How to Encode Semantic Knowledge: A Method for Meaning Representation and Computer-Aided Acquisition

Paola Velardi*
Universita of Ancona

Maria Teresa Pazienza†
University of Roma II

Michela Fasolo‡
Cervedomani

Natural language processing will not be able to compete with traditional information retrieval unless high-coverage techniques are developed. It is commonly agreed that a poor encoding of the semantic lexicon is the bottleneck of many existing systems. A hand encoding of semantic knowledge on an extensive basis is not realistic; hence, it is important to devise methods by which such knowledge can be acquired in part or entirely by a computer. But what type of semantic knowledge could be automatically learned, from which sources, and by what methods? This paper explores the above issues and proposes an algorithm to learn syncategorematic concepts from text exemplars. What is learned about a concept is not its defining features, such as kinship, but rather its patterns of use.

The knowledge acquisition method is based on learning by observations*; observations are examples of word co-occurrences (collocations) in a large corpus, detected by a morphosyntactic analyzer. A semantic bias is used to associate collocations with the appropriate meaning relation, if one exists. Based upon single or multiple examples, the acquired knowledge is then generalized to create semantic rules on concept uses.*

Interactive human intervention is required in the training phase, when the bias is defined and refined. The duration of this phase depends upon the semantic closure of the sublanguage on which the experiment is carried out. After training, final approval by a linguist is still needed for the acquired semantic rules. At the current stage of experimentation of this system, it is unclear whether and when human supervision could be further reduced.

1. Four Important Questions in Lexical Semantics

In the past few years there has been a growing interest in the field of lexical semantics. It is now commonly agreed that the the acquisition of the semantic lexicon is a central issue in computational linguistics. The major contributions on this topic are collected in *Computational Linguistics* (1987), Evens (1988), and Zernik (in press).

Research in lexical semantics is, however, rather heterogeneous in scope, methods, and results. Paraphrasing the famous "What, when and how?" questions introduced by Kathleen McKeown (1985) in her studies on language generation, we pose four similar questions that we believe important to understand the consequences of conceptual and

* Instituto di Informatica, via Brecce Bianche, Ancona
† Dipartimento di Elettronica, Roma
‡ corso Stati Uniti, Padova

linguistic decisions in lexical semantics:

1. *Why?* Why is lexical semantics useful in the first place? Linguists and psychologists are interested in the study of word senses to shed light on important aspects of human communication, such as concept formation and language use. Lexicographers need computational aids to analyze in a more compact and extensive way word definitions in dictionaries. Computer scientists need semantics for the purpose of natural language processing.

   The option and choices in lexical semantics are deeply related to the ultimate objective of research.[1]

2. *What?* What is encoded in a semantic lexicon? Each word is a world. Despite the interest that semantics has received from the scholars of different disciplines since the early history of humanity, a unifying theory of meaning does not exist. In practice, the type and quality of the expressed phenomena again depend upon the end user: a psychologist, a lexicographer, or a computer. For example, much work on lexical acquisition from dictionaries produces lexical entries in a format that is clearly useful for lexicographers (see the many examples in Byrd (1987) and Boguraev (1989)), but their utility for the purpose of automatic language processing is less evident. In fact, the output of a lexical analyzer is an entry in which several semantic and syntactic fields are identified, but many of these fields are raw text.

In psychology and linguistics, semantic knowledge is modeled with very deep, more or less formal, expressions (see Figure 1 in Section 2). Often semantic models focus on some very specific aspect of language communication, according to the scientific interest of a scholar.

In natural language processing (NLP hereafter), lexical entries typically express linguistic knowledge as commonsensically understood and used by humans. The entries are entirely formatted in some knowledge representation language and can be manipulated by a computer. Within the field of NLP, many options are still possible, as summarized in Section 2.

3. *Where?* What are the sources of lexical semantic knowledge?

   - *Introspection* is often a source of data, no matter what is the background of the scientist. However, introspection poses theoretical and implementation problems. Theoretical, because "different researchers with different theories would observe different things about their internal thoughts..." (Anderson 1989). Implementation, because consistency becomes a major problem when the size of the lexicon exceeds a few hundred entries.
   - *Psycholinguistic experiments*, such as verbal protocols, are more appropriate, because they produce observable and measurable

1 This is not to say that there is a clear cut among the interests of linguists, psychologists, and computer scientists. We believe that, just to take one example, a poor knowledge of the results and methods of lexicography, linguistics, and cognitive science in part motivates the nonstriking success of NLP in developing large-scale systems (Velardi 1990).

example 1 from (Leech 1981):
boy = +animate −adult +male

example 2 from (Melčuk 1987):
help =
Y carrying out Z, X uses his resources W in order for W to help Y to carry out Z; the use of resources by X and the carrying out of Z by Y are simultaneous

example 3 from (Schank 1972):
throw =
actor PROPELs and object from a source LOCation to a destination LOCation

Figure 1
Examples of conceptual meaning representation in the literature

data. One of the largest hand-encoded semantic lexicons (Dahlgren 1989 and Dahlgren, communication to AAAI90 Stanford seminars) was built by asking subjects to freelist features common to objects. Despite the scientific interest of such experiments, they cannot be extensively repeated for the purpose of acquiring several thousand word sense definitions.

- *On-line corpora and dictionaries* are widely available today and provide experimental evidence of word uses and word definitions. The major advantage of on-line resources is that in principle they provide the basis for very large experiments, even though at present the methods of analysis are not fully developed. This paper describes an experiment in lexical acquisition from corpora.

4. *How?* The final issue is implementation. Implementation may be thought of at various levels. It is the hard work of implementing a system in a real domain, or the more conceptual task of defining a mathematical framework to manipulate the objects defined within a linguistic model.

Quite obviously the "hows" in the literature are much more than the "wheres" and "whats". No classification of lexical semantic methods is attempted here for sake of brevity. In Section 5 we examine the research that is more closely related to the problem under examination in this paper, i.e. lexical acquisition for natural language processing.

2. "What?": Meaning Representation in a Semantic Lexicon

This section further considers the issue: what knowledge can be encoded in a semantic lexicon? We will not attempt an overall survey of the field of semantics, which has provided material for many fascinating books; rather, we will focus on the very practical problem of representing language expressions on a computer, in a way that can be useful for natural language processing applications, e.g. machine translation, information retrieval, user-friendly interfaces.

In the field of NLP, several approaches have been adopted to represent semantic knowledge. We are not concerned here with semantic languages, which are relatively well developed, but with meaning representation principles.

A thorough classification of meaning types was attempted by Leech in his book on semantics (Leech 1981). In surveying the meaning representation styles adopted in the computational linguistic literature, we found that many implemented natural language processors adopt one of the following two meaning types for the lexicon: conceptual (or deep) and collocative (or superficial).

1. *Conceptual meaning.* Conceptual meaning is the cognitive content of words; it can be expressed by *features* or by *primitives*: conceptual meaning is "deep" in that it expresses phenomena that are deeply embedded in language.
2. *Collocative meaning.* What is communicated through associations between words or word classes. Collocative meaning is "superficial" in that it does not seek "the real sense" of a word, but rather "describes" its uses in everyday language, or in some subworld language (economy, computers, etc.). It provides more than a simple analysis of co-occurrences, because it attempts an explanation of word associations in terms of *meaning relations* between a lexical item and other items or classes.

The conceptual vs. collocative distinction in computational linguistics closely corresponds to the defining vs. syncategorematic features distinction in psychology. Syncategorematic[2] concepts (Keil 1989) are those "almost entirely defined by their pattern of use, and (the) others by almost pure belief." The use of syncategorematic concepts is supported by evidence from psycholinguistic studies. Humans more naturally describe word senses with their characteristics and relations with other words than with kindship and other internal features. Not surprisingly, a similar "naturalness" is observed in the examples of collocative meaning descriptions. Examples show an evident similarity and are easily readable, whereas concept definitions in the conceptual framework are very different in different papers, and more obscure.[3]

Both conceptual (defining) and collocative (syncategorematic) features are formally represented in the NLP literature using some subjective, human-produced set of primitives (conceptual dependencies, semantic relations, conceptual categories) on which there is no shared agreement at the current time. As far as conceptual meaning is concerned, the quality and quantity of phenomena to be shown in a representation is subjective as well: the linguist relies mostly on his/her introspection. Collocative meaning can rely on the solid evidence represented by word associations; the interpretation of an association is subjective, but valid associations are an observable, even though vast, phenomenon.

In principle, the inferential power of collocative, or surface, meaning representation is lower than for conceptual meaning, because it does not account for many

2 Since the terms surface semantics, collocative meaning, and syncategorematic features all refer to the very fact that concepts (in humans and computers) are frequently described by lists of characteristics and usage types, we will use all the above terms interchangeably.

3 Notice in the examples of collocative meaning descriptions that the "is-a" relation is conceptual, not collocative, in nature. Kindship relations are very important in NLP because they provide the basis for inferences in semantic interpretation. However, the classification of word senses in type hierarchies is a conceptually very complex, almost unsolvable problem.

important aspects of human communication, such as beliefs, preconditions, and knowledge of cause-effect relations. These phenomena cannot be captured by an analysis of meaning relations between uttered words in a sentence.

Despite this, collocative meaning has been shown to be a useful knowledge scheme for a number of computer applications in language processing. In Niremburg (1987), the validity of this approach is demonstrated for TRANSLATOR, a system used for machine translation in the computer subworld. A semantic knowledge framework in the style of collocative meaning is also adopted in Ace (Jacobs 1987), which has been used by the TRUMP language analyzer in a variety of applications.

In our previous work on semantic knowledge representation (Pazienza 1988, Velardi 1988, Antonacci 1989) we showed that a semantic dictionary in the style of collocative meaning is a powerful basis for semantic interpretation.

The knowledge power provided by the semantic lexicon (about 1000 manually entered definitions) was measured by the capability of the language processor DANTE to answer a variety of questions concerning previously analyzed sentences (press agency releases on economics). It was found that, even though the system was unable to perform complex inferences, it could successfully answer more than 90% of the questions (Pazienza 1988).[4]

Representing word senses and sentences with surface semantics is hence useful (though not entirely sufficient) for many NLP applications.

An additional and very important advantage of surface semantics is that it makes it feasible to acquire large lexicons, as discussed in the following sections. "Acquirability" in our view is extremely important to evaluate a knowledge representation framework.

3. "Where?": Sources for Acquiring Lexical Semantic Knowledge

Acquiring semantic knowledge on a systematic basis is quite a complex task. One needs not to look at metaphors or idioms to find this; even the interpretation of apparently simple sentences is riddled with such difficulties that it is difficult to isolate a piece of the problem. A manual codification of the lexicon is a prohibitive task, regardless of the framework adopted for semantic knowledge representation; even when a large team of *knowledge enterers* is available, consistency and completeness are a major problem. We believe that automatic or semi-automatic acquisition of the lexicon is a critical factor in determining how widespread the use of natural language processors will be in the next few years.

Recently a few methods were presented for computer-aided semantic knowledge acquisition. The majority of these methods use standard on-line dictionaries as a source of data.

The information presented in a dictionary has in our view some intrinsic limitations:

- definitions are often *circular*; e.g., the definition of a term A may refer to a term B that in turn points to A;
- definitions are *not homogeneous* as far as the quality and quantity of information provided: they can be very sketchy, or give detailed

4 The test was performed over a six-month period on about 50 occasional visitors and staff members of the IBM Rome scientific center, unaware of the system capabilities and structure. The user would look at 60 different releases, previously analyzed by the system (or re-analyzed during the demo), and freely ask questions about the content of these texts. See the referenced papers for examples of sentences and of (answered and not answered) query types.

structural information, or list examples of use-types, or describe more internal features;

- a dictionary is the result of a conceptualization effort performed by some human specialist(s); this effort may not be consistent with, or suitable for, the objectives of an application for which a language processor is built.

A second approach is using *corpora* rather than human-oriented dictionary entries. Corpora provide experimental evidence of word uses, word associations, and such language phenomena as metaphors, idioms, and metonyms. Corpora are a genuine, "naive" example of language use, whereas dictionaries result from an effort of introspection performed by language experts, i.e. lexicographers. Corpora are more interesting than dictionaries as a source of linguistic knowledge, just as tribes are more interesting than "civilized" communities in anthropology.

The problem and at the same time the advantage of corpora is that they are raw texts; dictionary entries use some formal notation that facilitates the task of linguistic data processing.

No computer program may ever be able to derive formatted data from a completely unformatted source. Hence the ability to extract lexical semantic information from a corpus depends upon a powerful set of *mapping rules* between phrasal patterns and human-produced semantic primitives and relations. In machine learning, this is referred to as the *semantic bias*.

There is no evidence of innate conceptual primitives, apart from some very general ones (time, animacy, place, etc.), and even on these there is no shared agreement. We must hence accept the intrinsic limitation of using a bias whose source is the introspection of a single, or of a community of scientists. But even though the symbols we choose are arbitrary, their role is the prediction of basic statements, i.e. the processing of NL sentences in a way that is useful to some computational purpose, and should be evaluated on this ground.

4. A Method for the Acquisition and Interpretation of a Semantic Lexicon

Our research on lexical acquisition from corpora started in 1988, when a first version of the system was built as utility for the DANTE natural language processor (Velardi 1989), a system that analyzes press agency releases on finance and economics. The current version, described hereafter, is a self-contained tool on which we are running a large experiment using a nationwide corpus of enterprise descriptions (overall, more than one million descriptions). The objective is to derive a domain-dependent semantic lexicon of 10,000 entries to be used for information retrieval in the sub-domain of agricultural enterprises (Fasolo 1990). The project is a cooperative effort among the Universities of Ancona and Roma II, and the CERVED, the company that owns and manages the database of all commercial enterprises registered at the Chambers of Commerce in Italy.

In the current version, the system is able to acquire syncategorematic concepts, learning and interpreting patterns of use from text exemplars. Generated lexical entries are of the type shown in Figure 2 of Section 2. We do not exclude the possibility of acquiring defining features from dictionary entries and query collections at a later stage of this project. At present, however, we are interested in fully exploring the power of collocative semantics for NLP. We are also interested in the analysis of the linguistic material that is being produced by the system.

In what follows the methodology is described in detail.

example 1 from (Velardi 1988)

agreement =
is_a decision_act
participant person, organization
theme transaction
cause communication_exchange
manner interesting important effective ...

example 2 from (Niremburg 1987):
person =
isa creature
agent_of take put find speech-action mental-action
consist_of hand foot...
source_of speech-action
destination_of speech-action
power human
speed slow
mass human

Figure 2
Examples of collocative meaning representation in the literature

4.1 What Is Given

The input to the system is:

1. *a list of syntactic collocates*, e.g. subject-verb, verb-object, noun-preposition-noun, noun-adjective, etc. extracted through morphologic (Russo 1987) and syntactic analysis of the selected corpus. The level at which syntactic analysis should be performed to derive collocates is a matter of debate. In Smadja (1989) it is suggested that parsing can be avoided by simply examining the neighborhood of a word w at a distance of ±5. Our experience demonstrates that this algorithm produces too many collocations, of which a minority are actually semantically related.

   At the other extreme is full syntactic parsing, as performed in Velardi (1989). This is computationally too expensive for large corpora and fails to produce useful collocations when sentences are not fully grammatical, as for example in the agricultural businesses database. A typical text in this corpus is:

   vendita al minuto di legno da costruzione e manufatti in abete produzione mobili di legno e di metallo

   (*) retail sale of wood for construction and hand-manufactured in fir-tree production furniture of wood and of metal

   A better trade-off between speed and accuracy is to enable syntactic parsing of sentence parts and use some context-dependent heuristics to cut sentences into clauses (Fasolo 1990). However, the ±5 collocations are also collected for a reason that will be clarified later on.

2. *A semantic bias.* The semantic bias is the kernel of any learning algorithm, as no system can learn much more than what it already knows (Micalski 1983). This consists of:

   (a) a domain-dependent *concept hierarchy.* This is a many-to-many mapping from words to word sense *names* and an ordered list of conceptual categories.

   The hypothesis of hand-entering a type hierarchy would not require an unreasonable amount of time, because the task is comparable to entering a morphologic lexicon (Russo 1987). The problem is rather a conceptual one. The way humans categorize concepts in classes is far from being understood. Mere property inheritance seems to be inadequate at fully modeling categorization in humans (Lakoff 1987; Rosch 1975), and this very fact discouraged us from attempting some automatic hierarchy acquisition in the framework of machine learning (Gennari 1989). The limitations of current machine learning approaches when applied to language learning are discussed in Section 5.

   We consider the problem of acquiring type hierachies as an open issue, to which more thought and more research are being devoted in our current work.

   (b) a set of domain-dependent *conceptual relations,* and a many-to-many mapping (*synt-sem*) between syntactic relations and the corresponding conceptual relations (see Velardi 1988 and Antonacci 1989 for extensive examples of syntax-to-semantics mapping);

   (c) a set of coarse-grained *selectional restrictions* on the use of conceptual relations, represented by concept-relation-concept (CRC) triples. CRCs are expressed in Conceptual Graph notation (Sowa 1984).

4.2 The Output

The system produces two types of output:

1. a set of fine-grained CRCs, that are clustered around concepts or around conceptual relations;
2. an average-grained semantic knowledge base, organized in CRC triples.

The semantic knowledge base is acquired from a source sub-corpus and is used, before a final approval, for the semantic interpretation of sentences in a test-bed sub-corpus. The semantic interpreter is basically that described in Velardi (1988) and in other papers, to which the interested reader may refer.

Such terms as 'fine,' 'average,' and 'coarse' are obviously fuzzy. To the extent this makes sense, we ranged the above terms as follows:

i) fine-grained CRC are those in which concepts directly map into content words (e.g. [COW] ← (PATIENT) ← [BREED]). These CRCs are *true* because they are observed in the domain subworld.

ii) average-grained CRC are those in which concepts are fathers or grand-fathers of content-word concepts. These CRC are 'typically' true, as they may have a limited number of exceptions observed in the domain sublanguage (e.g. [ANIMAL] ← (PATIENT) ← [BREED] is typically true, even though breeding mosquitoes is quite odd).

iii) coarse-grained CRC are those in which concepts are at a higher level in the taxonomy (e.g. [ACTION] → (BENEFICIARY) → [ANIMATE_ENTITY]). They state necessary, but not sufficient, conditions on the use of conceptual relations.

The notion of "high-level" and "low-level" in a taxonomy is also relative to the application domain. For example, in a computer world, the concept COMPUTER_SOFTWARE may be rather high-level.

4.3 Learning Syncategorematic Concepts from Text Exemplars

To acquire syncategorematic knowledge on concepts, the algorithm proceeds as follows: For any syntactic collocate sc(w1,w2):

1. Restrict the set of conceptual relations that could correspond to the syntactic collocate using the synt-sem table.
2. Use coarse-grained knowledge and taxonomic knowledge to further restrict the hypotheses.
3. If no interpretation is found, reject the collocate. If one or more interpretations is found, put the resulting CRC(s) on a temporary knowledge base of fine-grained knowledge.
4. Generalize the result by replacing the concepts in the CRC with their closest supertypes, using the structural overcommitment principle (Webster 1989). Add the result to a temporary knowledge base of average-grained knowledge.
5. Repeat steps 1–4 for all the collocates of the same syntactic type, or (user choice) those including the same word W. Further generalize one step up in the hierarchy, based on at least three examples.
6. Present the results to the linguist for a final approval, then add to the permanent knowledge base.

In a first, training phase, the linguist is requested to inspect the system output in step 3, to verify and refine the semantic bias.

4.4 A Complete Example

Figure 3 shows the output of step 3. On the left-hand side of window 1 (ok-sema) and window 2 (no-sema) flow the syntactic collocates acquired during the syntactic analysis of the corpus. Syntactic collocates are couples, like adj-noun, noun-verb, verb-noun, or triples, like verb-prep-noun (G_V_P_N), noun-prep-noun (G_N_P_N) etc. In Figure 3 the triples are shown with the preposition "in".

For each collocate, the learner accesses the knowledge provided by the semantic bias and generates a CRC triple, corresponding to a plausible semantic interpretation

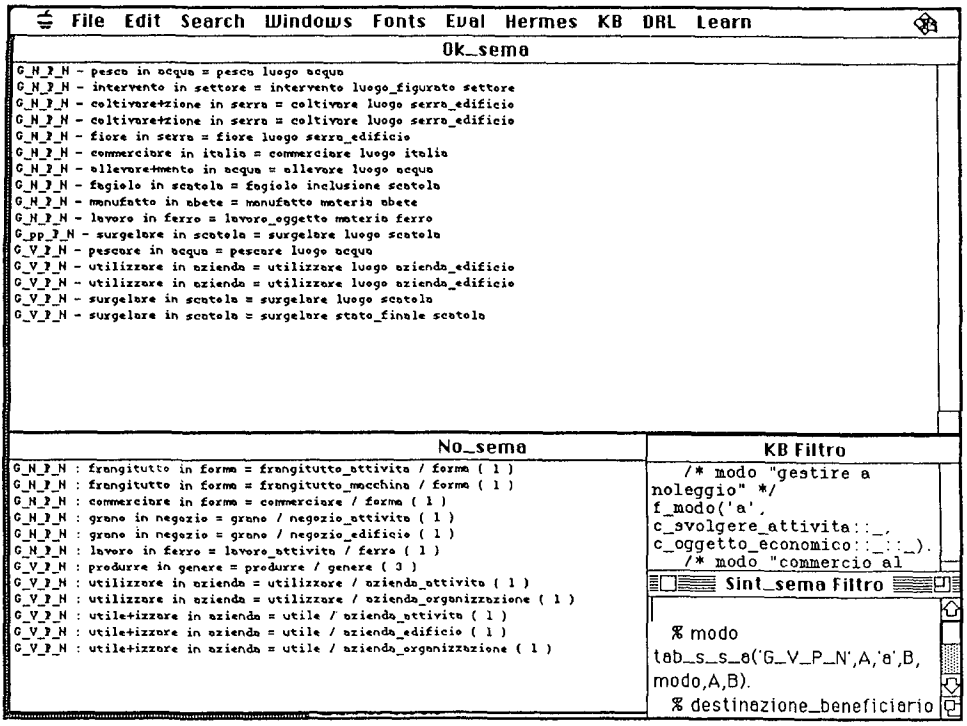

Figure 3
Production of CRC triples from syntactic collocates

of the syntactic collocate (shown to the right of the '=' in the window ok-sema). If an interpretation is not found, because the syntactic attachment is not semantically validated, the collocate is shown in the window no-sema, i.e., it is rejected.[5]

An example may clarify the steps:

1. acquired syntactic collocate: "coltivazione in serra" (farming in greenhouse)
2. possible subsumed conceptual relations (in Italian and in the specific domain) by the noun-"in"-noun collocation:
   luogo (location) (ex.: farming in greenhouse)
   stato_finale (final state) (ex.: trasformazione in vino = transformation in wine)
   materia (matter) (ex.: lavoro in ferro = craft in iron)
   luogo_figurato (figurative location) (ex.: intervento in settore = intervention in field)
   etc.

5 Window "KB Filtro" shows an excerpt of coarse-grained knowledge of conceptual relations and the window "sint_sema Filtro" is an excerpt of the synt-sem table.

3. coarse-grained knowledge used to select the most plausible interpretation(s), expressed in Conceptual Graph notation:
   1. [ACTIVITY]→(LOCATION)→[PLACE]
   2. [CHANGE]→(FINAL_STATE)→[PRODUCT]
   3. [ARTIFACT]→(MATTER)→[MATTER]
   4. [ACTIVITY]→(FIGURATIVE_LOCATION)→[AMBIT]
4. generated fine-grained CRC (applying rule 1):
   [FARMING]→(LOCATION)→[GREENHOUSE]
5. generated average-grained CRC:
   [AGRICULTURAL_ACTIVITY]→(LOCATION)→[BUILDING_FOR_CULTIVATION]

When the algorithm runs the first several times, the linguist user inspects the fine-grained output, as shown in Figure 3, to verify a correct partition of the collocates among the two windows, and a correct interpretation of the semantically plausible syntactic collocates. Errors are used to refine the bias. The bias is now stable, at least for what concerns the 'known' words (about 5000 root-form words). In our domain, this took two or three round steps through the algorithm (i.e. run the algorithm, verify, and correct the bias), for each type of syntactic collocation. We believe that the (relatively) low semantic ambiguity of the domain sub-language and the availability of a well-defined set of conceptual relations contributed to the result.

After this first system training phase, the linguist only overviews the average grained CRCs, which must be tested before final acquisition in the knowledge base. Whether and how human intervention can still be reduced is unclear at the present stage of the experiment. Figure 4 shows a generalization session. Given one or more examples, such as *allevamento di pesce* (fish breeding) that are interpreted by the PATIENT (=animate direct object) relation, the system proposes to acquire the rule (shown with a reverse video in Figure 4):

Rule 1
[BREED]→(PTNT)→[ANIMAL_FOR_FEEDING]

The user can acquire the rule by clicking on the "Acquisisce" button, or test the rule before approval. The test is performed by showing the possible implications of that rule. These are obtained by listing all low-level CRCs with non-zero probability of occurrence in the corpus,[6] shown in the lower window of Figure 4.

In case of exceptions, it is the choice of the linguist to reject the rule (button "Rifiuta"), or to explicitly account for the exception(s) in a negative-example knowledge base, by marking with an "n" the triples that are not semantically correct. The rule shown in Figure 4 above generates only correct associations. The subtypes of BREED, e.g. DRAIN and DRILL, would in principle produce odd, if not totally unreasonable, associations, such as "fish training." Such associations, however, are not listed in the lower window of Figure 4 because the words "fish" and "drill" never occur in the corpus at a distance of ±5.

6 For very large corpora, as the one used here, producing the complete list of possible implications would generate in some case hundreds of examples, especially when both concepts in a CRC triple are not terminal nodes in the hierarchy. The linguist is therefore presented only with the list of collocates that have a non-zero probability of occurrence. This list is the list of all the collocations found in the (sub-)corpus using the ±5 algorithm of Smadja (1989).

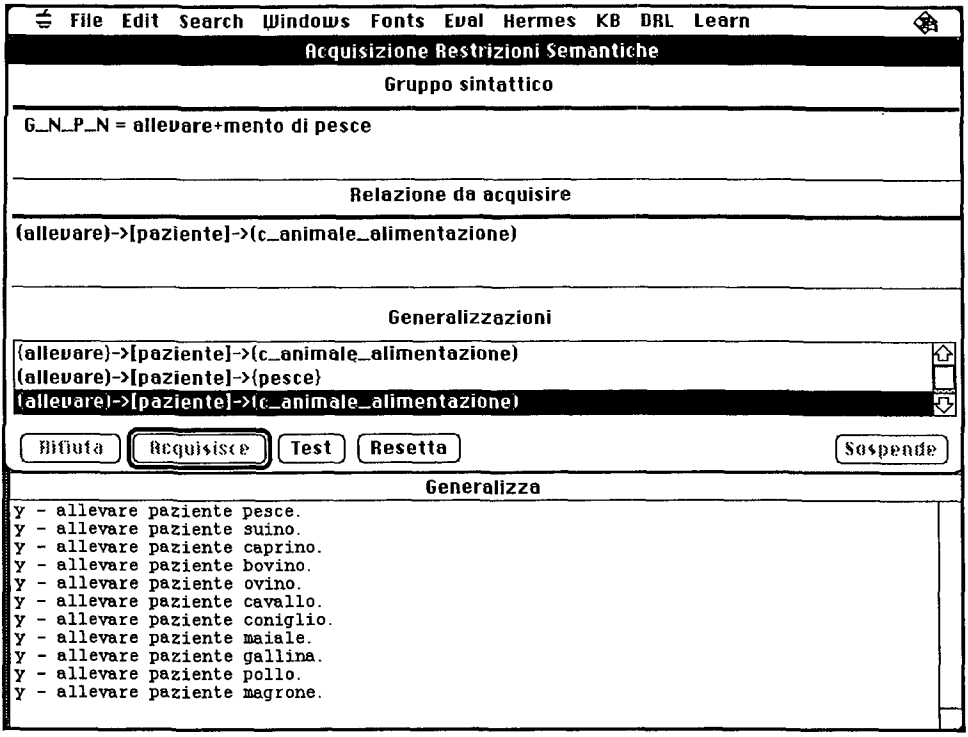

Figure 4
Example rule generalization session with the computer system

4.5 Discussion

This section describes the major problems and results of the acquisition algorithm.

4.5.1 Definition of the Semantic Bias. The most difficult task, both in the domain described in this paper and in the press agency releases domain described in Velardi (1989), is to define at the appropriate level of generality the selectional restrictions on conceptual relations. If restrictions are very coarse, they give rise to errors or (more frequently) to multiple interpretations at the end of step 2 in Section 4.3. If they are very selective, we get back to the case of a hand-encoded semantic lexicon.

In the first application (economy and finance) we used about 50 conceptual relations. The language domain was quite rich, and often it was not possible to express a selectional restriction with a single CRC triple; on the average, 1.5 CRCs per relation were necessary.

For example, the following are the selectional restrictions for the relation PARTICIPANT:

[HUMAN_ENTITY]←(PARTICIPANT)←[FLY],[SAIL],[MEETING],[AGREEMENT]

Examples of phrasal patterns interpreted by this relation are: John flies (to New York); an agreement between Fiat and Nissan; a contract among the companies, the assembly of the administrators, etc. (In the third example, notice that the word "contract" was classified in the economy domain both as an AGREEMENT and as a DOCUMENT.)

As for the agricultural enterprises, about 20 conceptual relations could provide enough expressive power to the semantic representation language. Most relations were defined by a single, coarse-grained CRC. Only a few required more detailed selectional restrictions, as for example the PART_OF relation:

[ANIMATE]→(PART_OF)→[BODY_PART]
[VEGETABLE]→(PART_OF)→[VEGETABLE_PART]
[CONCRETE_OBJ]→(PART_OF)→[CONCRETE_OBJ]

The major difficulty here is that a conceptual category is not always available to express a selectional restriction in a compact form. For example, the third restriction above is intended to capture patterns such as: "the pages of the book; the engine of the car; the top of the hill, etc." To express the relations among word couples such as (top,hill), (engine,car), (page,book) etc., a better CRC is:

[CONCRETE_OBJ]→(PART_OF)→[OBJECT_PART]

However, the type hierarchy defined for the agricultural domain does not have a class named OBJECT_PART, simply because this was not found useful, given the subworld lexicon. Hence, the third restriction looks very coarse, but turned out to be selective enough for the sub-language. Errors are of course possible, but remember that: first, the output of the acquisition process is supervised by a linguist for final approval (step 6 of Section 4.3); second, coarse-grained knowledge is used only for acquisition, not during the semantic interpretation of sentences. In other words, coarse knowledge is only a bias from which a more refined semantic knowledge is acquired.

4.5.2 Semantic Ambiguity. A second issue is semantic ambiguity. The system interprets word associations *outside a context*. This may give rise to several interpretations for a single syntactic collocate, even though the domain has very little lexical ambiguity.[7]

A good example of what could happen in a more general domain is the following: Consider the phrase "...run towards the bank...," that gives rise to the V_towards_N triple "run,towards,bank." Without a context, two CRCs are created: [RUN]→(DIR)→[RIVER_BANK] and [RUN]→(DIR)→[BANK_BUILDING]. The other interpretations, such as BANK_ORGANIZATION and BANK_ACTIVITY are rejected because their association in a "V_towards_N" position gives no plausible interpretation, if the semantic bias is "smart" enough. Now, the very fact that in the sentence from which the collocate was extracted "bank" was a river bank rather than a building simply does not matter. We are learning more than what the sentence was suggesting. In fact, we acquire two use types of the concepts RUN, RIVER_BANK and BANK_ACTIVITY (namely, that it is possible to run in the DIRection where a "*bank" is located), that are perfectly correct and can be used for semantic disambiguation. When a syntactic collocate is interpreted in a context, as during the semantic analysis of a sentence, the interpretation algorithm makes it possible to consider simultaneously

7 In fact, the word-to-concept table that maps words into concept names has fewer entries than the morphologic lexicon. For example, words that end with "zione" (tion) have the same concept type as the correspondent verb, e.g. production and produce. Other more complex examples of word to concept mappings are not mentioned for brevity. Ambiguous words are mostly those designating both an activity and a building where the activity takes place, as detailed in an example later on. But this is taken care of with a single metonymic rule (Lakoff 1987), rather than replicating the entries.

several restrictions (Velardi 1988); for example, the direct-indirect object relations, etc. It is in sentence interpretation that a full disambiguation is necessary, though not always possible.

4.5.3 Syntactic Disambiguation. The semantic knowledge base (SKB) acquired by our system is a large set of selectional restrictions on word uses, expressed by CRCs, that can be ordered either around concepts or around conceptual relations. The representation language is Conceptual Graphs, because we believe that this formalism has several advantages (Velardi 1988). However, standard logic could also be adopted. The majority of implemented NLP systems use selectional restrictions for syntactic and semantic disambiguation in a more or less standard way. We claim hence that the applicability of the algorithm presented in this paper *does not depend upon the specific algorithm used for semantic interpretation,* and the method can be adapted with minor changes to many NLP systems.

For the sake of completeness, we provide hereafter a brief summary of the semantic algorithm used in the DANTE system and in a system for information retrieval of agricultural businesses descriptions, focusing on the important problem of syntactic ambiguity. Details are given in the referenced papers.

The semantic interpreter proceeds bottom-up, in parallel with syntactic analysis. At the lower level nodes of the tree, it verifies whether it is possible to find the appropriate concepts and relations that interpret a given syntactic relation between words. While progressing up toward the root of the tree, it replaces syntactic relations between phrases by conceptual relations between partial Conceptual Graphs. It does this by following steps 1 and 2 of the acquisition algorithm (Section 4.3), but in step 2 rather than using selectional restrictions on conceptual relations (coarse-grained knowledge) it accesses the SKB. If at some point no interpretation is found, the system backtracks and selects a different syntactic attachment. For example, consider the following three sentences:

1. produrre vino in bottiglia (*to produce wine in bottle)
2. vendere uva all'ingrosso (*to sell grapes at wholesale)
3. produrre vino per i soci (*to supply wine for the shareholders).

All the above three phrases give rise to syntactic ambiguity, and precisely: 1:((V-N1)-prep-N2) or 2:(V-(N1-prep-N2)). For sentence 1 and tree 1, the interpreter first generates the graph: [PRODUCE]→(OBJ)→[WINE]. Then, a join is attempted between the head of the graph, [PRODUCE], with the rest of the sentence (in bottle). The preposition "in" corresponds to several conceptual relations (see the example in Section 4.4), but the selectional restrictions on concept uses available in the SKB do not suggest any valid interpretation. As no complete Conceptual Graph can be produced for tree 1, tree 2 is explored. Tree 2 generates first the graph:

[WINE]→(LOCATION)→[BOTTLE]

where the head concept is [WINE]. Then, a join is attempted between [PRODUCE] and [WINE]. This produces the final graph:

[PRODUCE]→(OBJ)→[WINE]→(LOCATION)→[BOTTLE].

Through a similar process, tree 1 is selected for sentence 2, that gives the graph:

[SALE] → (OBJ)→[GRAPE]
→ (MANNER)→[WHOLESALE]

In sentence 3, both trees are indeed plausible: the shareholders are the DESTINATION both of the wine and of the supply. In this case, for information retrieval purposes, it really does not matter which solution is preferred.

4.5.4 A First Evaluation. After some training and changes due to refinements in bias, the system was used to acquire the SKB for a prototype system for semantic codification of a test-bed set of agricultural texts, different from the one used for defining and testing the bias. The NLP system is described in Fasolo (1990) and is similar to the one described in Velardi (1988) and Antonacci (1989) and in the other papers on the DANTE system, except for the use of shallow methods in syntax and the ability to produce partial interpretations if parts of a text are not understood.

Many parts of this NLP system are still under development, but for what concerns the adequacy of the semantic knowledge base, the results are very encouraging. Currently, the test runs on about 3000 collocates, but these numbers keep changing. We plan to produce more accurate statistics after one full-year experimentation. The following is a very partial discussion of the results, useful for pinpointing the major problems.

1. About 10% of the syntactic collocates are rejected because one or both the words were unknown (no morphologic entry). Some of these are typos or abbreviations, and the others are really unknown. A major effort is being devoted toward an extension of the morphologic lexicon and the introduction of error correction algorithms. Recovery procedures such as the use of semantic information attached to word endings (e.g. "zione"="tion" always indicate an action, "x-ficio" is always a building for making some product x, etc.) can also be introduced.
2. Only about 1–2% correct syntactic collocates produce two interpretations where one is incorrect in whatever context. This is to some extent unavoidable, because the selectional restrictions on conceptual relations have obviously many exceptions, even though the probability of actually encountering such exceptions is low. It is unclear to what extent the presence of errors in the semantic knowledge base can induce errors in sentence interpretation, when multiple constraints are analyzed together. Currently, errors are purged by the linguist before final acquisition.
3. No incorrect syntactic collocates are erroneously given an interpretation.
4. Some collocates were rejected because they included pronouns. Anaphoric references are not handled in the current version of the NLP system.
5. One correct collocate was erroneously rejected showing the need for a new conceptual relation, FIGURATIVE_LOCATION (see Section 4.4).

There is no doubt that in order to fully validate this experiment we need to analyze thousands of texts, not hundreds.

This requires: first, that the morphologic lexicon be extended; second, that the NLP system that uses the acquired semantic knowledge be completed in several parts, primarily the question-answering module; third, and more importantly, that several linguistic issues be given more thorough solutions, in particular the definition and acquisition of conceptual categories.

5. Related Research

The method presented in this paper is at the frontier between machine learning and computational linguistics. It is closely related to two research areas: concept formation and lexical acquisition.

The similarities and differences of the problem discussed in this paper with that of concept formation are better understood by comparing our objectives with the formal tasks of concept formation (Fisher 1985; Gennari 1989):

1. Given: a sequential presentation of instances and their associated descriptions;
2. Find: a clustering of those instances in categories;
3. Find: an intensional definition for each category;
4. Find: a hierarchical organization for those categories.

Unlike the work on concept formation, instances (words in contexts) are not associated with descriptions (CRC triples), hence (1) is only in part a given. Conversely, (2) and (4) are a "given" rather than a "find." In summary:

1. Given: a sequential presentation of word associations;
2. Given: a many to many mapping from words to concept *types*;
3. Given: a hierarchical ordering of concept types;
4. Find: for each observation (two associated words), the concept types and the conceptual relation that interpret that observation (CRC);
5. Find: a definition for each concept that summarizes its instances (derived CRCs).

The more substantial difference is that the typical application dealt with in concept formation is clustering natural kinds (animals, biological species, etc.), whereas most of the terms we deal with are nominal kinds (verbs, etc.) and artifacts. Natural kinds are more easily described in terms of defining (internal) features, and their origin seem more basic to their "kindhood" (Keil 1989) than does the origin of nominal kinds. The latter are better described by their external relations with other objects. In other words, finding a type hierarchy is less basic but much more difficult for nominal kinds and artifacts than for natural kinds. This justifies the focus given to the derivation of concept descriptions rather than to conceptual clustering, even though we are aware that the hand-entering of the type hierarchy is the major limitation of our algorithm.

A second research area related to our work is lexical acquisition. An up-to-date survey of the most recent papers in this area is found in *Computational Linguistics* (1987) and Zernik (forthcoming). Many of the papers collected in the above two issues are more relevant to the fields of lexicography and cognition than to NLP. One of the

few lexical knowledge acquisition systems for NLP is described in Jacobs (1988) and Zernik (1989). When an unknown word is encountered, the system uses pre-existing knowledge of the context in which the word occurred to derive its conceptual category. The context is provided by on-line texts in the economic domain. For example, the unknown word *merger* in *"another merger offer"* is categorized as *merger-transaction* using semantic knowledge of the word *offer* and of pre-analyzed sentences referring to a previous offer event, as suggested by the word *another*. This method is interesting but allows a conceptual typing of unknown words only when everything else is known. It does not attack the problem of extensive lexical acquisition, but rather that of robust language processing.

In Binot (1987) prepositional attachments found in dictionary definitions are interpreted in terms of conceptual relations (e.g. WITH=INSTRUMENT) and used to solve syntactic ambiguity in parsing. The problem is that the information necessary to disambiguate is often not found, or requires several complex searches through the dictionary because of circularity and cross-references.

In Smadja (1989) the system EXTRACT, which uses shallow methods to extensively derive collocates from corpora, is described. The system produces a list of tuples (w1,w2,f), where w1 and w2 are two co-occurring words and f is the frequency of appearance in the corpus. No semantic interpretation is attempted, but it is claimed that mere co-occurrence knowledge can help language generators to correctly handle collocationally restricted sentences.

6. Concluding Remarks

The algorithm presented in this paper automatically detects co-occurrences in contexts and provides a semantic interpretation of the meaning relation between co-occurring words. Interpreted co-occurrences are used to build a semantic lexicon based on collocative meaning descriptions. The acquired concepts are syncategorematic; e.g., are completely defined by their pattern of use. It is assumed that such knowledge is sufficient to produce a surface semantic interpretation of raw text, i.e. a Conceptual Graph where content words and syntactic relations are replaced by the appropriate concept types and conceptual relations. This assumption indeed proved reasonable in our previous work on semantic interpretation, where the same type of semantic knowledge was hand-entered.

The observation of co-occurrences is language-dependent, context-dependent; the interpretation algorithm is (to some extent) language-independent, but relies upon human-derived primitives and relations that are in general context-dependent. No language model can prove to be objective, or even plausible. In principle, language rules and primitives do not exist. But even though symbols are arbitrary, their role is not to mimic human comprehension, but rather to produce some formal description of raw textual input, in a form that is ultimately useful for some relevant NLP application.

We feel that no human-invented semantic language will ever provide a full interpretation of language phenomena. We also strongly believe that more shallow methods such as the one discussed in this paper must be devised to give current NLP systems more breadth, as this will ultimately determine how widespread the use of NLP technology will be in the near future.

Acknowledgments

This work has been in part supported by the European Community, under grant PRO-ART 1989 and 1990, and in part by the CERVED.

References

Anderson, J. R. (1989). A Theory of the Origins of Human Knowledge. *Artificial Intelligence*. vol. 31–40, September.

Antonacci, Pazienza, and Russo, Velardi (1989). A System for Text Analysis and Lexical Knowledge Acquisition. *Data and Knowledge Engineering*. n. 4.

Binot, J. L., and Jensen, K. (1987). A Semantic Expert Using an On-line Standard Dictionary. *Proceedings of the IJCAI*. Milano.

Boguraev, B.; Byrd, R.; Klavans, J.; and Neff, M. (1989). From Machine Readable Dictionaries to Lexical Databases. *First Lexical Acquisition Workshop*, Zernik ed., Detroit.

Byrd, R.; Calzolari, N.; Chodorow, M.; Klavans, J.; Neff, M.; Rizk, O. (1987). Tools and Methods for Computational Lexicography. *Computational Linguistics*.

Computational Linguistics special issue on the Lexicon, Walker, D., Zampolli, A., and Calzolari, N., eds. July–December, 1987.

Dahlgren, K., and McDowell, J. (1989). Knowledge Representation for Commonsense Reasoning with Texts. *Computational Linguistics*, vol. 15, n. 3, September.

Evens, M. (ed.) (1988). *Relational Models of the Lexicon*. Cambridge University Press.

Fasolo, M.; Garbuio, L.; and Guarino, N. (1990). Comprensione di descrizioni di attivita' economico-produttive espresse in linguaggio naturale. GULP Conference on Logic Programming. Padova.

Fisher, D., and Langley, P. (1986). Conceptual Clustering and Its Relation to Numerical Taxonomy. *Artificial Intelligence and Statistics*. Addison-Wesley.

Gennari, J.; Langley, P.; and Fisher, D. (1989). Model of Incremental Concept Formation. *Artificial Intelligence*, vol. 31–40, September.

Jacobs, P. (1987). A Knowledge Framework for Natural Language Analysis, In *Proceedings of IJCAI87*. Milano, August.

Jacobs, P. (1988). Making Sense of Lexical Acquisition, In *Proceedings of AAAI88*. St. Paul, August.

Keil, F. (1989). Concepts, Kinds and Cognitive Development, In *Data and Knowledge Engineering*. The MIT Press.

Lakoff, G. (1987). *Woman, Fire and Dangerous Things*. The University of Chicago Press.

Leech, Geoffrey. (1981). *Semantics: The Study of Meaning*, Second edition, Penguin Books.

McKeown, K. (1985). *Text Generation*. Cambridge University Press.

Michalski, R. S.; Carbonell, J. C.; and Mitchell, T. M. (1983). *Machine Learning Vol I*. Tioga Publishing Company, Palo Alto.

Melčuk, I., and Polguere, A. (1987). A Formal Lexicon in Meaning-Text Theory (or How To Do Lexica with Words). *Computational Linguistics* **13**(3–4): 261–275.

Niremburg, S., and Raskin, V. (1987). The Subworld Concept Lexicon and the Lexicon Management System. *Computational Linguistics*.

Pazienza, M. T., and Velardi, P. (1988). Using a Semantic Knowledge Base to Support A Natural Language Interface to a Text Database. *7th International Conference on Entity-Relationship Approach*, Rome, November 16–18.

Rosch, E. (1975). Cognitive Representation of Semantic Categories. *Journal of Experimental Psychology* 104, 192–233.

Russo, M. (1987). A Generative Grammar Approach for the Morphologic and Morphosyntactic Analysis of Italian, In *Third Conference of the European Chapter of the ACL*, Copenhagen, April 1–3.

Schank, R. C. (1972). Conceptual Dependency: A Theory of Natural Language Understanding, In *Cognitive Psychology*, vol. 3.

Sowa, J. F. (1984). *Conceptual Structures in Mind and Machine*. Addison-Wesley.

Smadja, F. (1989). Lexical Co-occurrence: The Missing Link, In *Literary and Linguistic Computing*, Vol. 4, n. 3.

Velardi, P.; Pazienza, M. T.; and De Giovanetti, M. (1988). Conceptual Graphs for the Analysis and Generation of Sentences, In *IBM Journal of R&D*, special issue on language processing, March.

Velardi, P., and Pazienza, M. T. (1989). Computer-Aided Acquisition of Lexical Cooccurrences, In *Proceedings of the ACL 1989*. Vancouver.

Velardi, P. (1990). Why Human Translators Still Sleep in Peace? (Four Engineering and Linguistic Gaps in NLP), In *Proc. of COLING 90*. Helsinki, August.

Webster, M., and Marcus, M. (1989). Automatic Acquisition of the Lexical Semantics of Verbs from Sentence Frames. In *Proceedings of the ACL 1989*. Vancouver.

Zernik, U. (1989). Lexicon Acquisition: Learning from Corpus by Capitalizing on Lexical Categories, In *IJCAI 1989*, Detroit.

Zernik, U., ed. (in press) *Lexical Acquisition: Using On-line Resources to Build a Lexicon*, Lawrence Erlbaum.