# The Noisy Channel and the Braying Donkey

Roberto Basili (*), Maria Teresa Pazienza (*), Paola Velardi (·)

(*) Dept. of El. Engineering, University of Tor Vergata, Roma (ITALY),
{rbas,pazienza}@tovvxl.ccd.utovrm.it

(·) Istituto di Informatica, University of Ancona, (ITALY), vela@anvax2.cineca.it

**Abstract**
The title of this paper playfully contrasts two rather different approaches to language analysis. The "Noisy Channel" 's are the promoters of statistically based approaches to language learning. Many of these studies are based on the Shannons's Noisy Channel model. The "Braying Donkey" 's are those oriented towards theoretically motivated language models. They are interested in any type of language expressions (such as the famous "Donkey Sentences"), regardless of their frequency in real language, because the focus is the study of human communication.

In the past few years, we supported a more balanced approach. While our major concern is applicability to real NLP systems, we think that, after all, quantitative methods in Computational Linguistic should provide not only practical tools for language processing, but also some linguistic insight. Since, for sake of space, in this paper we cannot give any complete account of our research, we will present examples of "linguistically appealing", automatically acquired, lexical data (selectional restrictions of words) obtained trough an integrated use of knowledge-based and statistical techniques. We discuss the pros and cons of adding symbolic knowledge to the corpus linguistic recipe.

## 1. The "Noisy Channel" 's

All the researchers in the field of Computational Linguistics, no matter what their specific interest may be, must have noticed the impetuous advance of the promoter of statistically based methods in linguistics. This is evident not only because of the growing number of papers in many Computational Linguistic conferences and journals, but also because of the many specific initiatives, such as workshops, special issues, and interest groups.

An historical account of this "empirical renaissance" is provide in [Church and Mercer, 1993]. The general motivations are: availability of large on-line texts, on one side, emphasis on scalability and concrete deliverables, on the other side.

We agree on the claim, supported by the authors, that statistical methods potentially outperform knowledge based methods in terms of coverage and human cost. The human cost, however, is not zero. Most statistically based methods either rely on a more or less shallow level of linguistic pre-processing, or they need non trivial human intervention for an initial estimate of the parameters (training). This applies in particular to statistical methods based on Shannon's Noisy Channel Model (n-gram models). As far as coverage is concerned, so far no method described in literature could demonstrate an adequate coverage of the linguistic phenomena being studied. For example, in collocational analysis, statistically reliable associations are obtained only for a small fragment of the corpus. The problem of "low counts" (i.e. linguistic patterns that were never, or rarely found) has not been analyzed appropriately in most papers, as convincingly demonstrated in [Dunning, 1993].

In addition, there are other performance figures, such as adequacy, accuracy and "linguistic appeal" of the acquired knowledge for a given application, for which the supremacy of statistics is not entirely demonstrated. Our major objection to purely statistically based approaches is in fact that they treat language expressions like stings of signals. At its extreme, this perspective may lead to results that by no means have practical interest, but give no contribution to the study of language.

## 2...and the "Braying Donkey"'s

On the other side of the barricade, there are the supporters of more philosophical, and theoretically sound, models of language. We hope these scholars will excuse us for categorising their very serious work under such a funny label. Our point was to playfully emphasise that the principal interest in human models of language communication motivated the study of rather odd language expressions, like the famous "Donkey Sentences"[1]. The importance of these sentences is not their frequency in spoken, or written language (which is probably close to zero), but the specific linguistic phenomena they represent.

The supporters of theoretically based approaches cannot be said to ignore the problem of applicability and scalability but this is not a priority in their research. Some of these studies rely on statistical analyses to gain evidence of some phenomenon, or to support empirically a theoretical framework, but the depth of the lexical model posited eventually makes a truly automatic learning impossible or at least difficult on a vast scale.

The ensign of this approach is Pustejovsky, who defined a theory of lexical semantics making use of a rich knowledge representation framework, called the *qualia structure*. Words in the lexicon are proposed to encode all the important aspects of meaning, ranging from their argument structure, primitive decomposition, and conceptual organisation. The theory of *qualia* has been presented in several papers, but the reader may refer to [Pustejovsky and Boguraev, 1993], for a rather complete and recent account of this research. Pustejovsky confronted with the problem of automatic acquisition more extensively in [Pustejovsky et al. 1993]. The experiment described, besides producing limited results (as remarked by the author itself), is hardly reproducible on a large scale, since it presupposes the identification of an appropriate conceptual schema that generalises the semantics of the word being studied.

The difficulty to define scalable methods for lexical acquisition is an obvious drawback of using a rich lexical model. Admittedly, corpus research is seen by many authors in this area, as a tool to fine-tune lexical structures and support theoretical hypothesis.

## 3. Adding semantics to the corpus statistics recipe...

Indeed, the growing literature in lexical statistics demonstrates that much can be done using purely statistical methods. This is appealing, since the need for heavy human intervention precluded to NLP techniques a substantial impact on real world applications. However, we should not forget that one of the ultimate objectives of Computational Linguistic is to acquire some deeper insight of human communication. Knowledge-based, or symbolic, techniques should not be banished as impractical, since no computational system can ever learn anything interesting if it does not embed some, though primitive, semantic model[2].

In the last few years, we promoted a more integrated use of statistically and knowledge based models in language learning. Though our major concern is applicability and scalability in NLP systems, we do not believe that the human can be entirely kept out of the loop. However, his/her contribution in defining the semantic bias of a lexical learning system should ideally be reduced to a *limited amount of time constrained, well understood, actions, to be performed by easily founded professionals.* Similar constraints are commonly accepted when customising Information Retrieval and Database systems.

Since we are very much concerned with scalability *and* with what we call linguistic appeal, our effort has been to demonstrate that "some" semantic knowledge can be modelled at the price of limited human intervention, resulting in a higher informative power of the linguistic data extracted from corpora. With purely statistical approaches, the acquired lexical information has no linguistic content *per se* until a human analyst assigns the correct interpretation to the data. Semantic modelling can be more or less coarse, but in any case it provides a means to *categorise*

[1] where the poor donkey brays since it is beated all the time..

[2] this is often referred to as the *semantic bias* in Machine Learning

*language phenomena* rather that sinking the linguist in millions of data (collocations, n-grams, or the like), and it supports a more linguistically oriented large scale language analysis. Finally, symbolic computation, unlike for statistical computation, *adds predictive value* to the data, and ensures statistically reliable data even for relatively small corpora.

Since in the past 3-4 years all our research was centred on finding a better balance between shallow (statistically based) and deep (knowledge based) methods for lexical learning, we cannot give for sake of brevity any complete account of the methods and algorithms that we propose. The interest reader is referred to [Basili et al, 1993 b and c], for a summary of ARIOSTO, an integrated tool for extensive acquisition of lexical knowledge from corpora that we used to demonstrate and validate our approach.

The learning algorithms that we defined, acquire some useful type of lexical knowledge (*disambiguation cues, selectional restrictions, word categories*) through the statistical processing of syntactically *and* semantically tagged collocations. The statistical methods are based on distributional analysis (we defined a measure called *mutual conditioned plausibility*, a derivation of the well known *mutual information*), and cluster analysis (a COBWEB-like algorithm for word classification is presented in [Basili et al, 1993,a]). The knowledge based methods are morphosyntactic processing [Basili et al, 1992b] and some *shallow level of semantic categorisation.*

Since the use of syntactic processing in combination with probability calculus is rather well established in corpus linguistics, we will not discuss here the particular methods, and measures, that we defined. Rather, we will concentrate on semantic categorisation, since this aspect more closely relates to the focus of this workshop: *What knowledge can be represented symbolically and how can it be obtained on a large scale?*

The title of the workshop, *Combing symbolic and statistical approaches..*, presupposes that, indeed, one such combination is desirable, and this was not so evident in the literature so far. However, the *what-and-how* issue raised by the workshop organisers is a crucial one. It seems there is no way around: the more semantics, the less coverage. Is that so true? We think that in part, it is, but not completely. For example, categorizing collocations via semantic tagging, as we propose, *add predictive power* to the collected collocations, since it is possible to forecast the probability of collocations that have not been detected in the training corpus. Hence the coverage is, generally speaking, higher.

In the next section we will discuss the problem of finding the best source for semantic categorization. There are many open issues here, that we believe an intersting matter of discussion for the workshop.

In the last section we (briefly) present an example of very useful type of lexical knowledge that can be extracted by the use of semantic categorization in combination with statistical methods.

## 4. Sources of semantic categorization

We first presented the idea of adding semantic tags in corpus analysis in [Basili et al. 1991 and 1992a], but other contemporaneous and subsequent papers introduced some notion of semantic categorisation in corpus analysis. [Boggess et al, 1991] used rather fine-tuned selectional restrictions to classify word pairs and triples detected by an n-gram model based part of speech tagger. [Grishman 1992] generalises automatically acquired word triples using a manually prepared full word taxonomy. More recently, the idea of using some kind of semantics seems to gain a wider popularity. [Resnik and Hearst, 1993] use *Wordnet* categories to tag syntactic associations detected by a shallow parser. [Utsuro et al., 1993] categorise words using the "*Bunrui Goi Hyou*" (Japanese) thesaurus.

In ARIOSTO, we initially used hand assigned semantic categories for two italian corpora, since on-line thesaura are notcurrently available in Italian. For an English corpus, we later used *Wordnet.*

We mark with semantic tags all the words that are included at least in one collocation extracted from each application corpus.

In defining semantic tags, we pursued two contrasting requirements: portability and reduced manual cost, on one side, and the value-added to the data by the semantic markers, on the other side. The compromise we conformed to is to select about 10-15 "naive" tags, that mediate at best between generality and domain-appropriateness. Hand tagging was performed on a commercial and a legal domain (hereafter CD and LD), both in Italian. Examples of tags in the CD are: MACHINE (*grindstone, tractor, engine*), BY_PRODUCT (wine, *milk, juice*). In the LD, examples are: DOCUMENT (*law, comma, invoice)* and REAL_ESTATE (*field, building, house)*. There are categories in common between the two domains, such as HUMAN_ENTITY, PLACE, etc. The appropriate level of generality for categories, is roughly selected according to the criterion that words in a domain should be evenly distributed among categories. For example, BY_PRODUCT is not at the same level as HUMAN_ENITY in a domain general classification, but in the CD there is a very large number of words in this class.

For what concerns ambiguous words, many subtle ambiguities are eliminated because of the generality of the tags. Since all verbs are either ACTs or STATEs, one has no choices in classifying an ambiguous verb like *make*. This is obviously a simplification, and we will see later its consequences. On the other side, many ambiguous senses of *make* are not found in a given domain. For example, in the commercial domain, *make* essentially is used in the sense of manufacturing.

Despite the generality of the tags used, we experiment that, while the categorisation of animates and concrete entities is relatively simple, words that do not relate with bodily experience, such as abstract entities and the majority of verbs, pose hard problems.

An alternative to manual classification is using on-line thesaura, such as *Roget's* and *Wordnet* categories in English[3]. We experimented *Wordnet* on our English domain (remote sensing abstracts, RSD).

The use of domain-general categories, such as those found in thesaura, has its evident drawbacks, namely that the categorisation principles used by the linguists are inspired by philosophical concerns and personal intuitions, while the purpose of a type hierarchy in a NLP system is more practical, for example expressing at the highest level of generality the selectional constrains of words in a given domain. For one such practical objective, a suitable categorization principle is *similarity in words usage*. Though *Wordnet* categories rely also on a study of collocations in corpora (the *Brown* corpus), word similarity in contexts is only one of the classification principia adopted, surely not prevailing. For example, the words *resource, archive* and *file* are used in the RSD almost interchangeably (e.g. *access, use, read from resource, archive, file*). However, *resource* and *archive* have no common supertype in *Wordnet*.

Another problem is over-ambiguity. Given a specific application, *Wordnet* tags create many unnecessary ambiguity. For example, we were rather surprised to find the word *high* classified as a PERSON (*=soprano)* and as an ORGANIZATION (*=high school*). This wide-spectrum classification is very useful on a purely linguistic ground, but renders the classification unusable as it is, for most practical applications. In the RSD, we had 5311 different words of which 2796 are not classified in *Wordnet* because they are technical terms, proper nouns and labels. For the "known" words, the avergae ambiguity in Wordnet is 4.76 senses per word. In order to reduce part of the ambiguity, we (manually) selected 14 high-level *Wordnet* nodes, like for example: COGNITION, ARTIFACT, ABSTRACTION, PROPERTY, PERSON, that seemed appropriate for the domain. This reduced the average ambiguity to 1.67, which is still a bit too *high (soprano?),* i.e. it does not reflect the ambiguity actually present in the domain. There is clearly the need of using some context-driven disambiguation method to automatically reduce the ambiguity of *Wordnet* tags. For example, we are currently experimenting an algorithm to automatically select from *Wordnet* the "best level" categories for a given corpus, and eliminate part of the unwanted ambiguity. The algorithm is based on the Machine Learning method for word categorisation, inspired by the well known study on *basic-level* categories [Rosch, 1978], presented in [Basili et al, 1993a]. Other methods that seem applicable to the problem at hand have been presented in the literature [Yarowsky 1992].

[3] There is an on-going European initiative to translate Wordnet in other languages, among which Italian

## 5. *Producing wine, statements, and data*: on the acquisition of selectional restrictions in sub languages

Since our objective is to show that adding semantics to the standard corpus linguistics recipe (collocations + statistics) renders the acquired data more linguistically appealing, this section is devoted to the *linguistic analysis* of a case-based lexicon. The algorithm to acquire the lexicon, implemented in the ARIOSTO_LEX system, has been extensively described in [Basili et al, 1993c]. In short, the algorithms works as follows:
First, collocations extracted from the application corpus are clustered according to the semantic *and* syntactic tag[4] of one or both the co-occurring content words. The result are what we call *clustered association data*. For example, V_prep_N(*sell,to,shareholder*) and V_prep_N(*assign,to,tax-payer*), occurring with frequency *f1* and *f2* respectively, are merged into a unique association V_prep_N(*ACT,to,HUMAN_ENTITY*) with frequency *f1+f2*. The statistically relevant conceptual associations are presented to a linguist, that can replace syntactic patterns with the underlying conceptual relation (e.g. [ACT]->(*beneficiary*)->[HUMAN_ENTITY]).
These *coarse grained selectional restrictions* are later used in AIOSTO_LEX for a more refined lexical acquisition phase. We have shown in [Basili et al, 1992a] that in sub languages there are many unintuitive ways of relating concepts to each other, that would have been very hard to find without the help of an automatic procedure.

Then, for each content word *w*, we acquire all the collocations in which it participates. We select among ambiguous patterns using a preference method described in [Basili et al, 1993 b, d]. The detected collocations for a word *w* are then generalised using the coarse grained selectional restrictions acquired during the previous phase. For example, the following collocations including the word *measurement* in the RSD:

V_prep_N(*derive,from,measurement*),
V_prep_N(*determine,from,measurement*)
and V_prep_N(*infer,from,measurement*)

let the ARIOSTO_LEX system learn the following selectional restriction:

[COGNITION]
<-(*figurative_source*)<-[*measurement*],

where COGNITION is a Wordnet category for the verbs *determine, infer* and *derive*, and *figurative_source* is one of the conceptual relations used. Notice that *the use of conceptual relations is not strictly necessary*, though it adds semantic value to the data. One could simply store the syntactic subcategorization of each word along with the semantic restriction on the accompanying word in a collocation, e.g. something like: *measurement*: (V_prep_N *from*, COGNITION(V)). It is also possible to cluster, for each verb or verbal noun, all the syntactic subcategorization frames for which there is an evidence in the corpus. In this case, lexical acquisition is entirely automatic.

The selectional restrictions extensively acquired by ARIOSTO_LEX are a useful type of lexical knowledge that could be used virtually in any NLP system. Importantly, the linguistic material acquired is *linguistically appealing* since it provides evidence for a systematic study of sub languages. From a cross analysis of words usage in three different domains we gained evidence that *many linguistic patterns do not generalise across sub languages*. Hence, the application corpus is an ideal source for lexical acquisition.

In fig. 1 we show one of the screen out of ARIOSTO_LEX. The word shown is *measurement*, very frequent in the RSD, as presented to the linguist. Three windows show, respectively, the lexical entry that ARIOSTO_LEX proposes to acquire, a list of accepted patterns for which only one example was found (lexical patterns are generalized only when at least two similar patterns are found), and a list of rejected patterns. The linguist can modify or accept

[4] We did not discuss of syntactic tags for brevity. Our (not-so) shallow parser detects productive pairs and triples like verb subject and direct object (N_V and V_N, respectively), prepositional triples between non adjacent words (N_prep_N, V_prep_N), etc.

any of these choices. Each acquired selectional restriction is represented as follows:

*pre_sem_lex(word,conceptual_relation, semantic tag*[5]*, direction, SE,CF)*

the first four arguments identify the selectional restriction and the direction of the conceptual relation, i.e.:

[measurement]<-(OBJ)<-[COGNITION]
(e.g. *calculate, setup, compare...a measurement*)

[measurement]->(INSTRUMENT)->

[INSTRUMENTALITY]
(*e.g. measurement from satellite, aircraft, radar*)

*SE* and *CF* are two statistical measures of the *semantic expectation* and *confidence* of the acquired selectional restriction (see the aforementioned papers for details). ARIOSTO_LEX provides the linguist with several facilities to inspect and validate the acquired lexicon, such as examples of phrases from which a selectional restriction was derived, and other nice gadgets. For example, the central window in Figure 1 (opened only on demand) shows the Conceptual Graph of the acquired entry. The Conceptual Graph includes the extended *Wordnet* labels for each category.

One very interesting matter for linguistic analysis is provided by a cross-comparison of words, as used in the three domains. Many words, particularly verbs, exhibit completely different patterns of usage. Here are some examples:
The verb *produrre (to produce)* is relatively frequent in all the three domains, but exhibit very different selectional restrictions. In the RSD (Remote Sensing) we found the pattern:

produce
->(agent)->[ORGANIZATION, PERSON]
->(source)->[INSTRUMENTALITY]

or :

produce
->(theme)-> COGNITIVE_CONTENT
->(source)->INSTRUMENTALITY
e.g. *the satellite produced an image with high accuracy, the NASA produced the data..*

in the CD (commercial) we found:
produce
->(obj)-> ARTIFACT
->(agent)-> HUMAN_ENTITY
->(instrument)->MACHINE[6]
e.g.: *la ditta produce vino con macchinari propri (*the company produces wine with owned machinery)*

and in the LD (legal):
produce
->agent)->HUMAN_ENTITY
->(theme)-> DOCUMENT
e.g.: *il contribuente deve produrre la dichiarazione (the tax payer must produce a statement)*

It is interesting to see which company the word "*ground*" keeps in the three domains. The RSD is mostly concerned with its physical properties, since we find patterns like:
measure
->(obj)-> PROPERTY/ATTRIBUTE
<-(characteristic)<- ground
(e.g. *to measure the feature, emissivity, image ,surface of ground*)

In the CD, *terreno* (=*ground*) is the direct object of physical ACTs such as *cultivate, reclaim, plough*, etc. But is also found in patterns like:
BY_PRODUCT ->(source)-> terreno
(e.g. *patate, carote ed altri prodotti del terreno = potatoes, carrots and other ground products*)

in the LD, *terreno* is a real estate, object of transactions, and taxable as such. The generalised pattern is:

[5] Labels of *Wordnet* classes are sometimes denoted by abbreviations, e.g. CGN = 'cognition, knowledge'.

[6] MACHINE is the same as the *Wordnet* class INSTRUMENTALITY. Notice that we used *Wordnet* categories for our English corpus only later in our research. Perhaps we could find a *Wordnet* tag name for each of our previous manually assigned tags in the two Italian domains, but this would be only useful for presentation purposes. In fact, since there is not as yet an Italian version of *Wordnet* (though it will be available soon), we cannot classify automatically.

TRANSACTION->(obj)-> terreno
(*vendere, acquistare, permutare terreno = sell, buy, exchange a ground*)

AMOUNT <-(source)<- terreno
(e.g. *costo, rendita, di terreno= price, revenue of (=deriving from the ownership of) ground*)

And what is *managed* in the three domains? In the RSD, one manages COGNITIVE_CONTENT , such as *image, information, data* etc. The manager is a human ORGANIZATION, but also an ARTIFACT (a *system, an archive*).
In the CD, the pattern is:

manage ->(agent)->HUMAN_ENTITY
->(theme)->ACT
->(location)-> BUILDING

or

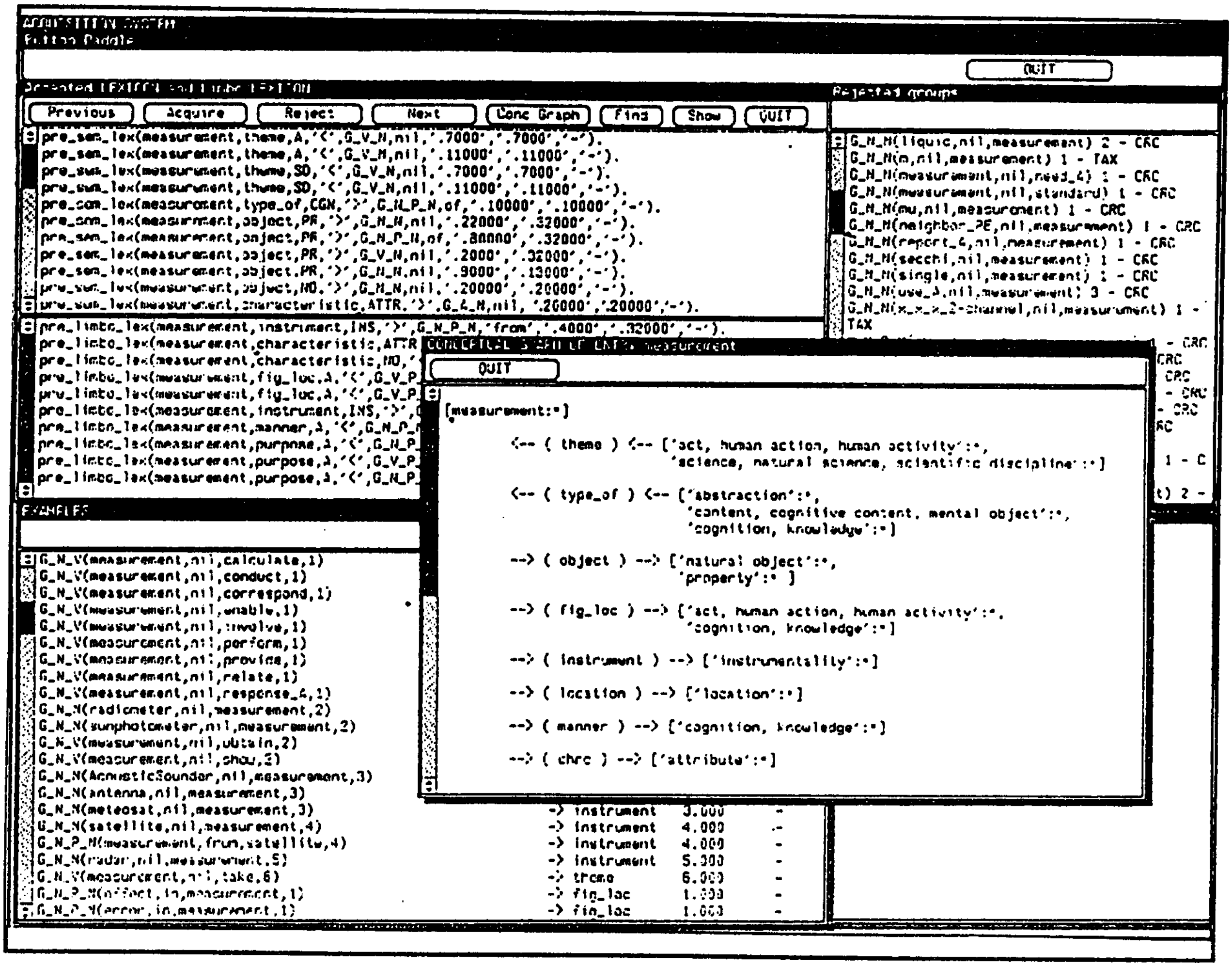
manage ->(agent)->HUMAN_ENTITY
->(obj)->BUILDING

(e.g. *gestire la vendita di alimentari nel negozio.. = to manage the sale of food in shops* )

Finally, in the LD, the pattern is:

manage
->(agent)->HUMAN_ENTITY
->(obj)->[AMOUNT,ABSTRACTION]

(e.g. *gestione di tributi, fondi, credito, debito etc. = management of taxes, funding, credit,debit*)



```
pre_sem_lex(measurement,theme,A,'<',G_V_N,nil,'.7000','.7000','-').
pre_sem_lex(measurement,theme,A,'<',G_V_N,nil,'.11000','.11000','-').
pre_sem_lex(measurement,theme,SO,'<',G_V_N,nil,'.7000','.7000','-').
pre_sem_lex(measurement,theme,SO,'<',G_V_N,nil,'.11000','.11000','-').
pre_sem_lex(measurement,type_of,CGN,'>',G_N_P_N,of,'.10000','.10000','-').
pre_sem_lex(measurement,object,PR,'>',G_N_N,nil,'.22000','.32000','-').
pre_sem_lex(measurement,object,PR,'>',G_N_P_N,of,'.80000','.32000','-').
pre_sem_lex(measurement,object,PR,'>',G_V_N,nil,'.2000','.32000','-').
pre_sem_lex(measurement,object,PR,'>',G_N_N,nil,'.9000','.13000','-').
pre_sem_lex(measurement,object,NO,'>',G_N_N,nil,'.20000','.20000','-').
pre_sem_lex(measurement,characteristic,ATTR,'>',G_A_N,nil, '.20000','.20000','-').
```

```
pre_limbo_lex(measurement,instrument,INS,'>',G_N_P_N,'from','.4000','.32000','-').
pre_limbo_lex(measurement,characteristic,ATTR
pre_limbo_lex(measurement,characteristic,NO,'
pre_limbo_lex(measurement,fig_loc,A,'<',G_V_P
pre_limbo_lex(measurement,fig_loc,A,'<',G_V_P
pre_limbo_lex(measurement,instrument,INS,'>',G
pre_limbo_lex(measurement,manner,A,'<',G_N_P_N
pre_limbo_lex(measurement,purpose,A,'<',G_N_P
pre_limbo_lex(measurement,purpose,A,'<',G_V_P
pre_limbo_lex(measurement,purpose,A,'<',G_N_P
```

```
G_N_N(liquic,nil,measurement) 2 - CRC
G_N_N(m,nil,measurement) 1 - TAX
G_N_N(measurement,nil,need_4) 1 - CRC
G_N_N(measurement,nil,standard) 1 - CRC
G_N_N(mu,nil,measurement) 1 - CRC
G_N_N(neighbor_PE,nil,measurement) 1 - CRC
G_N_N(report_4,nil,measurement) 1 - CRC
G_N_N(secchi,nil,measurement) 1 - CRC
G_N_N(single,nil,measurement) 1 - CRC
G_N_N(use_3,nil,measurement) 3 - CRC
G_N_N(x_x_x_2-channel,nil,measurement) 1 -
TAX
                                          - CRC
                                          CRC
                                           CRC
                                          - CRC
                                          - CRC
                                          RC

                                           1 - C

                                          t) 2 -
```

```
G_N_V(measurement,nil,calculate,1)
G_N_V(measurement,nil,conduct,1)
G_N_V(measurement,nil,correspond,1)
G_N_V(measurement,nil,enable,1)
G_N_V(measurement,nil,involve,1)
G_N_V(measurement,nil,perform,1)
G_N_V(measurement,nil,provide,1)
G_N_V(measurement,nil,relate,1)
G_N_V(measurement,nil,response_4,1)
G_N_N(radiometer,nil,measurement,2)
G_N_N(sunphotometer,nil,measurement,2)
G_N_V(measurement,nil,obtain,2)
G_N_V(measurement,nil,show,2)
G_N_N(AcousticSounder,nil,measurement,3)
G_N_N(antenna,nil,measurement,3)
G_N_N(meteosat,nil,measurement,3)
G_N_N(satellite,nil,measurement,4)
G_N_P_N(measurement,from,satellite,4)
G_N_N(radar,nil,measurement,5)
G_N_V(measurement,nil,take,6)
G_N_P_N(effect,in,measurement,1)
G_N_P_N(error,in,measurement,1)
```

```
-> instrument    3.000    -
-> instrument    4.000    -
-> instrument    4.000    -
-> instrument    5.000    -
-> theme         6.000    -
-> fig_loc       1.000    -
-> fig_loc       1.000    -
```

```
[measurement:*]

      <-- ( theme ) <-- ['act, human action, human activity':*,
                         'science, natural science, scientific discipline':*]

      <-- ( type_of ) <-- ['abstraction':*,
                           'content, cognitive content, mental object':*,
                           'cognition, knowledge':*]

      --> ( object ) --> ['natural object':*,
                          'property':* ]

      --> ( fig_loc ) --> ['act, human action, human activity':*,
                           'cognition, knowledge':*]

      --> ( instrument ) --> ['instrumentality':*]

      --> ( location ) --> ['location':*]

      --> ( manner ) --> ['cognition, knowledge':*]

      --> ( chrc ) --> ['attribute':*]
```

**Fig.1 The screenout for the lexical entry of the word "measurement"**

It is interesting to see how little in common these patterns have. Clearly, this information could not be derived from dictionaries or thesaura. Though the categories used to cluster patterns of use are very high level (especially for verbs), still they capture very well the specific phenomena of each sublanguage.

## 6. Concluding remarks

In this paper we supported the idea that "some amount" of symbolic knowledge (high-level semantic markers) can be added to the standard lexical statistics recipe with several advantages, among which categorization, predictive power, and linguitic appeal of the acquired knowledge. For sake of space, we could not provide all the evidence (algorithms, data and performance evaluation) to support our arguments. We briefly discussed, and gave examples, of our system for the semi-automatic acquisition, on a large scale, of selectionl restrictions. ARIOSTO_LEX has its merits and limitations. The merit is that it acquires extensively, with limited manual cost, a very useful type of semantic knowledge, usable virtually in any NLP system. We demonstrated with several examples that selectional restrictions do not generalize across sublanguages, and acquiring them by hand is often inintuitive and very time-consuming.
The limitation is that the choice of the appropriate conceptual types is non trivial, even when selecting very high-level tags. On the other hand, selecting categories from on-line thesaura poses many problems, particularly because the categorization principia adopted, may not be adequate for the practical purposes of a NLP system.

### References

[Basili et. al , 1991] R. Basili, M.T. Pazienza, P.Velardi, Using word association for syntactic disambiguation, in "Trends in Artificial Intelligence", Ardizzone, Gaglio, Sorbello Eds., in Lecture Notes in AI, Springer Verlag, 1991.

[Basili et al, 1992a] Basili, R., Pazienza, M.T., Velardi, P., Computational Lexicons: the Neat Examples and the Odd Exemplars, Proc. of Third Int. Conf. on Applied Natural Language Processing, Trento, Italy, 1-3 April, 1992.

[Basili et al, 1992 b] Basili, R., M.T. Pazienza, P. Velardi, A shallow syntactic analyzer to extract word associations from corpora, Literary and Linguistic Computing, 1992, vol. 7, n. 2, 114-124.

[Basili et al. 1993 a] Basili, R., Pazienza, M.T., Velardi, Hierarchical clustering of verbs, ACL-SIGLEX Workshop on Lexical Acquisition, Columbus Ohio, June, 1993.

[Basili et al, 1993b] Basili, R., M.T. Pazienza, P. Velardi, What can be learned from raw texts ?, forthcoming on Journal of Machine Translation.

[Basili et al, 1993c] Basili, R., M.T. Pazienza, P. Velardi, Acquisition of selectional patterns, forthcoming on Journal of Machine Translation.

[Basili et al, 1993d] Basili, R., M.T. Pazienza, P. Velardi, Semi-automatic extraction of linguisitc information for syntactic disambiguation, Applied Artificial Intelligence, vol. 4, 1993.

[Bogges et al. 1992] L. Boggess, R. Agarwal, R. Davis, Disambiguation of Prepositional Phrases in Automatically Labeled Technical Texts, Proc. of AAAI 1991

[Church and Mercer, 1993] K. Church and L. Mercer, Introduction to the sepcial issue on Computational Linguistics using large corpora, Computational Linguistics, vol. 19. n.1, 1993

[Dunning, 1993] T. Dunning, Accurate methods for statistical surprise and coincidence, Computational Linguistics, vol. 19. n.1, 1993

[Grishman, 1992] R. Grishman, Acquisition of selectional patterns, Proc. of COLING-92, Nantes, 1992

[Yarowsky 1992] D. Yarowsky, Word-sense disambiguation using statistical models of Roget's categories trained on large corpora, Proc. of COLING-92, Nantes, 1992

[Pustejovsky and Boguraev, 1993] J. Pustejovsky and B. Boguraev, Lexical Knowledge representation and natural language processing, Artificial Intelligence, vol.63, n. 1-2, pp. 193-224, October 1993

[Pustejovski et al, 1993] J. Pustejovsky, S. Bergler, and P. Anick, Lexical semantic techniques for corpus analysis, Computational Linguistics, vol. 19, n.2. 1993

[Rosch, 1978 ] E. Rosch, Principle of categorization, in Cognition and Categorization, Erlbaum 1978.

[Resnik and Hearst, 1993] P. Resnik and M. Hearst, Structural ambiguity and conceptual relations, ACL workshop on very large corpora, Ohio State University, June 22, 1993

[Utsuro et al, 1993] T. Utsuro, Y. Matsumoto, M. Nagao, Verbal case frame acquisition from bilingual corpora, proc. of IJCAI-93