpful comments on a draft version of this article.

References

AGROVOC. 1995. *Thésaurus Agricole Multilingue.* Organisation de Nations Unies pour l'Alimentation et l'Agriculture, Roma.

Roberto Basili, Maria Teresa Pazienza, and Paola Velardi. 1993. Acquisition of selectional patterns in sublanguages. *Machine Translation*, 8:175–201.

Didier Bourigault. 1993. An endogeneous corpus-based method for structural noun phrase disambiguation. In *EACL'93*, pages 81–86, Utrecht.

Béatrice Daille. 1996. Study and implementation of combined techniques for automatic extraction of terminology. In Judith L. Klavans and Philip Resnik, editors, *The Balancing Act: Combining Symbolic and Statistical*

[6] (AGROVOC, 1995) is composed of 15,800 descriptors but only single-word terms found in the corpus [AGRO] are used in this evaluation (1,580 descriptors). From these descriptors, 168 terms representing 4 topics (cultivation, plant anatomy, plant products and flavorings) are selected for the purpose of evaluation.

Table 2: Precision of the projection of corpus-based links

Classes	**Specialization**			**Transfer**		
	# Occ.	**Correct occ.**	**Precision**	**# Occ.**	**Correct occ.**	**Precision**
trees	0	-	-	3	3	100.0%
chemical elements	8	4	50.0%	101	99	98.0%
cereals	6	1	16.7%	76	65	85.5%
enzymes	3	3	100.0%	29	20	69.0%
fruits	32	20	62.5%	214	172	80.4%
olives	4	1	25.0%	10	8	80.0%
apples	4	1	25.0%	16	12	75.0%
vegetables	3	2	66.7%	3	3	100.0%
polyols	0	-	-	0	-	-
polysaccharides	3	1	33.3%	13	11	84.6%
proteins	0	-	-	8	6	75.0%
sugars	13	11	84.6%	34	26	76.5%
Total	77	45	58.4%	507	425	83.8%

Table 3: Production of new terms and correct links through the projection of links

	Corpus-based links		**Thesaurus-based links**	
	Terms	**Relations**	**Terms**	**Relations**
Initial links	96	94	162	159
Specialization	72	30	49	18
Transfer	345	167	256	70
Total	417	197	305	88

Approaches to Language, pages 49–66. MIT Press, Cambridge, MA.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database.* MIT Press, Cambridge, MA.

Gregory Grefenstette. 1994. *Explorations in Automatic Thesaurus Discovery.* Kluwer Academic Publisher, Boston, MA.

Thierry Hamon, Adeline Nazarenko, and Cécile Gros. 1998. A step towards the detection of semantic variants of terms in technical documents. In *COLING-ACL'98*, pages 498–504, Montreal.

Marti A. Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In *COLING'92*, pages 539–545, Nantes.

Marti A. Hearst. 1998. Automated discovery of wordnet relations. In Christiane Fellbaum, editor, *WordNet: An Electronic Lexical Database.* MIT Press, Cambridge, MA.

Christian Jacquemin. 1999. Syntagmatic and paradigmatic representation of term variation. In *ACL'99*, University of Maryland.

John S. Justeson and Slava M. Katz. 1995. Technical terminology: some linguistic properties and an algorithm for identification in text. *Natural Language Engineering*, 1(1):9–27.

Emmanuel Morin. 1999. Using Lexico-syntactic Patterns to Extract Semantic Relations between Terms from Technical Corpus. In *Proceedings, 5th International Congress on Terminology and Knowledge Engineering (TKE'99)*, Innsbrück.

Ellen Riloff. 1993. Automatically constructing a dictionay for information extraction tasks. In *Proceedings, 11th National Conference on Artificial Intelligence*, pages 811–816, Cambridge, MA. MIT Press.

Hinrich Schütze. 1993. Word space. In Stephen J. Hanson, Jack D. Cowan, and Lee Giles, editors, *Advances in Neural Information Processing Systems 5.* Morgan Kauffmann, San Mateo, CA.