Computational Lexicons: the Neat Examples and the Odd Exemplars

Roberto Basili , Maria Teresa Pazienza
Dip. di Ingegneria Elettronica, Universita'
"Tor Vergata", Roma, Italy

Paola Velardi
Ist. di Informatica, Universita' di Ancona,
Ancona, Italy

Abstract

When implementing computational lexicons it is important to keep in mind the texts that a NLP system must deal with. Words relate to each other in many different, often queer, ways: this information is rarely found in dictionaries, and it is quite hard to be invented a priori, despite the imagination that linguists exhibit at inventing esoteric examples.

In this paper we present the results of an experiment in learning from corpora the frequent selectional restrictions holding between content words. The method is based on the analysis of word associations augmented with syntactic markers and semantic tags. Word pairs are extracted by a morphosyntactic analyzer and clustered according to their semantic tags. A statistical measure is applied to the data to evaluate the significance of a detected relation.

Clustered association data render the study of word associations more interesting with several respects: data are more reliable even for smaller corpora, more easy to interpret, and have many practical applications in NLP.

1. Introduction

One of the fundamental property of computational lexicons is an account of the relations between verbs and its arguments. Arguments are identified by their position in a predicate-argument structure, or by conceptual relations names (e.g. *agent*, *purpose*, *location*, etc). Arguments are annotated with *selectional restrictions*, that impose type constraints on the set of content words that may fill a relation. Selectional restrictions often do not provide all the semantic information that is necessary in NLP systems, however they are at the basis of the majority of computational approaches to syntactic and semantic disambiguation.

It has been noticed that representing only the semantics of verbs may be inadequate (Velardi *et al.* 1988; Boguraev 1991; Macpherson 1991). The notion of *spreading the semantic load* supports the idea that every content word should be represented in the lexicon as the union of all the situations in which it could potentially participate. Unfortunately, hand writing selectional restrictions is not an easy matter, because it is time consuming and it is hard to keep consistency among the data when the lexicon has several hundred or thousand words. However the major difficulty is that words relate to each other in many different, often domain dependent ways. The nowadays vast literature on computational lexicons is filled with neat examples of the *eat(animate,food)* flavour, but in practice in many language domains selectional constraints between words are quite odd. It is not just a matter of violating the semantic expectations, such as in "*kill the process*" or "*my car drinks gasoline*", neither it is that kind of fancifulness that linguists exhibit at finding queer sentences. Rather, there exist statistically relevant linguistic relations that are hard to imagine a-priori, almost never found in dictionaries, and even harder to assign to the appropriate slot in the whatever conceptual structure adopted for lexical representation. Several examples of such relations are shown throughout this paper.

Ideally, knowledge on word relations should be acquired directly from massive amounts of texts, rather than from hand-crafted rules. This idea is at the basis of many recent studies on word associations. The results of these studies have important applications in lexicography, to detect lexico-syntactic regularities (Church and Hanks, 1990) (Calzolari and Bindi,1990), such as, for example, support verbs (e.g. "make-decision") prepositional verbs (e.g. "rely-upon") idioms, semantic relations (e.g. "part_of") and fixed expressions (e.g. "kick the bucket"). In (Hindle,1990; Zernik, 1989; Webster et Marcus, 1989) cooccurrence analyses augmented with syntactic parsing is used for the purpose of word classification. All these studies are based on the (strong) assumption that syntactic similarity in word patterns implies semantic similarity. In (Guthrie *et al.*, 1991), sets of consistently contiguous words ("neighbourhood") are extracted from machine-readable dictionaries, to help semantic disambiguation in information retrieval. In (Smadja and McKeown, 1990) statistically collected associations provide pragmatic cues for lexical choice in sentence generation. For example, we can learn that "make decision" is a better choice than, say "have decision" or "take decision". (Hindle and Rooths, 1991) proposes that a syntactic disambiguation criterion can be gathered by comparing the probability of occurrence of noun-preposition and verb-preposition pairs in V NP PP structures.

In general word associations are collected by extracting word pairs in a +-5 window. In (Calzolari and Bindi, 1990), (Church and Hanks, 1990) the significance of an association (x,y) is measured by the mutual information I(x,y), i.e. the probability of observing x and y together, compared with the probability of observing x and y independently. In (Smadja, 1989), (Zernik and Jacobs, 1990), the associations are filtered by selecting the word pairs (x,y) whose frequency of occurrence is above f+ks, where f is the average appearance, s is the standard deviation, and k is an empirically determined factor. (Hindle, 1990; Hindle and Rooths,1991) and (Smadja, 1991) use syntactic markers to increase the significance of the data. (Guthrie *et al.*, 1991] uses the subject classification given in machine-readable dictionaries (e.g. *economics, engineering*, etc.) to reinforce cooccurence links.

Despite the use of these methods to add evidence to the data, the major problem with word-pairs collections is that reliable results are obtained only for a small subset of high-frequency words on very large corpora, otherwise the association ratio becomes unstable. For example, Church run his experiment on a corpus with over 20-30 millions words, and Hindle reports 6 millions words as not being an adequate corpus. In many practical NLP/IR applications corpora are not so large, and typically span from 500,000 to a few million words. The analysis of associations could be done on wider domains, but a part for very general words, it is much more desirable to collect data from the application corpus. Information collected from other sources could add noise rather than strengthening the data, because in most applications jargon, technical words, and domain-dependent associations are the norm. In (Smadja, 1989b) it is shown a table of operational pairs like adjective-noun and verb-object, from which clearly emerges the very different nature of the two source domains (Unix Usenet and Jerusalem Post). For example, the noun-noun pairs with "tree" include associations such as "parse, grammar, decision" and "olive, Christmas". If the NLP/IR application is about the computer world, associations such as "olive tree" or "Christmas tree" are (at best) useless.

A second problem with statistically collected word pairs is that an analysis based simply on surface distribution may produce data at a level of granularity too fine. For example, a purely distributional analysis for word classification, such as those cited above, might place two verbs into distinct classes because one is used primarily with an object *olive* and the other with the object *grape*. This may not be appropriate given the application. Abstraction via semantic classes (e.g. VEGETABLE), would ensure that the ontology found is appropriate for the domain. The model of prepositional attachment preference proposed by Hindle is also too weak if applied only to verb-preposition and noun-preposition pairs. A preposition may or may not be related to a verb, even if it frequently cooccurs with it, depending upon the underlying semantic relation. It is the semantic category of the noun following a preposition that determines the nature of the semantic link (e.g. for+ HUMAN_ENTITY = *beneficiary*, for+ACTION = *purpose*), and ultimately influences the choice of the proper attachment. Semantic abstraction also renders the data more readable. Millions of simple word cooccurrences let the experimenter sink in an ocean of data, without providing much insight of the conceptual nature of the detected associations.

In this paper, we present a study on word associations augmented with syntactic markers and semantic tagging. We call these data **clustered associations.** Clustered association data are syntactic pairs or triples (e.g. N_V(*John,go*) V_prep_N(*go,to,Boston*), N_prep_N(*Boston,by,bus*)[1]) in which one or both content words are replaced by their semantic tag (e.g. V_prep_N(PHYSICAL_ACT-to-PLACE),N_prep_N(PLACE-by-MACHINE) etc.). Semantic tags are very high-level in order to reduce the cost of hand-tagging.

Clustered association data have several advantages:

- First, statistically meaningful data can be gathered from (relatively) small corpora;
- Second, data are presented in a compact form and are much more readable;
- Third, clustered association data are useful for many interesting NLP applications, such as conceptual clustering, syntactic and semantic disambiguation, and semi-automatic learning of the relevant selectional restrictions in a given language domain.

In this paper we discuss the results of an experiment in learning selectional restrictions, to provide support for the design of computational lexicons. Other results are presented in (Basili *et al.*, 1991; Fabrizi *et al.*, forthcoming].

[1] We did not want to schock the reader with queer examples since the introduction.

The method is applied to a corpus of economic enterprise descriptions, registered at the Chambers of Commerce in Italy. The database of these descriptions (in total over 1,000,000 descriptions, each spanning from 1 to 100-200 words) is managed in Italy by the Company CERVED. Sentences describe one or several commercial enterprises carried out by a given Company. Examples of these descriptions are provided throughout the text. In our experiment, we used only 25,000 descriptions, including about 500,000 words. A second experiment on a legal corpus is under preparation and will be ready shortly.

2 Acquiring syntactic associations

Clustered association data are collected by first extracting from the corpus all the syntactically related word pairs.

Combining statistical and parsing methods has been done by (Hindle, 1990; Hindle and Rooths,1991) and (Smadja and McKewon, 1990; Smadja,1991). The novel aspect of our study is that we collect not only operational pairs, but triples, such as N_prep_N, V_prep_N etc. In fact, the preposition convey important information on the nature of the semantic link between syntactically related content words. By looking at the preposition, it is possible to restrict the set of semantic relations underlying a syntactic relation (e.g. *for=purpose,beneficiary*).

To extract syntactic associations two methods have been adopted in the literature. Smadja attempts to apply syntactic information to a set of automatically collected collocations (statistics-first). Hindle performs syntactic parsing before collocational analysis (syntax-first). In our study, we decided to adopt the syntax-first approach, because:

- as remarked above, it is important to extract not only syntactic pairs, but also triples;
- statistically collected associations miss some syntactic relation between distant words in coordinate constructions (usually the window in which word pairs are extracted is +-5) and couple many words that are not syntactically related. Even though (Smadja,1991) reports good performances of his system, it must be noticed that the precision and efficiency figures of the parser apply to a set of data that have been already (statistically) processed. Thus the actual precision and efficiency in extracting syntactically related words from the source corpus may be lower than expected.

As in other similar works, the syntactic analyzer used in this study does not rely on a complete Italian grammar. The parser only detects the **surface syntactic relations** between words. A full description of the analyzer is outside the scope of this paper (see (Marziali, 1991) for details). In short, the parser consists of a segmentation algorithm to cut texts into phrases (NP, PP, VP etc), and a phrase parser that is able to detect the following 15 links: N_V, V_N, N_ADJ, N_N, N_prep_N, V_prep_N, N_prep_V, V_prep_V, N_cong_N, ADJ_cong_ADJ, V_ADV, ADV_cong_ADV, V_cong_V, N_prep_ADJ, V_prep_ADJ.

The segmentation algorithm is very simple. If the domain sublanguage is good Italian, sentence cutting is based on the presence of verbs, punctuation, adverbs such as *when, if, because,* etc. For more jergal domains, such as the economic enterprise corpus, text cutting is based on heuristics such as the detection of a word classified as "activity" (Fasolo *et al.*,1990). In fact, this domain is characterized by absence of punctuation, ill formed sentences, long nested coordinate constructions.

The phrase parser is based on DCG (Pereira and Warren,1980), the most complex part of which is the treatment of coordination. The grammar consists of about 20 rules. Rather than a parse tree, the output is a "flat" set of syntactic relations between content words. For example, parsing the sentence:

fabbrica di scarpe per uomo e per bambino
*(*manufacture of shoes for man and child)*
produces the following relations:

N_prep_N(fabbrica,di,scarpe)
N_prep_N(fabbrica,per,uomo)
N_prep_N(fabbrica,per,bambino)
N_prep_N(scarpe,per,uomo)
N_prep_N(scarpe,per,bambino)
N_cong_N(uomo,e,bambino)

Unlike Church and Hindle, we are not interested in collecting binary or ternary relations between words *within a sentence*, but rather in detecting recurring binary syntactic associations in the corpus. For this purpose it is unnecessary to retrieve even partial parse trees.

The complexity of the grammar is O(n^2), that makes it computationally attractive for parsing large corpora. In (Marziali,1991) the efficiency and precision of this grammar with respect to the full set of surface syntactic links detectable by a complete DCG grammar are evaluated to be 85% and 90%. The reference output adopted to perform the evaluation is a *syntactic graph* (Seo and Simmons,1989). Syntactic graphs include in a unique graph the set of all possible parse trees. The evaluation was hand-made over a set of 100 sentences belonging to three domains: the economic corpus, the legal corpus, and a novel. The performances are better for the legal corpus and the novel, due to the ungrammaticality of the economic corpus.

The relatively high efficiency rate, as compared with the figures reported in (Brent, 1991), are due to the fact that Italian morphology is far more complex than English. Once a good morphologic analyzer is available (the one used in our work is very well tested, and has first described in (Russo,1987)), problems such as verb detection, raised in (Brent, 1991), are negligible. In addition, the text-cutting algorithm has positive effects on the precision.

Despite this, we verified that about a 35% of the syntactic associations extracted from the economic corpus are semantically unrelated, due to syntactic ambiguity. As shown in the following sections, semantic clustering in part solves this problem, because semantically unrelated word pairs do not accumulate statistical relevance, except for very rare and unfortunate cases.

In any case, we need more experiments to verify the effect of a more severe sentence cutting algorithm on the precision at detecting semantically related pairs. This issue is particularly relevant for ungrammatical texts, as in the economic corpus.

3. Assigning semantic tags

The set of syntactic associations extracted by the DCG parser are first clustered according to the cooccurring words **and** the type of syntactic link. A further clustering is performed based on the semantic tag associated to the cooccurring words.

Clustering association data through semantic tagging has two important advantages:

First, it improves significantly the reliability of association data, even for small corpora;

Second, and more importantly, semantic tags make it explicit the **semantic nature** of word relations.

Manually adding semantic tags to words may appear very expensive, but in fact it is not, if very broad, domain-dependent classes are selected.

In our application, the following 13 categories were adopted:

PHYSICAL_ACT (packaging, travel, build, etc.)
MENTAL_ACT(sell, organize, handle, teach, etc.)
HUMAN_ENTITY (shareholder, company, person, farmer, tailor, etc.)
ANIMAL (cow, sheep, etc.)
VEGETABLE (carrots, grape, rubber, coffee, etc.)
MATERIAL (wood, iron, water, cement, etc.)
BUILDING (mill, shop, house, grocery, etc.)
BY_PRODUCT (jam, milk, wine, drink, hide, etc.)
ARTIFACT (item, brickwork, toy, table, wears, etc.)
MACHINE (engine,tractor,grindstone,computer, etc.)
PLACE (ground, field, territory, Italy, sea, etc.)
QUALITY (green, chemical, coaxial, flexible, etc.)
MANNER (chemically, by-hand, retail, etc.)

These categories classify well enough the words which are found in the selected sub-corpus as a test-bed for our research. Some words received two tags: for example, there are sentences in which a BUILDING metonymically refers to the commercial ACT held in that building (e.g. *"trade mills for the production.."*); some word is both a BY_PRODUCT (e.g. *"wood carving"*) or a MATERIAL (e.g. *"handicrafts in wood"*). Because the domain is very specific, double-tagging is never due to polisemy.

Once the definition of a semantic class is clearly stated, and with the help of a simple user interface, hand tagging a word *is a matter of seconds.* We adopted the policy that, whenever assigning a tag is not obvious, or none of the available tags seems adequate, the word is simply skipped. Unclassified words are less than 10% in our corpus. Overall, we classified over 5000 words (lemmata). The activity of classification was absolutely negligible in comparison with all the other activities in this research, both on the ground of time and required skill.

Domain-dependent tags render the classification task more simple and ensure that the clustered association data are *appropriate* given the application. An obvious drawback is that it is necessary to re-classify many words if the application domain changes significantly. For example, we are currently preparing a new experiment on a legal corpus. The domain is semantically more rich, hence we needed 15 classes. A first estimate revealed that about 30-40% of the words need to be re-classified using more appropriate semantic tags.

4 Acquisition of selectional restrictions

Clustered association data are at the basis of our method to detect the important selectional restrictions that hold in a given sublanguage. The statistical significance of the detected relations is measured by the probability of cooccurrence of two classes C_1 and C_2 in the pattern C_1 *synt-rel* C_2, where *synt-rel* is one of the syntactic relations detectable by the parser summarized in Section 2.

Rather than evaluating the probability Pr(C_1 synt-rel C_2), we computed the **conditioned probability** P(C_1,C_2/synt-rel) estimated by:

$$(1) \qquad \frac{f(C_1,synt\_rel,C_2)}{f(synt\_rel)}$$

The reason for using (1) rather than other measures proposed in the literature, is that what matters here is to detect all the statistically relevant phenomena, not necessarily all the meaningful associations. Such measures as the *mutual information* and the *t-score* (Church *et al.*, 1991) give emphasis to the infrequent phenomena, because the statistical significance of the coupling between C_1 and C_2 is related to the probability of occurrence of C_1 and C_2 independently from each other. This would be useful at detecting rare but meaningful relations if one could rely on the correctness of the data. Unfortunately, due to syntactic ambiguity and errors in parsing, many syntactic associations **are not semantically related**, i.e. there exists no plausible selectional relations between the to cooccurring words. Using the mutual information, such relations could accumulate statistical evidence. The (1) is more conservative, but ensures more reliable results. In any case, we run several experiments with different statistical measures, without being entirely happy with any of these. Finding more appropriate statistical methods is one of the future objectives of our work.

Clustered association data are used to build tables, one for each syntactic structure, whose element (x,y) represents the statistical significance in the corpus of a concept pair C_1 C_2. All the relevant couplings among classes are identified by a human operator who inspects the tables, and labels concept pairs by the appropriate conceptual relation. Finding an appropriate set of conceptual relations is not an easy task. In labeling concept pairs, we relied on our preceding work on semantic representation with Conceptual Graph [Pazienza and Velardi, 1987].

Four of these tables are presented in the Appendix. The data have been collected from a corpus of about 500,000 words. The morphosyntactic analyzer takes about 6-10 hours on a Spark station. Clustering the data takes about as much. At first, we extracted only the V_N, N_prep_N and V_prep_N associations, for a total of 52,155 *different* syntactic associations. The average is 5 occurrences for each association.

At first glance, the data seem quite odd even to an Italian reader, but it turns out that the tables show exactly what the corpus includes. Let us briefly go through the 4 tables.

Table 1 summarizes the relations C_1-per-C_2 (*per=for*). Some of the significant associations are:

ARTIFACT - PHYSICAL_ACT (e.g.: *articoli per lo sport (*items for sport), attrezzi per giardinaggio (*tools for gardening))*

ARTIFACT - BUILDING (e.g. *biancheria per la casa (*linens for the house), mobili per negozi (*furnitures for shops))*

MACHINE-BUILDING (e.g. *macchinari per laboratori (*equipments for laboratories), macine per mulini (*grindstones for mills)*)

All the above relations subsume the *usage* (or *figurative_destination*) relation.

Notice that the "advertised" *beneficiary* relation is not very significant in the corpus. The only statistically relevant *beneficiary* relations are ARTIFACT-for-HUMAN_ENTITY ((e.g. *calzature per uomo (*shoes for man), biancheria per signora (*linens for lady)*) and HUMAN_ENTITY_for_HUMAN_ENTITY (e.g. *parrucchire per signora (*hairdresser for lady*). It appears that in the considered domain, verbs, except for some, poorly relate with the preposition *for* (this is the first surprise!).

Table 2 shows the C_1-in-C_2 relations. Two relations represent the large majority:

ARTIFACT-in- BY_PRODUCT (e.g. *calzature in pelle (*shoes in leather), guarnizioni in gomma (packings in rubber))*

ARTIFACT-in-MATERIAL (e.g. *oggetti in legnc (*handicrafts in wood) ringhiere in ferro (*banisters in iron)*)

both subsume a *matter* relation (this is one of the few "expected" associations we found).

Less frequent but interesting are the following relations:

MENTAL_ACT-in-MENTAL_ACT (e.g. *concedere ir appalto (*to grant in bid) acquistare in leasing (*tc buy in leasing)*)

ARTIFACT-in-ARTIFACT (e.g. *prodotti in scatolc (*products in can)*)

While the second is a "reassuring" *location* relatior (subsumed also by the in-PLACE associations in the last column), we are less sure about the semantic interpretation of the first relation. Tentatively, we choosed the *manner* relation. The same type o relation is also found in the C_1-a-C_2 (*a=to,on*) table (Table 3): MENTAL_ACT-a-MENTAL_ACT (e.g *acquistare a credito (*to buy to (=on) credit) abilitare all'ottenimento (*qualifying to th attainment), assistenza all'insegnamento (assistenc to the teaching)*)

The first example (*on credit*) clearly subsumes the same relation as for *in leasing;* the following two seem of a different nature. We used *figurative-destination* to label these relations. This may or may not be the best interpretation: however, what matters here is not so much the human interpretation of the data, but rather the ability of the system at detecting the relevant semantic associations, whatever their name could be.

Once again in this table, notice that the common relations, like *recipient* (to-HUMAN_ENTITY) and *destination* (to-PLACE) are less frequent than expected.

Table 4 shows the C_1-da-C_2 relations (da=from,for,to). The most frequent relations are:

MATERIAL-da-PHYSICAL_ACT (e.g *materiali da imballaggio (*material from (=for) packing) legna da ardere (*wood from (=to) burn)*)

ARTIFACT-da-ARTIFACT (e.g *cera da pavimenti (*wax for floors), lenzuola da letto (*sheets for bed)*)

ARTIFACT-da-PLACE (e.g *giocattoli da mare (*toys for sea) abiti da montagna (*wears for mountain)*

MENTAL_ACT-da-BUILDING (e.g *acquistare da fabbrica (*to buy from firms), comprare da oleifici (*to buy from oil-mills)*)

The first three relations, very frequent in the corpus, all subsume the *usage* relation. It is interesting to notice that in Italian "*da*+PLACE" commonly subsumes a *source* relation, just as in English (*from*+PLACE). The table however shows that this is not the case, at least when "from-PLACE" cooccurs with classes such as ARTIFACT and MATERIAL. The classical *source* sense is found in the fourth example. BUILDINGs here metonymically refer to the human organization that manages an activity in the building.

Currently we are unable to analyze the preposition *di* (*of*), because in the corpus it is used for the large majority to subsume the direct-object syntactic relation (e.g. *vendita di frutta *sale of fruit*). It turns out that the distribution of C_1-di-C_2 is too even to allow an analysis of the data. Perhaps a less crude parsing strategy could help at ruling out these associations.

A new domain is now under examination (a legal corpus on taxation norms). A first analysis of the data shows that even for this corpus, despite it is much less jergal, several unconventional relations hold between content words.

5. Final Remarks

We spent some time at illustrating the tables to make the reader more confident with the data, and to show with several practical examples the thesis of this paper, i.e. that selectional restrictions are more fanciful than what usually appears from the literature on computational lexicons (and from dictionaries as well). The reader should not think that we selected for our application the oddest domain we could find: similar (as for fancifulness) data are being extracted from a legal corpus which is currently under examination.

The (semi-)automatic acquisition of selectional restrictions is only one ot the things that can be learned using clustered association data. In a forthcoming paper (Basili *et al.*, forthcoming) the same data, clustered only by the right-hand word, are at the basis of a very reliable algorithm for syntactic disambiguation. We are also experimenting concept formation algorithms for verb and noun classification (Fabrizi *et al.*, forthcoming).

In summary, clustered associations in our view greatly improve the reliability and applicability of studies on word associations. More work is necessary, because of semantic tagging, but there is an evident payoff. In any case, semantic tagging is not at all the most painful manual activity in association studies.

Acknowledgements.

We thank CERVEDomani for making us available the corpus of enterprise descriptions. This work has been in part supported by the European Community (PRO-ART and NOMOS ESPRIT 1991 n. 5330).

References

ACL 1990, Proceedings of ACL '90, Pittsburgh, Pennsylvania, 1990.

ACL 1991, Proceedings of ACL '91, Berkley, California, 1991

R. Basili, M. T. Pazienza, P. Velardi, Using word association for syntactic disambiguation, 2nd. Congress of the Italian Association for Artificial Intelligence, Palermo, 1991

B. Boguraev, Building a Lexicon: the Contribution of Computers, IBM Report, T.J. Watson Research Center, 1991

M. Brent, Automatic Aquisition of Subcategorization frames from Untagged Texts, in (ACL, 1991)

N. Calzolari, R. Bindi, Acquisition of Lexical Information from Corpus, in (COL,1990)

K. W. Church, P. Hanks, Word Association Norms, Mutual Information, and Lexicography, Computational Linguistics, vol. 16, n. 1, March 1990.

K. Church, W. Gale, P. Hanks, D. Hindle, Using Statistics in Lexical Analysis, in (Zernik, 1991).

S. Fabrizi, M.T.Pazienza, P. Velardi, A corpus-driven clustering algorithm for the acquisition of word ontologies, forthcoming.

M.Fasolo, L.Garbuio, N.Guarino, Comprensione di descrizioni di attivita' economico-produttive espresse in linguaggio naturale, Proc. of GULP Conference, Padova 1990.

J. Guthrie, L. Guthrie, Y. Wilks, H. Aidinejad, Subject-dependent Co-occurrence and Word Sense Disambiguation, in (ACL, 1991).

D. Hindle, Noun classification from predicate argument structures, in (ACL ,1990).

D. Hindle, M. Rooths, Structural Ambiguity and Lexical Relations, in (ACL, 1991).

"Lexical Semantics and Knowledge Representation" Proc. of a workshop sponsored by the Special Interest Group on the Lexicon of the ACL, Ed. J. Pustejovsky, June 1991

M. Macpherson, Redefining the Level of the Word, in (Lexical, 1991).

A. Marziali, "Laurea" dissertation, University of Roma II, Dept. of Electrical Engineering, in preparation

M.T. Pazienza, P. Velardi, A structured Representation of Word Senses for Semantic Analysis, Third Conference of the European Chapter of ACL, Copenhagen, April 1-3, 1987.

F. Pereira, D. Warren, Definite Clause Grammars for Language Analysis - A Survey of the Formalism and a Comparison with Augmented Transition Networks, in Artificial Intelligence, n. 13, 1980.

M. Russo, A generative grammar approach for the morphologic and morphosyntactic analysis of the Italian language, 3rd. Conf. of the European Chapter of the ACL, Copenhaghen, April 1-3 1987

J. Seo, R.F. Simmons, Syntactic Graphs a Representation for the Union of All Ambigous Parse Trees, Computational Linguistics, Vol. 15, n.1, March, 1989.

F. A. Smadja, Lexical Co-occurrence The Missing Link, Literary and Linguistic Computing, vol. 4, n.3, 1989.

F. Smadja, Macrocoding the Lexicon with Co-occurrence Knowledge, First Lexical Acquisition Workshop, August 1989, Detroit, and in (Zernik,1991).

F. Smadja, K. McKewon, Automatically extracting and representing collocations for language generation, in (ACL,1990).

F. Smadja, From N-Grams to Collocations an evaluation of XTRACT, in (ACL,1991).

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

M. Webster M. Marcus, Automatic Acquisition of lexical semantics of verbs from sentence frames, Proc. of ACL89, Vancouver 1989

U. Zernik, Lexical acquisition Learning from Corpus by capitalizing on Lexical categories, Proc. of IJCAI 1989, Detroit 1989

U. Zernik, P. Jacobs, Tagging for Learning Collecting Thematic Relations from Corpus, Proc. of COLING 90, Helsinki, August 1990.

U. Zernik ed. "Lexical Acquisition Using On-line Resources to Build a Lexicon", Lawrence Erlbaum Ass.,1991

Appendix: Examples of acquired conceptual associations

per	1)	2)	3)	4)	5)	6)	7)	8)	9)	10)	11)
1) att_mat	0.121	0.075	0.031	0.022	0.000	0.039	0.002	0.002	0.001	0.010	0.004
2) att_ment	0.041	0.054	0.023	0.018	0.000	0.023	0.001	0.000	-	0.005	0.003
3) manufatto	0.094	0.068	0.045	0.026	0.001	0.057	0.005	0.005	0.000	0.011	0.004
4) entita_umana	0.016	0.024	0.002	0.024	0.000	0.014	0.000	0.001	-	0.001	0.001
5) vegetale	0.003	0.001	0.002	0.000	-	0.000	-	-	-	0.001	0.000
6) costruzione	0.061	0.030	0.012	0.010	-	0.013	0.000	0.000	-	0.001	0.001
7) derivato	0.013	0.004	0.006	0.001	0.000	0.004	0.001	0.001	0.001	0.002	0.001
8) materiale	0.027	0.008	0.011	0.001	0.000	0.015	0.002	0.002	-	0.003	0.001
9) animali	0.000	0.001	-	-	-	0.000	-	-	-	-	0.000
10) macchinario	0.032	0.036	0.010	0.002	0.001	0.033	0.001	0.001	-	0.004	0.001
11) luoghi	0.004	0.004	0.000	0.000	-	0.001	-	0.000	-	-	0.000

Table 1: C_1-per-C_2

in	1)	2)	3)	4)	5)	6)	7)	8)	9)	10)	11)
1) att_mat	0.019	0.048	0.023	0.013	0.002	0.013	0.041	0.085	-	0.000	0.053
2) att_ment	0.033	0.088	0.038	0.050	0.001	0.007	0.016	0.045	-	0.004	0.074
3) manufatto	0.014	0.006	0.066	0.001	0.007	0.007	0.092	0.121	-	0.001	0.009
4) entita_umana	0.004	0.009	0.000	-	0.002	0.001	0.004	0.009	-	-	0.010
5) vegetale	-	-	0.010	-	0.001	0.009	0.006	0.000	-	-	-
6) costruzione	0.008	0.013	0.017	0.000	-	0.004	0.009	0.035	-	0.001	0.020
7) derivato	0.001	0.001	0.006	-	0.001	0.000	0.005	0.005	-	-	0.004
8) materiale	0.002	-	0.008	-	-	0.001	0.002	0.005	-	-	0.008
9) animali	-	-	-	-	-	-	-	-	-	-	-
10) macchinario	-	-	0.001	-	-	-	-	0.003	-	0.001	0.000
11) luoghi	0.001	0.003	0.001	-	-	-	0.001	0.002	-	-	0.000

Table 2: C_1-in-C_2

a	1)	2)	3)	4)	5)	6)	7)	8)	9)	10)	11)
1) att_mat	0.020	0.084	0.032	0.020	-	0.018	0.002	0.006	0.001	0.025	0.036
2) att_ment	0.037	0.351	0.077	0.055	0.008	0.019	0.016	0.005	0.001	0.024	0.049
3) manufatto	0.005	0.013	0.019	0.000	-	0.004	0.001	0.002	0.000	0.005	0.005
4) entita_umana	0.001	0.009	0.001	0.002	-	0.002	0.000	0.002	-	0.001	0.008
5) vegetale	0.001	0.001	0.002	-	-	-	-	0.001	-	0.001	-
6) costruzione	0.007	0.009	0.002	0.002	-	0.002	0.001	0.007	-	0.004	0.004
7) derivato	0.000	0.001	0.000	0.001	-	0.001	0.001	-	-	0.004	0.003
8) materiale	0.002	0.011	-	-	-	-	-	0.000	-	-	-
9) animali	-	0.000	-	-	-	-	-	-	0.001	-	-
10) macchinario	0.004	0.005	0.002	-	-	-	0.002	0.002	-	0.003	0.002
11) luoghi	0.001	0.005	-	0.007	-	-	-	-	-	0.001	0.017

Table 3: C_1-a-C_2

da	1)	2)	3)	4)	5)	6)	7)	8)	9)	10)	11)
1) att_mat	0.046	0.023	0.036	0.033	0.001	0.023	0.010	0.008	-	0.005	0.021
2) att_ment	0.022	0.012	0.047	0.052	0.001	0.037	0.001	0.001	-	0.004	0.032
3) manufatto	0.023	0.009	0.251	0.009	0.002	0.036	0.007	0.002	-	0.007	0.059
4) entita_umana	0.003	0.004	0.010	0.004	0.001	0.005	0.002	-	-	-	0.004
5) vegetale	0.002	-	0.001	-	-	0.001	0.002	-	-	-	0.016
6) costruzione	0.008	0.007	0.036	0.003	0.003	0.023	0.004	0.001	-	-	0.025
7) derivato	0.012	0.001	0.010	-	0.001	0.001	0.001	0.005	-	-	0.009
8) materiale	0.012	0.001	0.010	-	0.001	0.001	0.001	0.005	-	-	0.009
9) animali	0.003	-	0.001	-	0.003	0.003	0.008	-	-	-	0.006
10) macchinario	0.004	0.001	0.007	-	0.002	0.001	-	0.001	-	0.007	0.001
11) luoghi	0.003	-	-	0.001	-	0.002	-	-	-	-	0.001

Table 4: C_1-da-C_2

Legenda:

1) att_mat = PHYSICAL_ACT
2) att_ment = MENTAL_ACT
3) manufatto = ARTIFACT
4) entita_umana = HUMAN_ENTITY
5) vegetale = VEGETABLE
6) costruzione = BUILDING
7) derivato = BY_PRODUCT
8) materiale = MATTER
9) animali = ANIMALS
10) macchinario = MACHINE
11) luoghi = PLACES