ELSEVIER

Artificial Intelligence 85 (1996) 59–99

Artificial Intelligence

An empirical symbolic approach to natural language processing

Roberto Basili[a,*], Maria Teresa Pazienza[a,1], Paola Velardi[b,2]

[a]*Department of Computer Science, Systems and Production, University of Tor Vergata, Via della Ricerca Scientifica, 00133 Roma, Italy*

[b]*Istituto di Informatica, University of Ancona, Ancona, Italy*

Received October 1995; revised November 1995

Abstract

Empirical methods in the field of natural language processing (NLP) are usually based on a probabilistic model of language. These methods recently gained popularity because of the claim that they provide a better coverage of language phenomena. Though this claim is not entirely proved, empirical methods certainly outperform in this regard rationalist, or symbolic, methods. However, empirical methods provide a probabilistic, not conceptual, explanation of the analyzed linguistic phenomena. Probabilistic systems do "work" in real applications, and this is meritorious, but in our view they are intrinsically unable to provide insight into the mechanisms of human communication, because the output is represented by plain words, or word clusters, with attached probabilities. Eventually, a human analyst must make sense of these data. In the past few years, we explored the possibility of combining the advantages of empirical and rationalist approaches in NLP. Our objective was to define methods for lexical knowledge acquisition that are both scalable *and* linguistically "appealing", that is, amenable to a theoretically founded analysis of language. In this paper we describe and evaluate the results of a large-scale lexical learning system, ARIOSTO_LEX, that uses a combination of probabilistic and knowledge-based methods for the acquisition of selectional restrictions of words in sublanguages. We present many experimental data obtained from different corpora in different domains and languages, and show that the acquired lexical data not only have practical applications in NLP, but they are indeed useful for a comparative analysis of sublanguages. Importantly, ARIOSTO_LEX shed light on recurrent linguistic phenomena that have a problematic impact on the large-scale applicability of commonly used NLP techniques.

* Corresponding author. E-mail: basili@info.utovrm.it.

[1] E-mail: pazienza@info.utovrm.it.

[2] E-mail: velardi@anvax1.cineca.it.

SSDI 0004-3702(95)00116-6

1. The empiricist resurgence in natural language processing

At the beginning of the 1990s the field of natural language processing, whether we are willing to admit it or not, was at an impass. In 1989 the *Financial Times* [20] presented an overview of commercial and research systems based on NLP technology. The panorama of commercial systems was quite discouraging: the editorial was spread with sentences like: "*not yet robust enough*", ". . . *coverage is modest*", "*no computer has the background knowledge to resolve enough linguistic ambiguities . . .*" and concluded: "*the computer that can sustain a natural free-flowing conversation on a subject of your choice is unlikely to exist for several decades*".

Though perhaps no scientist in the field of NLP was seriously pursuing such an ambitious objective, we admittedly failed even with a much less complex task, i.e., that of building computer programs based on NLP technology that could increase the acceptability of computers in everyday situations. There are a variety of commercially important applications, such as on-line translation help, information retrieval, interfaces to databases, for which a deep understanding of the text is not required, and yet an even partial use of NLP technologies would have been greatly innovative. But even this limited objective was missed, since the actual impact of NLP systems on industrial applications was poor.

The major limitation of NLP was the limited coverage that language processors could exhibit when applied to real systems. The manual acquisition and codification of the various aspects of lexical knowledge was unrealistic as a systematic basis for most applications of practical interest. On the other hand, there was a proliferation of theories for lexical knowledge representation and analysis, whose merits and deficiencies could not be fully demonstrated in real applications.

In this panorama, two scientific events marked, in our view, the beginning of a new era, which we may call the *empiricists resurgence*:

- In 1990, a symposium was held in Stanford titled: *Text-based Intelligent Systems*. In his introductory note, Jacobs [33] baptised with this new name "*systems that combine artificial intelligence techniques with more robust but 'shallower' methods*". The meeting was a success, and so was the idea of improving the coverage of NLP systems by the use of shallow techniques.
- In the same year, a milestone paper was published on Computational Linguistics [15]. In this paper a group of researchers from IBM Thomas Watson Research Center proposed to use stochastic methods (widely used in the field of speech processing) in machine translation. The paper—and the proposed method—was a striking success.

The core idea of statistically based methods in NLP is to learn a predictive model of language through the extensive analysis of word patterns (and word translations, when available) from on-line resources. The first statistical methods were rather crude, generally limited to word counts in texts; current statistical methods are rather sophisticated.

The quantitative methods used in NLP can be roughly grouped according to the mathematical model adopted:

- A first research stream is concerned with the analysis of word co-occurrences. One popular measure used for co-occurrence analysis is *mutual information*, or one of its derivations (*t-score*, *lexical association*, etc.). The mutual information of two co-occurring words $W1$ and $W2$ compares the probability of observation of $W1$ and $W2$ together with the probability of observing them independently. An overview of statistical measures for co-occurrence analysis is in [17].
- A second stream recast the language modelling problem as one of computing the probability of a single word W given all the words that preceded W in a sentence. These statistical models are based on Shannon's Noisy Channel Model. An overview is found in [18].
- A third group, currently a minority, use machine learning methods to categorize language phenomena. In [40] the objectives and methods of this approach are summarized.

Probabilistic techniques have been applied with encouraging results to a variety of natural language processing problems, like syntactic [3] and semantic [47], disambiguation, part-of-speech tagging [36], word classification [28], automatic translation [15], etc.[3] The literature in this area has recently grown to a point that it is difficult to read everything that is published. Our contribution to the field will be summarized in the next sections.

One of the major claims of the followers of probabilistic approaches in NLP is scalability. Though we have been among the supporters of large-scale methods in linguistics [46] we think that there is no strict equivalence between probability calculus and scalability. Many probabilistic models in fact require quite an amount of manual work (for initial training and optimal parameter setting). Often, statistically reliable results are obtained only for a small fragment of the data, and in some case we suspect that the problem could have been handled more easily by hand. One example is represented by statistically based methods for sense disambiguation. Many methods described in the literature just do not scale up, because they require manual training[4] of the statistical model, for every ambiguous word. Perhaps an approach based on manually defined heuristic rules would be more general, though these types of comparative studies are not found in the literature.

Another problem with statistically based methods is that their output is represented by words, word strings, bi-lingual word correspondences, word clusters, etc., with attached probabilities. A conceptual explanation of the results is not provided. Eventually, a human analyzer must make sense of the data, to gain some linguistic insight into the matter. But a manual analysis is almost as complex as an inspection of raw tests, since there might be thousands or millions of different observed word patterns. For example, word clustering methods [22, 30,38] create word groups whose similarity on a linguistic ground can only be

[3] We selected here for brevity only one among the most representative papers for each application.
[4] That is, given a learning set of sentences including an ambiguous word, each occurrence of the word must be manually assigned to one of its senses.

evaluated by inspection. No conceptual description of a cluster is provided, in contrast to machine learning conceptual clustering methods.

In conclusion, though we agree that, in general, empirical methods in NLP outperform rationalist methods as far as coverage is concerned, we also believe that this claim should not be taken for granted. Furthermore, empirical methods are very much concerned with applications, which is meritorious, but they seem to be inherently inadequate for, or, in any case, poorly concerned with, a theoretical analysis of the linguistic material produced.

The thesis of this paper, and, we would say, of all our recent work, is that *pure symbolic methods may not scale up and pure quantitative methods may not dig deep*. An integration of the two is necessary to obtain both domain coverage *and* linguistic insight. We believe that the problem of balancing qualitative and quantitative methods, that we consider in the area of NLP, have an interest also for artificial intelligence as a field.

In the next sections, we describe ARIOSTO_LEX, a system that we developed in the past few years, using a combination of probabilistic and knowledge-based methods. ARIOSTO_LEX extensively acquires the selectional restrictions of words in sublanguages. The approach that we have undertaken in ARIOSTO_LEX is "empirical" in a more general sense than that adopted in the area of NLP. It is empirical because we start from the data (word observations in corpora) and then we derive an interpretation (the selectional restrictions). The interpretation of the data is not founded only on a probabilistic model, like in most corpus-based studies, but also on a "naive" semantic model, represented by a system of high-level semantic categories and relations. The semantic model is the bias of the lexical learning system. Its definition requires a minimal, relatively well-specified, human activity, which can be further reduced by the use of on-line thesaura.

The acquisition of an unrestricted case-based semantic lexicon was a rather challenging objective. The choices and methodologies that we adopted during the development of ARIOSTO_LEX implies the analysis of several aspects, such as knowledge representation, cognitive modelling, design and balancing of symbolic and probabilistic methods. However, in this paper we give emphasis to a qualitative and quantitative evaluation of the results. Though we provide a summary of the main processing steps, our aim will be to critically discuss many experimental data obtained from different corpora in different domains and languages. We show that the approach that we present, that is, a combination of symbolic and numeric methods, allows us to acquire lexical data that not only have practical applications in NLP, but are indeed useful for a comparative analysis of sublanguages. Furthermore, our experimental findings impact on the applicability of many popular NLP techniques.

2. An overview of ARIOSTO_LEX

ARIOSTO_LEX is a part of a corpus-based lexical learning system, ARIOSTO [8], that acquires several types of linguistic knowledge, like syntactic disambiguation

criteria [7] and conceptual clusters of words [6,10]. ARIOSTO has been applied so far to the fields of information retrieval and hypertextual navigation [3].

The general objectives of the ARIOSTO project are, on the linguistic side, to shed light on interesting language phenomena that are recurrent in sublanguages, and on the computational side, to demonstrate the advantages of combining probabilistic and knowledge-based techniques for large-scale lexical acquisition.

We used ARIOSTO to study several sublanguages, such as:

- a legal domain (LD) of taxation norms, in Italian;
- a commercial domain (CD) of agricultural activities, in Italian;
- a collection of remote sensing (RSD) abstracts, in English.

We are in the process of analyzing a medical domain in English and an environmental domain in Italian. In the near future, we plan a more systematic cross-analysis and categorization of sublanguage types.

ARIOSTO_LEX has the objective of acquiring extensively a lexicon of word sense selectional restrictions from application corpora. The lexicon is acknowledged as one of the major components of NLP and machine translation (MT) systems. It is broadly agreed that the most successful implementations of NLP-based systems so far have been those based on the lexicon. However, hand-built lexicons have obvious problems of size extension beyond, say, the 7–8k word barrier. Therefore, we may expect industrial interest in automatically sizable lexicons.

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

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. The major difficulty is that words relate to each other in many different, often domain-dependent ways. The current 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 unintuitive. It is not just a matter of violating the semantic expectations, as in "kill the process" or "my car drinks gasoline". Rather, there exist statistically relevant linguistic relations that are hard to imagine a priori, as almost never found in dictionaries, and are even harder to assign to the appropriate slot in the whatever conceptual structure is adopted for lexical representation. The key idea of ARIOSTO_LEX is to *tune* lexical knowledge, so that it expresses the precise semantic relationships present in a sublanguage, rather than the standard relationships found in general dictionaries.

A short description of the algorithm is presented here to summarize the main steps of analysis. The subsequent sections provide details on each step.

(1) The first step is to identify in an application corpus the statistically prevailing generalized semantic patterns (e.g. ACT-*with*-INSTRUMENTALITY). Generalized semantic patterns are detected by a shallow syntactic analyzer and by a semantic tagger. High-level semantic tags are assigned to words manually or by an on-line thesaurus, when available. We show that many detected semantic patterns are very unintuitive, and do not generalize across sublanguages.

(2) Then, generalized patterns are used by a linguist to identify the relevant selectional restrictions in sublanguages. The linguist must replace syntactic links with the appropriate conceptual relation, e.g. [ACT]→(INSTRUMENT)→[INSTRUMENTALITY].[5] We see no automatic way to identify conceptual relations in a sublanguage, though the descriptive power of a posited set of relations can be evaluated a posteriori. In any case, the use of conceptual relations is not essential in the subsequent acquisition step (though obviously they add informative power).

(3) Finally, we use the semi-automatically acquired "coarse" selectional restrictions as the "semantic bias" of an algorithm for the automatic acquisition of a case-based semantic lexicon. The algorithm extracts domain relevant selectional restrictions for *all* the content words w_i in a sublanguage, or at least for a statistically significant fragment of the sublanguage. Selectional restrictions are weighted by a statistical measure of the strength of their expectation in sentences including w_i.

Though more details on each step are needed, this brief description highlights some important advantages of this approach with respect to pure quantitative methods.

- *Digging deep*: Generalized co-occurrence patterns are linguistic material amenable to sublanguage analysis, while probabilistic methods let a human analyst sink into an ocean of data.
- *Scaling up*: Since the detected linguistic patterns are generalized, the method is less sensitive to the problem of low counts. Quantitative methods defined in the literature are unreliable when applied to linguistic patterns that have been rarely observed (this is a serious drawback since rare patterns are the majority). To obtain relatively good coverage, it is necessary to use learning corpora of several million words, that are rarely available. Instead, generalized patterns have a predictive power. It is possible to interpret word patterns that have never been observed in the learning corpus.

In Sections 2.1 and 2.2 we illustrate the method by which "coarse" co-occurrence patterns are extracted from corpora, and we provide a detailed discussion of the results, comparing these sublanguages. In Section 2.3 we describe the algorithm for the acquisition of a case-based semantic lexicon.

[5] Hereafter we will use the conceptual graph [44] notation to express selectional restrictions.

Finally, Section 2.4 provides a linguistic (though partial for sake of brevity) analysis of the data. A formal method for performance evaluation is presented in Section 3.

2.1. Acquisition of syntactically and semantically tagged word co-occurrences

The input to ARIOSTO_LEX is provided by a corpus pre-processing module, which is part of the ARIOSTO system. There are two phases:

(1) First, the corpus is analyzed by a grammar-based part of speech tagger and by a surface syntactic analyzer, described in [5,9]. The syntactic analyzer produces a database of productive word pairs and triples, which identify surface syntactic relations, like for example *N_V* (i.e., the subject relation), *V_N* (the direct object relation), *V_prep_N*, *N_prep_N* (prepositional phrases) etc. We call these triples elementary syntactic links (*esl*s). An *esl* has the following structure:

$$esl(w1, prep, w1)$$

where *prep* expresses the syntactic relation. In some case, $prep = nil$ (e.g. in *V_N* relations).

Each detected *esl* is weighted by a measure called *plausibility*, which is formally defined in [4]. To simplify, the plausibility of a detected *esl* is inversely proportional to the number of mutually excluding syntactic structures in a sentence. For example, in the sentence *John flies to Rome by plane*, the colliding *esl*s (or *collision sets*) are:

$$\{N\_prep\_N(Rome, by, plane), V\_prep\_N(fly, by, plane)\} .$$

The *evidence* of a given *esl* type in a corpus is computed as the sum of all the plausibility values of identical *esl*s (i.e., same words and same syntactic relation).

(2) Second, each word included in an *esl* is semantically tagged using a set of domain appropriate, high-level categories, like HUMAN_ENTITY, INSTRUMENTALITY, ABSTRACTION, etc. Though the selection of a domain appropriate set of categories is best performed manually,[6] the actual tagging may be automatic if an on-line thesaurus is available, like for example WordNet [12] in the English language.

There are several motivations for using high-level categories, that are briefly summarized hereafter:

First, high-level tags are less ambiguous and more intuitive. Hence, assigning a word to one (or more) categories is a relatively simple task

[6] In [29] a method is proposed to automatically create a flat set of categories from WordNet with a controlled upper and lower bound of the category size. However, since our objective was to select very few (about 12–15) domain appropriate categories, we found it more easy and reliable to perform manually the choice of the categories.

(when thesauri are not available for automatic tagging).[7] Second, high-level tags support a psychologically plausible model of *semantic bootstrapping* [39] in human language learning. They represent the semantic bias of our lexical learning system.

Semantically tagged *esl*s, e.g.

N_prep_N(*temperature*/PROPERTY, *in*, *water*/NATURAL_OBJECT)

are the input to ARIOSTO_LEX.

One important advantage of semantically tagged *esl*s is that they *significantly reduce the problem of low counts*, that is one of the major limitations of corpus-based statistical methods. In fact, the evidence of a pattern that is found only once in a corpus may be increased by the observation of syntactically and semantically similar patterns. For example, the pattern above is similar to the following:

N_prep_N(*emissivity*/PROPERTY, *in*, *air*/NATURAL_OBJECT) .

To evaluate numerically the advantages of semantically and syntactically marked co-occurrences with respect to the problem of low counts, we performed an experiment summarized in Table 1.

Table 1 shows the data obtained by extracting from the legal domain LD all the co-occurrences including the word *reddito* (*income*), using three methods.

(1) *Distance-based associations* are derived by extracting all the pairs where the second word is no more than 5 words apart from *income* (excluding articles and conjunctions, but not prepositions). Such "windowing" techniques are very popular in corpus-based literature.

(2) *Syntactic co-occurrences* are the *esl*s extracted by our syntactic analyzer, including the word *income*. In the literature, surface parsers are also used to detect co-occurrences.

(3) *Semantic co-occurrences* are obtained by the set of syntactic co-occurrences, where one (or more if ambiguous) semantic tag(s) is assigned to the word other than *income*.

For example, consider the sentence fragment:

Table 1
Statistics on different methods to detect co-occurrences

Method	Total	Different	Frequencies >3	% preserved information
(1) Distance-based	4044	623	3546	0.87
(2) Syntactic co-occurrences	5609	1454	4272	0.76
(3) Semantic co-occurrences	7048	312	6869	0.97

[7] In any case, on-line thesauri are produced by humans, and reflect the difficulties of manually defining word ontologies. In general, the classification choices relative to high-level categories are more "acceptable".

I redditi(1) *di*(2) *gestione*(3) *di*(4) *imprese*(5) *di*(6) *navigazione*(7) *marittima*(8) *o aerea*(9) *con*(10) *sede*(11) *di*(12) *direzione*(13) *effettiva*(14) *in*(15) *uno Stato*(16), [*sono imponibili*](17) *solo*(18) *nello*(19) *Stato*(20) . . .

The word-by-word translation is:

The income(1) *of*(2) (*means*: *deriving from the*) *management*(3) *of*(4) *companies*(5) *of*(6) *maritime*(8) *or aerial*(9) *navigation*(7) *with*(10) *office*(11) *of*(12) *actual*(14) *management*(13) (means: *that have a primary management office*) *in*(15) *a State*(16), [*is eligible for taxation*](17) *only*(18) *in*(19) *the State*(20) . . .

With method (1), the following associations are obtained, including the word *income*:

[1-2, 1-3, 1-4, 1-5, 1-6]

and with method (2):

1-2-3, 1-4-5, 1-6-7, 1-10-11, 1-12-13, 1-15-16, 1-17.

Notice that, though the surface syntactic parser detects syntactic relations that are not semantically correct in the sentence context (e.g. *reddito di imprese*, *reddito di navigazione*, etc.) it also detects a subject–verb relation (e.g. *the income is eligible*) that would be missed by most distance-based methods and surface parsers as well, because the two words are very far apart. Most statistical methods use a posteriori filtering techniques to reduce noise. However filtering techniques can increase the *precision* of the collected data, but not the *recall*. Therefore it is important to preserve as much as possible the initial information. The filtering techniques that we use are described next.

With method (3), some of the associations detected with method (2) are grouped together, because they are semantically similar. For example, *reddito di gestione* and *reddito di navigazione* are similar because the words *gestione* e *navigazione* (*management* and *navigation*) both belong to the same semantic category *activity* (ACT). This example is interesting, since though *reddito di navigazione* is not a correct attachment in the sentence context, it is semantically correct in general, that is to say, an income can be originated by an activity of navigation. Since we preserve the information on the structure of a detected pattern, we can use patterns that are not locally correct to improve domain knowledge.

In Table 1, the first column shows the total number of detected associations. It is seen that the number of syntactic and semantic associations is higher than that of distance-based associations, despite the fact that we collect triples, not only pairs. This result is explained by the fact that the legal corpus has many coordinations and nested prepositional sentences. Many related words are located at a distance higher than 5, as in the example before. Notice also that the total number of semantic associations is higher than that of syntactic associations, because of semantic ambiguity.

The second column shows the total number of different clusters. Distance-based associations are clustered when they have the same co-occurring words in the same order. Syntactic associations are clustered when they have the same words and the same *esl* type. The number of distinct semantic clusters is significantly low, because these are obtained by further clustering syntactic clusters of words with the same semantic tag, in the same order, like *reddito di navigazione* and *reddito di gestione* in the previous example.

A commonly used criterion is to consider "statistically reliable" associations that are detected with a frequency higher than 3 in a corpus. The term reliable obviously does not mean "correct". A language pattern may be extremely rare, but perfectly correct, while error prone language analyzers may repeat several times the same error, thus accumulating statistical evidence of a wrong pattern. However, the common wisdom suggests that noise is mostly originated by rare patterns, therefore it is more "reliable" to preserve only observations that have been detected more than, say, three times.[8]

A rather crude, yet meaningful, estimate of the amount of information that may be reliably acquired with the three methods is therefore given by the following measure:

$$\%\text{ of preserved information (PI)} = \frac{(\#\text{ associations with freq.} > 3)}{\#\text{ associations}}.$$

It is seen in column 4 of Table 1 that semantic associations allow it to preserve almost all (97%) the initial information, though all three methods perform rather well. This is because the word *income* is one of the most frequent words in the legal domain, and it has the tendency to appear in similar patterns. We hence extended our analysis to groups of words with different frequencies.

Fig. 1 plots the relation between PI (% of preserved information) and frequency range for the LD. The figure shows that even for low frequencies, semantic clustering allows it to preserve over 50% of the detected associations. Though clusters with more than three observations may well include noisy data, these can be filtered out a posteriori.

2.2. Acquisition of coarse selectional patterns

In order to identify concept patterns that are typical of a sublanguage, in ARIOSTO_LEX the statistical significance of a concept pair occurring with a given syntactic pattern (e.g. PROPERTY-in-NATURAL-OBJECT) is computed. More formally, we measure the probability of co-occurrence 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 (*els*) extracted by the shallow syntactic analyzer.

For each pattern C_1-*synt_rel*-C_2, we computed the *conditioned plausibility* $CP(C_1, synt\_rel, C_2)$ defined as

[8] Recently, so-called smoothing techniques have been adopted to reduce in part the problem of low counts [22,28].

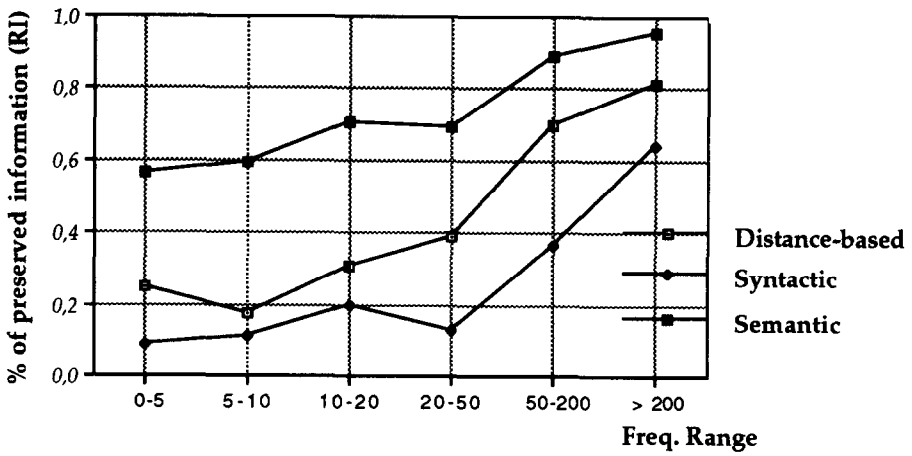

Fig. 1. % of preserved information.

$$CP(C_1, synt\_rel, C_2) = \frac{\sum_{w_1 \in C_1, w_2 \in C_2} pl(esl(w_1, synt\_rel, w_2))}{\sum_{w_1, w_2\, any} pl(esl(w_1, synt\_rel, w_2))}.$$

The numerator is the sum of the plausibility values of all the *esl*s of type *synt_rel* that relate word pairs belonging to the conceptual categories C_1 and C_2. The denominator is the sum of the plausibility values of all the *esl*s of type *synt_rel*. The reason for using the *CP* rather than other measures, like for example the popular *mutual information* used in [17] and in other studies, is that what matters here is to *detect all the statistically relevant phenomena*, and present them to a linguist.

Clustered *esl*s are used to build tables, one for each syntactic structure, whose element (x_i, y_j) represents the statistical significance in the corpus of a generalized pattern C_i-*synt_rel*-C_j. All the statistically prevailing couplings among classes are submitted to a linguist who replaces *synt_rel* with the appropriate (according to his/her intuition) *conceptual relation*.

A linguistic analysis of these tables is indeed useful to illustrate the advantages of the method. In what follows three tables are discussed, for three domains.[9]

Each table shows the distribution in a given corpus of the C_i-*synt_rel*-C_j associations. In illustrating groups of similar conceptual patterns detected by the system, we propose an interpretation of the type of conceptual relation subsumed (e.g. the preposition *to* followed by HUMAN_ENTITY subsumes a *beneficiary* or *recipient* case relation).

[9] The tables and a list of the semantic tags used are in Appendix A.

Table A.1 summarizes the relations C_1-*per*-C_2 (*per* = *for*), in the commercial corpus (for brevity, CD in what follows). Some of the pertinent associations are:

- ARTEFACT *for* ACT (e.g. *articoli per lo sport* (**items for sport*),[10] *attrezzi per giardinaggio* (**tools for gardening*)),
- ARTEFACT *for* BUILDING (e.g. *biancheria per la case* (**linens for the house*), *mobili per negozi* (**furnitures for shops*)),
- MACHINE *for* BUILDING (e.g. *macchinari per laboratori* (**equipment for laboratories*), *macine per mulini* (**grindstones for mills*)).

All the above relations subsume the *use* relation (e.g. *tools used for gardening*).

Notice that the most intuitive senses of the preposition *for*, that is, *beneficiary* and *purpose* are not the most frequent in the corpus. The only statistically relevant *beneficiary* relations hold between ARTEFACT and HUMAN_ENTITY (e.g. *calzature per uomo* (**shoes for man*), *biancheria per signora* (**linens for lady*)) and HUMAN_ENTITY-HUMAN_ENTITY (e.g. *parrucchiere per signora* (**hairdresser for lady*)).

The proposition *for* has a relatively more conventional use in the legal corpus, as seen in Table A.2. Examples of frequent relations are:

- ACT *for* ACT (e.g. *pagare per prestazione* (*to pay for a job*)), interpreted by the *cause* relation,
- ACT *for* ABSTRACTION (e.g. *assegnare per categoria* (**to assign for* (= *by*) *category*)), interpreted by the *manner* relation,
- ACT *for* AMOUNT (e.g. *disintinguere per aliquota* (**distinguish for* (= *by*) *rates*), interpreted by the *manner* relation.

In the RSD (Table A.3) the following uses of the preposition *for* are frequent:

- {ACT, COGNITIVE_PROCESS, ABSTRACTION} *for* {ACT, COGNITIVE_PROCESS, DISCIPLINE} (*method for evaluation*, *technique for remote-sensing*, etc.), interpreted by the *purpose* relation,
- {ACT, COGNITIVE_PROCESS} *for* {LOCATION, INSTRUMENT, NATURAL_OBJECT} (*analysis for atmosphere*, *calculation for satellite*, *data for Orgeon*, etc.), in which the underlying relation is *reference* (e.g. *analysis concerning*, *referring to*, *the atmosphere*),
- {COGNITIVE_PROCESS} *for* {INDIVIDUAL, ORGANIZATION} (*study for university*), interpreted by the *recipient* relation, and
- {INSTRUMENTALITY, ARTEFACT} *for* {ACT, COGNITIVE_PROCESS} (*spectrometer for the analysis*) that subsumes the *use relation*.

We spent some time illustrating the tables to support[11] to one of the results of this study, i.e., that selectional restrictions are less intuitive than what usually appears in the literature on computational lexicons (and general dictionaries). There is a remarkable difference in the use of the same prepositions in the three domains, and in the way words relate to each other. Many patterns do not generalize across sublanguages.

[10] The asterisk indicates a literal translation. In many cases, the English translation would be a compound, e.g. *sport items*.

[11] Many other examples have been discussed in [8].

In the examples, we also provided a semantic interpretation of some C_i-*synt_rel-*C_j pattern. As we remarked earlier, the task of associating with a syntactic pattern an appropriate conceptual relation cannot be automated. The problem is that, notwithstanding the large consensus on the use of relational models for the lexicon, there is the greatest disagreement on the number and type of relations posited [23]. To define a set of relations for our sublanguages, we relied as much as possible on previous work on semantic networks, for example [44]. We used "commonly agreed" relations, i.e., relations frequently mentioned in the computational linguistic literature, like *agent*, *object*, *theme*, *patient*, *location*, *purpose*, *instrument*, etc., plus a small set of relations that we found necessary to interpret specific word patterns in the three sublanguages, like *figurative_location* (*to include in the program*), *reference* (*the data for Oregon*) etc. Overall, we used about 30 conceptual relations, but we did not use exactly the same set of relations for the three sublanguages. For each domain, we prepared a list of correspondences between C_i-*synt_rel-*C_j patterns and concept–relation–concept triples (CRC). On average, there are 10–30 different CRC for each syntactic pattern type (compounds in English have the highest number of interpretations). There are prepositions that have only one straightforward semantic interpretation (e.g. *by means of* = *manner*).

In any case, we do not consider the task of preparing CRC tables as particularly relevant for the purpose of demonstrating the thesis of this paper. The choice of a conceptual relation name may sound more or less appropriate to the reader, but what matters here is not our personal view of a relational model, but rather the possibility of *automatically detecting generalized word patterns that are frequent in a given sublanguage. These patterns are highly variable within and among sublanguages. It would be very difficult for a linguist to identify them all using only his/her intuition of a language domain.*

In the next sections, we will show how the automatically detected conceptual patterns can be used as the *semantic bias* of a system for the acquisition of a semantic lexicon.

2.3. Acquisition of a case-based lexicon

The CRC are the "semantic bias" of ARIOSTO_LEX. The process of acquiring word sense selectional restrictions is summarized in what follows:
For any word W in the corpus and any sense of W:

(1) *Select* all the *esl*s that include W as one of the arguments, e.g. if W = *data* in the RSD: *N_prep_N*(*data*, *from*, *satellite*) *N_prep_N*(*data*, *from*, *radar*), *V_N*(*analyze*, *data*), *N_V*(*data*, *demonstrate*), *N_prep_N*(*data*, *for*, *Oregon*), *N_prep_N*(*data*, *from*, *code*), etc. The *esl*s are collected with their global plausibility value in the corpus.

(2) For any *esl*, *given* the semantic tag(s) assigned to W and to its co-occurring word, and *given* the type of syntactic relation, *find* the appropriate semantic interpretation(s) or reject the *esl*. Put rejected patterns (e.g.

(*data*, *from*, *code*)) in a list R, and accepted in a list A. Both lists are available for inspection to a linguist.

Example: one of the existing CRC interpretations of the preposition *from*, in the RSD, is: [MENTAL_OBJECT]→(SOURCE)→[INSTRUMENTALITY], since "*data*" is classified in WordNet as a MENTAL_OBJECT and "*satellite*" as an INSTRUMENTALITY, *N_p_N*(*data*, *from*, *satellite*) receives the interpretation: [data]→(SOURCE)→[satellite].

(3) If at least two *esl*s are interpreted by the same CRC, *generalize* and create a new selectional restriction for W.

E.g. If in the corpus we find: [data]→(SOURCE)→[satellite], [data]→(SOURCE)→[radar] the following *selectional restriction* is acquired for the word sense *data*/MENTAL_OBJECT: [data]→(SOURCE)→[INSTRUMENTALITY].

(4) For each acquired selectional restriction of W, $SR(W)$, *compute* two statistical figures:

- The *semantic expectation* $SE(SR, W)$, given by $SE(SR, W) = freq(SR(W))/freq(W)$, where $freq(SR(W))$ is the (plausibility-weighted) frequency of *esl*s including W that are interpreted by the selectional restriction SR, and $freq(W)$ is the absolute frequency of phrases including W in the corpus.

  High values of the semantic expectation suggest that, when parsing a sentence including the word W, that particular case role must be filled in order for the parse to succeed. The semantic expectation is particularly important for verbs and verbal nouns, whose thematic structure is used in many semantic interpretation algorithms to guide parsing. However, we will see later that most verbs have shallow expectations.
- The *certainty factor* $CF(SR, W)$ is a Boolean measure. CF is set to 1 whenever a selectional restriction $SR(W)$ has been learned from at least one unambiguous sentence structure, elsewhere it is set to 0. For example, if the program analyzes the sentence: *sottomettere all'autorita' competente* (*to submit to the competent authority*), it will have no doubt that *submit to authority*/HUMAN_ENTITY is indeed a valid pattern. In other words, *short unambiguous sentences in the corpus are used to increase the reliability of the acquired relations*.

The lexicon derived by ARIOSTO_LEX is not intended to be the "final" lexicon to be used in a NL or MT system. A post-editing by a linguist is suggested, and ARIOSTO_LEX provides a nice environment for doing so (see next section). One problem that is partially handled at this stage is semantic ambiguity. The over-general semantic tags make possible to discriminate among strongly different interpretations, e.g. *bank*/BUILDING, from *bank*/LOCATION. This discriminating power is "good enough" for many concrete nouns, especially in restricted sublanguages, but does not capture the subtle ambiguities of verbs. Therefore, different senses of verbs may collapse into a single entry in the lexicon. The problem of systematic sense disambiguation, especially for verbs, is a very complex one. See [10] for a first investigation on the matter.

2.4. Linguistic analysis of the lexicon

ARIOSTO_LEX acquires the selectional restrictions of all the content words in the corpus that are found frequently enough to generalize some of their patterns of usage (step (3)). However, step (2) allows it to interpret patterns that are found just once in a corpus. This is an important advantage over empirical methods that do not rely on any semantic model.

In this section we give a brief account of the linguistic material produced when processing the three domains (CD, LD and RSD). The reported data have been acquired from learning corpora of about 500,000 words each, belonging to the three domains.

Fig. 2 shows the lexical entry for the verb *to obtain*, that has a relatively high frequency in the RSD, as presented to the linguist. Four windows show, respectively, the selectional restrictions that ARIOSTO_LEX proposes to acquire (window 1, upper left), a list of accepted selectional restrictions for which only one example was found (window 2, middle left) called *limbo*, a list of rejected *esl*s (window 3, upper right), and the list of accepted *esl*s with the underlying conceptual relation (window 4, lower left). ARIOSTO_LEX is based on shallow NLP and statistical techniques, therefore it may collect some noisy relations. An environment is hence provided to review the output and optimally tune the system. The linguist can modify or finally accept any of the selectional restrictions in windows 1 and 2. For any rejected pattern in window 3, an explanation is provided. A pattern can be rejected because no CRC could interpret the pattern (CRC), or because a word in the *esl* was not found in the morphologic lexicon (MOR), or it was not classified (TAX). In each case, the linguist may decide to

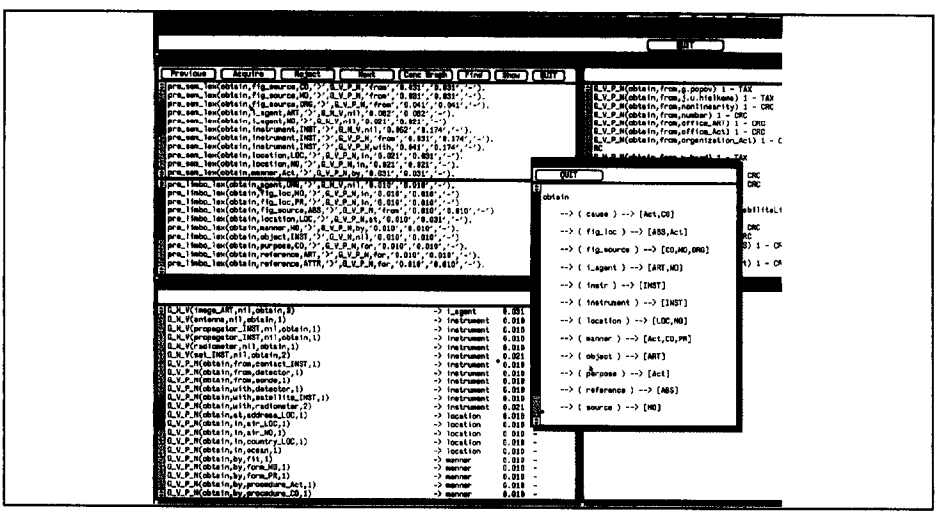

Fig. 2. ARIOSTO_LEX: Screendump of a validation session for the entry *to_obtain*.

update the CRC, or the morphologic lexicon, or the taxonomy, within the same environment. He/she may also browse the corpus and the database of *esl*.

Each acquired selectional restriction is represented as follows:

pre_sem_lex(*word*, *conceptual_relation*, *semantic tag*, *direction*, *esl type*, *preposition*, *SYNE*, *SE*, *CF*).

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

[obtain] ← (PURPOSE) ← [COGNITIVE_PROCESS]
{e.g. *variance is computed in order to obtain*}

or

[obtain] → (INSTRUMENT) → [INSTRUMENTALITY]
{e.g. *obtain an image from the satellite*}.

SYNE is the syntactic expectation,[12] *SE* and *CF* have been described earlier.

In this form, the lexical entry is rather sparse, because syntactic and statistical data are also shown. The same selectional restriction may be generated by different syntactic patterns. However, the central right window in Fig. 2 (opened only on demand) provides a more compact semantic representation of the lexical entry, that is the conceptual graph.

Fig. 2 is a good summary of some recurrent findings of our research: The verb *to obtain* in the RSD, like many others, has a rather technical use. The most frequent subject for this verb does not belong, as expected, to the category HUMAN_ENTITY (e.g. INDIVIDUAL + ORGANIZATION). This pattern was found only once, therefore it appears in window 2 (first line of the limbo). Commonly, the subjects are words belonging to abstract categories such as COGNITIVE_PROCESS (CO), e.g. *the following analysis obtained results* . . . , or concrete nouns, such as ARTEFACT (ART), INSTRUMENTALITY (INST) or NATURAL_OBJECT (NO), e.g. *the antenna obtains* . . . . These patterns of use could not be deduced from a standard dictionary definitions. For example, the Webster's dictionary gives the following definition for the verb *to obtain*: "1. *To gain possession of, to acquire.* 2. *To be widely acceptable.*"

It also is interesting to observe, in Fig. 2, that the *SE* values are relatively low. The highest expectation seen in Fig. 2 is associated to the *instrument* relation. In fact, most sentences with the verb *to obtain* in the RSD specify some instrument used to obtain information or an artifact (e.g. *to obtain an image from satellite*). As defined, *SE* is 1 only if a given selectional restriction for a word is found with plausibility 1 in *every* sentence including that word. But the cases of $SE = 1$ are

[12] Its definition follows straightforwardly that of *SE*: it is the probability that *word* is used in the corpus with the *conceptual relation* expressed by a syntactic relation represented by *esl type* and *preposition*.

very rare. Even verbs that always take a preposition, like *associate with*, *go to*, *relate to*, etc., have a prepositional modifier belonging to different semantic categories. For example, one can *relate to data*/MENTAL_OBJECT or *relate to the turbulence*/PROCESS. We experimentally determined that *SE* values higher than 0.5 indicate highly expected relations in the RSD, but even these values are rare. Strong semantic expectations are conveyed only by a very restricted number of verbs, some of which have the tendency to appear not only within similar contexts, but within almost identical expressions. Verbs with strong expectations are frequent in the commercial domain, which adopts a telegraphic, stereotypical style. In contrast, in the legal domain, *SE* values are even lower, because there is the highest syntactic ambiguity of sentences (hence plausibility values are low), and because the style is less technical.

Another observation emerging from an analysis of Fig. 2 is that the relational patterns of words may be highly variable, despite the fact that high-level semantic categories are an inherent limit to the detection of very fine-tuned relations.

Table 2 shows the general validity of this finding. The table lists, for the LD in Italian and RSD in English, the average number of detected selectional restrictions for verbs. Verbs are grouped in three classes according to their frequency in the corpus.

Though the number of detected selectional restrictions is obviously lower when we had fewer examples of a verb context, the total number of different detected selectional restriction types is about the same (around 80) for each frequency range and for the two domains. This is because the class of lower frequency verbs is more populated. It is also remarkable that the average number of different relations attached to a verb is around 30, whenever it is possible to gain enough evidence on its patterns of use in a sublanguage.

The data in Table 2 provide experimental evidence to justify the well-known difficulty of defining (manually or automatically) a verb taxonomy, based on the identification of common relational patterns of verbs.

Interesting matter for linguistic analysis emerges from a comparison between the three sublanguages. Many verbs exhibit completely different patterns of use. For example, the verb *produrre* (*to produce*) is relatively frequent in all the three domains, but occurs in very different contexts.

In the RSD we found for example:

Table 2
Average number of selectional restrictions per verb in two sublanguages

Verb frequency ranges	Legal domain		Remote sensing domain	
	Average # relations per verb	# of different detected relations	Average # relations per verb	# of different detected relations
$x < 10$	3.76	81	6.4	76
$10 < x < 100$	9.8	85	17.9	80
$x > 100$	27	85	35.2	80

ORGANIZATION produce ABSTRACTION or MENTAL_OBJECT

or

INSTRUMENT produce ABSTRACTION or MENTAL_OBJECT

e.g.

the NASA/ORGANIZATION *produced the image*/MENTAL_OBJECT
the satellite/INSTRUMENTALITY *produced data*/MENTAL_OBJECT
with high accuracy .

In the CD we found:

ORGANIZATION produce ARTEFACT or INSTRUMENTALITY with INSTRUMENTALITY

e.g.

la ditta produce articoli in pelle con macchinari propri
(*the company*/*ORGANIZATION produces items*/ARTEFACT *in leather with owned machinery*/INSTRUMENTALITY)

and in the LD:

ORGANIZATION produce DOCUMENT

e.g.

la società deve produrre un attestato
(*the company*/ORGANIZATION *must produce a form*/DOCUMENT) .

Once again, it appears that word patterns do not generalize across sublanguages. Often, words are used in a much narrower (and sometimes unintuitive) sense than that suggested by dictionaries or common sense.

Finally, our data provide more insight into the problem of relating conceptual roles and syntactic structures. Notice for example, in window 4 of Fig. 2, that different syntactic patterns may have the same semantic interpretation (e.g. *instrument*).

Another example is provided again by the verb *to produce*.

In the RSD, one typical conceptual pattern is:

[produce]-
(INSTRUMENT)→[INSTRUMENTALITY]
(OBJ)→[MENTAL_OBJECT]
(MANNER)→[PROPERTY]

as in

the satellite produces an image with high accuracy

whereas in the CD a typical pattern is:

[produce]-
(AGENT)→[ORGANIZATION]
(OBJ)→[BY_PRODUCT]
(INSTRUMENT)→[INSTRUMENTALITY]

as in:

la ditta produce vino con macchinari propri

(*the company produces wine with owned machinery*).

These examples show that different syntactic relations subsume the same semantic relation *instrument* (*the satellite produces* versus *produces with machinery*), while the same syntactic relation (*produce with accuracy* versus *produce with machinery*) has different semantic interpretations, namely *manner* and *instrument*.

This example (as many others that can be found) demonstrates that generalizing word patterns on the basis of syntactic similarity may cause problems. *Syntactic similarity is not always a systematic marker of semantic similarity*. This motivates our choice of using conceptual relations, at the price of some additional human labour.

In this section, we reported and commented on a variety of linguistic phenomena that are recurrent across different technical domains, styles and languages. Our findings can be summarized as follows.

- Word uses may be very different across sublanguages. Often, the patterns of use cannot be deduced from standard dictionary definitions.
- Syntactic similarity is not a reliable marker of semantic similarity.
- The case structure of words, especially of verbs, is highly variegated and poorly overlapping. Furthermore, it is difficult to identify strongly expected patterns of use for a verb.

The last statement is the most problematic. One consequence is that there are inherent limitations to the possibility of defining a language independent ontology for verbs, at least at the lowest levels. Another consequence is that the effectiveness of expectation-driven semantic interpreters seems to be limited to applications in which strong semantic expectations are imposed from the outside world, such as for example in NL interfaces to databases.

3. Performance evaluation

In the previous section we analyzed the results of ARIOSTO_LEX on linguistic grounds, and we discussed the impact of our findings on commonly used NLP techniques. This section provides a quantitative evaluation of ARIOSTO_LEX.

There are two issues related to this task.

- The first is that the problem of lexicon evaluation is theoretically underdetermined.
- The second is that the evaluation frameworks adopted in the area of computational linguistics are not fully adequate to measure the complexity of the lexical learning system.

As for the first issue, we must notice that in the literature *there are no effective evaluation mechanisms for lexicons*. Frequently, attention is paid to the problem of making a lexicon consistent and "provably correct", but, to agree with Yorick Wilks, "lexicons are inconsistent and could not be proved consistent even if they happen to be so". Furthermore "comparative evaluation is impossible, since the lexical component of a (NLP) system cannot be alternated with an alternative to compare the final output, while retaining the rest of the system constant".[13] Finally, we believe that the evaluation of a lexicon should not be confused with the evaluation of a system that *uses* the lexicon. For example, there are international conferences, like MUC (Message Understanding Conference) and TREC (Text Retrieval Conference), in which different systems that use a given technology are compared against some common task of language understanding. However, the linguistic components of an NLP-based system are many, and it is difficult to establish, for example, whether a sentence misunderstanding was originated by the lexicon, or by some deductive component, or by the grammar, etc.

Therefore, we see two possible ways to attack this problem: The first, which we pursue in this section, is to consider ARIOSTO_LEX as a self-standing knowledge base of lexical facts, and to evaluate its performances limited to some specific and intuitive task. We call "intuitive" a task for which a human can (relatively) easily judge the decision of the system.

The second, is to evaluate ARIOSTO_LEX *within* some application, for example an information extraction system. In this second case, we do not see the possibility of performing a comparative experiment, for the motivations stated above. Rather, since we propose a dynamic, adaptable lexicon, we could compare the system performance over time, at different stages of lexical tuning to a corpus.[14] This second experiment requires an experimental setup and a design effort that we hopefully will be able to carry on in cooperation with other research sites, within a larger project that is now under final definition.

As far as the evaluation framework is concerned, in the NLP literature, the most popular evaluation parameters to compare the performance of *inductive systems* (from statistically based syntactic analyzers to IR systems, or pattern classifiers) are *accuracy* (also referred as *efficiency* or *recall*) and *precision*. Accuracy measures the number of correct decisions over the number of expected decisions (or test cases). Precision is the number of correct decisions over the global number of automatic decisions.

---

[13] Wilks, personal communication, January 1995.

[14] We are in debt with Yorick Wilks for his intuitions on this problem.

As already observed by different authors, accuracy and precision do not provide a good measure of the quality of different inductive systems, since they are not sensitive to the *complexity of the decision task* at hand [34].

This criticism can be intuitively understood with the following example: Say we wish to assign instances of an observed phenomenon to one of n classes. The "blind" probability to assign the correct class in absence of any decision strategy is $1/n$. Clearly, an inductive system that selects the correct class over n, with an 80% precision, does in fact a much better job than one that exhibits a 90% precision at selecting among $n/2$ classes.

The absence of a notion of blind probability, called in [34] the *a priori*, or *prior*, probability, renders the evaluation of classification methods presented by different authors very hard to compare. In this section, it is shown that the information gain is more adequate than other popular evaluation parameters, like recall and precision, since it takes into account the complexity of the learning task at hand. In what follows we perform a systematic evaluation of ARIOSTO_LEX, limited to a task of syntactic disambiguation.

The problem that we analyze numerically is whether the use of very high-level semantic tags still provides good interpretative power. In fact, clustering word associations by very coarse classes could create unacceptable noise due to cumulative effects of morphology and parsing errors, syntactic ambiguity, and polysemy. Though obviously initial noise can be reduced through an interaction with the linguist, who may refine the CRC, add a new syntactic rule, and revise some of the classification choices, the problem of over-generality is inherent to the adopted lexical acquisition model.

This evaluation, though concerning only one aspect of ARIOSTO_LEX, has two advantages: First, it is not necessary to have a complete NLP system available, but only a syntactic analyzer.[15] Second, the results can be compared with other disambiguation methods presented in the literature, since no evaluation data are available for the computational lexicons developed (or being developed) in the literature (see Section 4).

3.1. Measuring the information gain of inductive linguistic methods

In this section we show that modeling the PP attachment problem as a classification task favours the definition of a more principled, and uniform, notion of performance. This evaluation framework provides a unified view of the different disambiguation methods, and allows a systematic study of performances in terms of *information gain*.

However, the linguistic nature of the classification problem at hand renders a formal definition of the prior (and posterior) probability rather more complex than for other "standard" classification tasks. For example, the prior probability of a given class is often evaluated as the number of available training instances in

[15] This is not an easy matter, but we developed a full (i.e., not shallow) grammar of the legal language that produced complete parse trees for over 400 complex sentences in the LD.

the class [34]. Furthermore, the test set is relatively small and well assessed. In the case of lexical learning, instead, the acquisition of information is triggered by a noisy, unsupervised set of instances, i.e., the training corpus.

In this section we devise a formal definition of the information gain of the PP disambiguation task. The definition will be applied to the evaluation of two syntactic disambiguation methods, precisely the popular *lexical association* [31] and the *semantic expectation* computed in the ARIOSTO_LEX system. The evaluation method can easily be extended to other lexical learning tasks.

First we show how syntactic disambiguation can be modelled as a classification task. Any sample set of sentences Σ is characterized by one (or more) underlying parse tree $t(s)$. A given grammar Γ provides, for each sentence $s \in \Sigma$, a related forest $\gamma(s)$. In general, $\gamma(s)$ includes also the correct parse $t(s)$, i.e., $t(s) \in \gamma(s)$.

Furthermore, all the parse trees in $\gamma(s)$ are partitioned in two disjoint classes: G, the class of correct (or meaningful) trees, that express the structure assigned to s by any language user, and E, the class of wrong, or meaningless, parses, generated by some deficiency of the adopted grammatical model.

More generally, for all the available sentences Σ, we have a sample space $\Omega = \bigcup_{s \in \Sigma} \gamma(s)$ of all the derived trees that is partitioned into two disjoint classes:

$$G = \{t \in \Omega \mid \exists s\colon t(s) = t\}\,, \qquad E = \Omega - G\,.$$

Given a universe Ω, the syntactic analysis of a sentence s is equivalent to classifying members of $\gamma(s)$ into the set E or G. The evaluation of a syntactic disambiguation method is reduced to the evaluation of the related classifier. Note that several types of syntactic substructures (not only trees) can be modelled in a given universe Ω. Simpler syntactic relations (e.g. subject-verb, or noun-preposition-noun) are correct or wrong with respect to the source sentences and disambiguation is a classification into the class of correct or wrong ones (within the space of structures generated by a sentence). The notion of correctness of an elementary syntactic structure will be formalized in the next section.

Most disambiguation methods used by the corpus linguistics community cannot be considered categorical classifiers [34], since they produce a probability distribution over the ambiguous instances rather than selecting one class for each instance. In other words, corpus driven disambiguation methods assign confidence factors to competing candidates: these factors can be seen as probabilities of the correctness of the related interpretations. Decisions are generally undertaken only when one of the competing interpretations is significantly more confident than the others (as for example in [31]). In ambiguous sentences like *VN_prep_N* structures, as for example

watching girls with binoculars

competing readings

(1) (*girls-with-binoculars*) *and*

(2) (*to_watch-with-binocular*)

may have a computed confidence factor α_1 and α_2, respectively. Given this distribution $\{\langle 1, \alpha_1 \rangle, \langle 2, \alpha_2 \rangle\}$ we can normalize the α, thus obtaining a probabili-

ty distribution $\{\langle 1, P'(1)\rangle, \langle 2, P'(2)\rangle\}$, where $P'(1)$ (or $P'(2)$) is simply the confidence that the system assigns to the following statements: (1) (*or* (2)) *is the correct reading* that we will denote $P'(1 \in G)$ (or $P'(2 \in G)$).

In absence of any lexical or distributional knowledge about *to_watch*, *girl*, and *binocular* the blind confidence that an uninformed classifier can assign to (1) and (2) is simply

$$P(1 \in G) = P(2 \in G) = 0.5 .$$

With this notation P and P' represent respectively the prior (or blind) and posterior probability that a given syntactic structure is correct. A lexical acquisition system will perform as well on P' as P. $P'(t \in G)$ is thus expected to increase over the correct structures t ($t = (1)$ in the example) and to lower over wrong ones ($t = (2)$, in the example). Vice versa, $P'(t' \in E)$ is expected to increase over wrong structures t' because the classification is binary and $P'(E) = 1 - P'(G)$ as well as $P(E) = 1 - P(G)$. Before defining $P(G)$ and $P'(G)$ more formally, let us characterize the information gain in our lexical learning framework.

The *information gain* of an inductive task is defined in information theory as the average reduction in number of bits necessary to describe the correct classification/disambiguation. Our definition of *information score* for the PP disambiguation problem follows closely the general definition provided, for example, in [34].

Definition 1 (*Information score*). Given a sample space Ω, if the classifier performs a correct (*useful*) classification (i.e., $P'(G) > P(G)$ when $t \in G$ or $P'(E) > P(E)$ when $t \in E$) of t, then the *information score* (I) is

$$I = -\log P(G) + \log P'(G) ,$$

or (for a correct classification in E):

$$I = -\log P(E) + \log P'(E) .$$

If the corresponding decision is *misleading* (i.e., $P'(G) < P(G)$ when $t \in G$ or $P'(E) < P(E)$ when $t \in E$) then the information score (I) is a penalty, whose magnitude is:

$$-\log(1 - P(G)) + \log(-P'(G))$$

or (for a wrong classification in E):

$$-\log(1 - P(E)) + \log(1 - P'(E))$$

and thus the information score is

$$I = \log(1 - P(G)) - \log(1 - P'(G)$$
$$(\text{or } I = \log(1 - P(E)) - \log(1 - P'(E))) .$$

The overall performance index I_Σ over the test set Σ is the sum of the information scores I of all the testing cases averaged by their cardinality $|\Sigma|$.

The important aspect of this definition is that *it assigns to an inductive step a score that is as positive as the classification is correct and the complexity of the task was high, and a penalty as strong as the classification is incorrect and the complexity of the task is low.*

3.2. Syntactic disambiguation as a classification task

In order to evaluate the information gain that a given lexical acquisition algorithm provides, it is necessary to carefully model what we called *prior probability* ($P(G)$ or $P(E)$) in the previous sections. In the very specific perspective of the linguistic task at hand (i.e., syntactic disambiguation), the prior probability of *any* syntactic structure (not necessarily a tree) is simply the blind confidence that a system has in the correctness of the structure, without any further information. Correspondingly, the posterior probability is just the same confidence gained by virtue of some model of the test set (i.e., source corpus) or the syntactic structure itself (i.e., the semantics of its constituents).

3.2.1. Modelling prior probability

In order to define suitable notions of prior and posterior probability we must: (1) define a set of syntactic phenomena that can be observed; (2) establish the sample space in which the events are actually observed; (3) determine the prior probabilities of such events in the sample space: (4) interpret lexical disambiguation methods (e.g. [31]) as posterior probabilities in the sample space.

In corpus linguistics, it is a common practice to extrapolate the properties of a sublanguage $\mathscr{L}$, by analyzing a reference corpus $\mathscr{C}$. We can say that $\mathscr{C}$ embodies the model of use of $\mathscr{L}$. The majority of methods for automatic syntactic disambiguation, whose evaluation is the concern of this section, rely on the following hypothesis:

(H1) *The more we observe a phenomenon in the corpus $\mathscr{C}$ the more we can rely on the assertion that it is meaningful (= semantically plausible) for the related language $\mathscr{L}$. Vice versa: rare observations may be markers of inconsistency (or noise during the observation phase).*

The observed phenomena, in the majority of methods proposed in the literature, are word collocations, augmented by syntactic markers. The type of syntactic structures observed may vary significantly, but in general they are simpler than complete parse trees. For example, they are productive triples like subject-verb-object (SVO) or verb-prepositional modifier (V_prep_N). We denote these structures as *elementary syntactic structures* (*esss*).[16]

In order to evaluate the different methods we will rely on the set of structures that they observe from the corpus $\mathscr{C}$. This set will be our global sample universe Ω. If $ESS(s)$ denotes the set of elementary syntactic structures that can be

[16] Though it may be confusing, we use the term *ess* to indicate surface syntactic structures in a more general sense, that includes SVO, *esl*, and other surface structures defined in the literature.

observed (and are actually observed in ARIOSTO) in sentences of the corpus, then

$$\Omega = \bigcup_{s \in \mathscr{C}} ESS(s) . \tag{1}$$

Note that different sentences may generate the same *ess*. However, multiple versions of the same *ess* are different in Ω, as they are indexed by the source sentence s. Correctness is in fact a local property of an *ess* since it depends on the context. Therefore multiple occurrences of the same *ess* should be classified separately. Note also that (1) locates the performance evaluation methods within the sample space produced by an underlying grammatical model. The global ambiguity is thus a function of the adopted grammar. This restriction is necessary, but it applies uniformly to the different disambiguation methods being compared.

As a consequence of (1) the overall set of correct *esss*, which we denote with $G \subseteq \Omega$, is the union of the correct *esss* of each sentence s, i.e.,

$$G = \bigcup_{s \in \mathscr{C}} G_s \tag{2}$$

where G_s is the set of correct *esss* for the sentence s. Clearly the set of wrong (i.e., locally meaningless) syntactic structures E is simply given by:

$$E = \bigcup_{s \in \mathscr{C}} (ESS(s) - G_s) = \Omega - G . \tag{3}$$

As an example consider a corpus restricted to the single sentence

(s1) *the system acquires data from the satellite*

It follows that:

$$\begin{aligned} \Omega = ESS(s) = \{ \ & \text{(i) } (system,\ to\_acquire), \\ & \text{(ii) } (to\_acquire,\ data), \\ & \text{(iii) } (to\_acquire,\ from,\ satellite), \\ & \text{(iv) } (data,\ from,\ satellite)\} \end{aligned}$$

where $G = \{(\text{i}), (\text{ii}), (\text{iii})\}$ and $E = \{(\text{iv})\}$.

Note that this classification is local to (s1). In fact in the sentence *the program processes data from satellite*, (iv) would belong to the class G of correct *esss*.

In order to derive a prior probability for the different *esss* we can simply count the number of correct *esss* over the cardinality of the whole sample universe Ω. This process however cannot be replicated for all the sentences of a real corpus, where noisy unsupervised learning is performed. In fact we simply do not know a priori which *esss* are correct.

A better solution for modelling the prior probability would be simply to approximate the average number of correct *esss* in a sentence. Each sentence produces one or more collision sets, that is, groups of structures that cannot belong to the same reading of the sentence. Trivial collision sets are single, non-ambiguous structures. In (s1), the elementary syntactic structures (iii) and (iv) form a nontrivial collision set. For each sentence s, $ESS(s)$ can be partitioned

into the set of its collision sets that will be denoted by $ESS(s)/\rho$.[17] In the example, the elements of $ESS(s)/\rho$ are the following collision sets:

$$C1 = \{(\mathrm{i})\} = \{(system, acquires)\},$$
$$C2 = \{(\mathrm{ii})\} = \{(acquires, data)\},$$
$$C3 = \{(\mathrm{iii}), (\mathrm{iv})\} = \{(acquires, from, satellite), (data, from, satellite)\}.$$

Let us now model the prior probability $P(G)$ of the correct *ess* as follows:

$$P(G) = \frac{\sum_{s \in \mathscr{C}} P(G_s)}{|\mathscr{C}|}. \tag{4}$$

$P(G_s)$ is the probability of elementary syntactic structures being locally correct. This value is dependent on the collision sets that are generated from the sentence s.

Let us point out that in general, there is only one correct *ess* (i.e., $\in G_s$), in each collision set. This hypothesis allows the following definition for $P(G_s)$:

$$P(G_s) = \frac{|ESS(s)/\rho|}{|ESS(s)|} = \frac{\#\text{ of collision sets}}{\#\text{ of } esss}. \tag{5}$$

In the example $P(G) = 3/4$ is the (blind) probability that any elementary structure *ess* is correct, given $\mathscr{C} = \{s\}$. Correspondingly,

$$P(E_s) = 1 - P(G_s). \tag{6}$$

The definition of $P(G)$ clearly follows from the observation that *esss* of different sentences are disjoint, i.e.,

$$ESS(s) \cap ESS(s') = \{\emptyset\} \quad \forall s \neq s'.$$

Thus, $P(G \mid \Omega) = P(G)$ can be derived from each $P(G_s)$ as follows:

$$P(G) = \sum_{s \in \mathscr{C}} P(G_s) P(s \mid \mathscr{C}) = \frac{1}{|\mathscr{C}|} \sum_{s \in \mathscr{C}} P(G_s). \tag{7}$$

3.2.2. Modelling posterior probability

The posterior probability is the probability of being in a class (G_s or E_s), as assigned by the trained system to the test instances (i.e., *esss*) of a sentence s. All corpus driven methods for automatic disambiguation assign some probability driven score to extensive collections of data, i.e., syntactic patterns extracted from the raw texts. In order to use the *information gain* as a performance index of these methods we must express in a more appropriate form the preference scores

[17] ρ is used to indicate the equivalence relation that holds between conflicting syntactic structures. The collision sets in fact are the elements of the quotient set, $ESS(s)/\rho$. More details about relation ρ can be found in [2–4, 9].

defined for each method. The problem here is that each syntactic disambiguation operator in the literature has a probabilistic flavour *but is not a probability*. Yet this is essential for the evaluation of the information gain.

Given any disambiguation score σ, the posterior probability P'_σ derived from the preference function σ, i.e., the probability of correctness according to σ, should obey the following condition:

$$\sum_{ess \in coll} P'_\sigma(ess \in G_s \mid coll) = 1 \quad \forall coll \in ESS(s)/\rho \tag{8}$$

where *coll* denotes a collision set for the sentence s.

In fact the global probability of being correct within a given collision set *coll* is 1. Note that $ESS(s)/\rho$ is a partition of $ESS(s)$ so that the collision set *coll*(*ess*) to which a given *ess* belongs is unique. Within a collision set, the syntactic disambiguation scores provide a redistribution of preference among the members. For this reason the probability that a given *ess* is correct in a sentence s is given by:

$$\begin{aligned} P'_\sigma(ess \in G_s) &= \sum_{coll \in ESS(s)/\rho} P'_\sigma(ess \in G_s \mid coll) P'_\sigma(coll \mid s) \\ &= P'_\sigma(ess \in G_s \mid coll(ess)) P'_\sigma(coll(ess) \mid s) \end{aligned} \tag{9}$$

since *ess* belongs only to the collision set *coll*(*ess*), and thus

$$P'_\sigma(ess \in G_s \mid coll) = 0 \quad \forall coll \neq coll(ess) .$$

A definition of $P'_\sigma(ess \in G_s \mid coll(ess))$ that applies to any disambiguation score σ is the following:

$$P'_\sigma(ess \in G_s \mid coll(ess)) = \frac{\sigma(ess)}{\sum_{ess' \in coll(ess)} \sigma(ess')} . \tag{10}$$

$P'_\sigma(coll(ess) \mid s)$ is the probability that *coll*(*ess*) is correct in a sentence (i.e., it includes at least one correct *ess*), thus implying:

$$P'_\sigma(coll(ess) \mid s) = 1 . \tag{11}$$

For example let us assume that a hypothetical disambiguation score σ computed the following scores for the example (s1):

$$\sigma((to\_acquire, from, satellite)) = 0.5 ,$$
$$\sigma((data, from, satellite)) = 0.1 .$$

Using (9), (10), (11), the posterior probabilities are computed as follows:

$$P'_\sigma((system, to\_acquire) \in G_s) = \frac{\sigma((system, to\_acquire))}{\sigma((system, to\_acquire))} P(C1 \mid s) = 1 ,$$

$$P'_\sigma(to\_acquire, data) \in G_s) = \frac{\sigma((to\_acquire, data))}{\sigma((to\_acquire, data))} P(C2 \mid s) = 1 ,$$

$$P'_{\sigma}((to\_acquire, from, satellite) \in G_s)$$
$$= \frac{\sigma((to\_acquire, from, satellite))}{\sigma((to\_acquire, from, satellite)) + \sigma((data, from, satellite))} P(C3 \mid s)$$
$$= \frac{0.5}{0.5 + 0.1} \cong 0.83 .$$
$$P'_{\sigma}((data, from, satellite) \in G_s)$$
$$= \frac{\sigma((data, from, satellite))}{\sigma((data, from, satellite)) + \sigma((data, from, satellite))} P(C3 \mid s)$$
$$= \frac{0.5}{0.5 + 0.1} \cong 0.17 .$$

In this example the role of the information gain as a measure of the improvement that P' implies over P can be easily seen. Note that the prior probability $P(ess \in G) = |\{C1, C2, C3\}|/(|C1| + |C2| + |C3|) = 0.75$. The distribution P' assigns to both the *ess*s the following scores:

$$P'((to\_acquire, from, satellite)) = 0.83 .$$
$$P'((data, from, satellite)) = 0.17 .$$

As a result both decisions are useful ($P'(C) > P(C)$ for the correct classification in C) and the following values are obtained:

$$I = -\log(P((\text{iii}) \in G)) + \log(P'(\text{iii}) \in G)) = \log(0.83/0.17) = 2.3 ,$$
$$I = -\log(P((\text{iv}) \in E)) + \log(P'(\text{iv}) \in E)) = \log(0.83/0.17) = 2.3 .$$

(9), (10) and (11) allow to model *any disambiguation method*, given that an appropriate notion of collision set is defined. In the following, we evaluate the *lexical association* (*LA*) [31] and the *semantic expectation* (*SE*) (see Section 2.3)) modelled in terms of probability distributions over sets of training instances. Clearly, the applicability of this framework is not limited to these methods, since (9), (10), and (11) can be easily extended to other corpus-based language learning algorithms.

3.3. Posterior probability based on lexical association

The lexical association (*LA*) is a preference score introduced in [31] as an extension of the more classical *t*-score measure [48]. *LA* is used in ambiguous sentence frameworks like

verb dir_obj prep noun (12)

to select the correct referent of the PP constituent (*prep noun*). If the score is (significantly) positive, preference is given to the verb attachment (*verb prep noun*); when negative, it suggests the opposite attachment (*dir_obj prep noun*).

To summarize, we will recall the definition that models syntactic preference in sentences like (12). The LA value is defined [31] by:

$$LA(verb\ \ dir\_obj\ \ prep) = \log_2 \frac{P(prep \mid verb)P(NULL \mid dir\_obj)}{P(prep \mid dir\_obj)} \tag{13}$$

where $P(NULL \mid dir\_obj)$ is the probability of observing no prepositional modifier for *dir_obj*. In this form, (13) is useful only within ambiguous frame sentences like (12).[18] In order to extend its coverage to other ambiguous structures within the test set, we must model also sentence structures like

$$word_1 \ldots [prep_1]\, word_2 \ldots [prep_2]\, word_3 \ldots prep_3\, word_4 \tag{14}$$

(i.e., chains of prepositional modifiers). Since in these cases alternative readings may give rise to more than two referents, we cannot use a single LA value. According to the definition (13), we will express a modified preference for a referent as follows. This preference score, which is slightly different from LA, fits our experimental purposes, and will be denoted by σ_{LA}. When the referent of a post modifier, i.e., ($prep_3$ $word_4$), is the closest word, i.e., $word_3$, then the preference score follows the philosophy used for the denominator in (13):

$$\sigma_{LA}(word_3, prep) = P(prep \mid word_3) \tag{15}$$

else, the score is given, as in the numerator in (13), by:

$$\sigma_{LA}(word_i, prep) = P(prep \mid word_i)P(NULL \mid word_3)\,, \quad i \neq 3 \tag{16}$$

where $word_i$ is the preferred referent in (14). This definition is a close approximation of the LA, and is appropriate for our evaluation purposes. When the preference score σ_{LA} is fully defined, the associated posterior probability can be derived as follows, by the use of (10).

Definition 2 (*Posterior probability distributions based on lexical association*). Given a sentence s, any syntactic structure $ess(w_1, p, w_2) \in ESS(s)$ has a posterior probability based on LA, $P'_{LA}(ess(w_1, p, w_2) \in G_s)$, given by:

$$\begin{aligned} P'_{LA}(ess \in G_s) &= P'_{LA}(ess \in G_s \mid coll(ess))P(coll(ess) \mid s) \\ &= \frac{\sigma_{LA}(w_1, p)}{\sum_{ess(w,p,\_) \in coll(ess)} \sigma_{LA}(w, p)} \end{aligned} \tag{17}$$

where $coll(ess)$ is the actual collision set of $ess(w_1, p, w_2)$, and _ stands for any word.

[18] The authors derive $P(prep \mid verb)$, $P(NULL \mid dir\_obj)$, $P(prep \mid dir\_obj)$ from partial parses of the source corpus, that are manually validated.

3.4. Posterior probability based on semantic expectation

The disambiguation method based on the semantic expectation is guided by the case-based lexicon acquired by ARIOSTO_LEX. Syntactic relations are validated according to their semantic expectation in the lexicon. In Section 2.4 we described a typical lexical entry acquired from the corpus by:

$$pre\_sem\_lex(word, conceptual\_relation, semantic\_class, direction, esl\_type, prep, SYNT, SE, CF)\,. \tag{18}$$

The semantic expectation (SE) is the probability that *conceptual_relation* between *word* and any word belonging to *semantic_class* occurs in the corpus (Section 2.4). In the following, we will denote the SE factor in (18) as

$$SE(word, prep, conceptual\_relation, semantic\_tag)\,. \tag{19}$$

Given a sentence s, and a collision set $coll(ess)$ derived from an ambiguous PP referent, as in (12) or (14), the preference score based on the semantic expectation of any elementary syntactic structure in $coll(ess)$, is defined as follows:

$$\begin{aligned} &SE(ess(word_i, prep, word_j)) \\ &\quad = \max_{C_k \supseteq word_k} (SE(word_i, prep, conc\_rel, C), SE(word_j, prep, conc\_rel, C_i))\,, \\ &\quad k = i, j\,, \end{aligned} \tag{20}$$

where $ess(word_i, prep, word_j) \in coll(ess)$, and $\{C_k\}$ is the set of all supertypes (i.e., semantic tags) of $word_k$.

Note that the algorithm in (20) considers all the interpretations (i.e., conceptual relations) of the ess, according to the preposition *prep* and the possible generalizations of $word_i$ and $word_j$. Clearly, when no semantic interpretation is found for a given ess, SE is 0.

According to (10), the posterior probability associated with the SE score is correspondingly derived:

$$\begin{aligned} &P'_{SE}(ess(w_i, prep, w_j) \in G_s) \\ &\quad = P'_{SE}(ess(w_i, prep, w_j) \in G_s \mid coll(ess)) P'_{SE}(coll(ess) \mid s) \\ &\quad = \frac{SE(ess(w_i, prep, w_j))}{\sum_{ess' \in coll(ess)} SE(ess')} \end{aligned} \tag{21}$$

where again $coll(ess)$ is the collision set of $ess(word_i, prep, word_j)$.

3.5. Evaluation results

We selected a test set and evaluated the accuracy and information score derived by using LA, and SE in the disambiguation phase.

The test corpus (which we used also in other evaluation experiments) Σ includes 232 complex sentences from which 7871 parse trees have been derived (an average of 33 parse trees per sentence). Sentences belonged to the legal domain in Italian. The Italian grammar adopted in the test is a definite clause grammar, including 103 rules. In the 7871 parse trees, 4117 elementary syntactic structures have been detected. The classes of elementary syntactic structures of interest are the following: *N_prep_N* (noun-preposition-noun), *V_prep_N* (verb-preposition-noun), *V_N* (verb-noun), *N_V* (noun-verb), *N_nil* (noun-NULL), *V_nil* (verb-NULL), ... The test set has been manually validated (i.e., class E and G have been defined for the collision sets of each sentence) by three human judges. The number of correct (according to the judges) parses approximates the number of sentences (there is an average of 1.3 correct trees per sentence), while the number of correct elementary syntactic links (with repetitions) is 1540. The prior probability of any *ess* being correct in the learning corpus is, according to (5) and (7), 0.66.[19] Fig. 3 provides examples of sentence fragments, and the related collision sets. The fragment of each sentence enclosed in brackets represents the segment that originates the collision set. Bold characters indicate the ambiguous PP, and Italic characters indicate the words that compete for the

---

1. Examples of Simple Collision sets:

1.1 Minimal Attachment:

su richiesta del ministro per le finanze , il [(*servizio* di *vigilanza* sulle *aziende)* **di credito**] (* *service of control of agencies of credit*) controlla l esattezza delle attestazioni contenute nel certificato .'

```
meas([g_N_prep_N(2,azienda,di,credito)],0.945)          %agency-of-credit
meas([g_N_prep_N(4,vigilanza,di,credito)],0.000)        %control-of-credit
meas([g_N_prep_N(6,servizio,di,credito)],0.0006)        %service-of-credit
```

1.2 Non-Minimal Attachment

i sostituti d imposta devono [(*presentare* la *dichiarazione* di_cui_a quarto *comma* dell *articolo* 9 , *relativamente* ai pagamenti fatti e agli utili *distribuiti* nell anno 1974) **entro il 15_aprile_1975**]. (* *must present the declaration of which at comma 4th of item 9, relatively to the payment done and the profit distributed in the year 1974, within april 15, 1974*)

```
meas([g_N_prep_N(17,articolo,entro,x_15_aprile_1975)],0.000)          %item-wihtin-April_15th
meas([g_V_prep_N(7,distribuire,entro,x_15_aprile_1975)],0.166)        %to_distribute-within-April_15th
meas([g_Adv_prep_N(14,relativamente,entro,x_15_aprile_1975)],0.000)   %relatively-within-April_15th
meas([g_N_prep_N(19,comma,entro,x_15_aprile_1975)],0.000)             %comma-within-April_15th
meas([g_N_prep_N(22,dichiarazione,entro,x_15_aprile_1975)],0.107)     %declaration-within-April_15th
meas([g_V_prep_N(24,presentare,entro,x_15_aprile_1975)],0.166)        %to_present-within-April_15th
```

2. ComplexCollision Set (i.e. non singletons members)

gli organi del tributario possono dichiarare non dovute le pene pecuniarie quando la violazione [e' *giustificata* da obiettive *condizioni* di *incertezza* **sulla portata e sull ambito**] (* *it is justified by objective conditions of uncertainty on the scope and ambit*) di applicazione delle disposizioni alle quali si riferisce.

```
meas([g_N_prep_N(2,incertezza,su,portata),g_N_prep_N(5,incertezza,su,ambito)],0.181)
                  %uncertainty-on-scope              %uncertainty-on-ambit
meas([g_N_prep_N(4,condizione,su,portata),g_N_prep_N(7,condizione,su,ambito)],0.012)
                  %condition-on-scope                %condition-on-ambit
meas([g_V_prep_N(7,giustificare,su,portata),g_V_prep_N(10,giustificare,su,ambito)],0.000)
                  %to_justify-on-scope               %to_justify-on-ambit
```

Fig. 3. Examples of collision sets in the test set.

[19] In other corpora we found slightly different values, 0.60 for a domain of ecological newspaper articles, and 0.57 for the RSD in English.

PP attachment. A word-by-word English translation of the segment is also provided. For each competing *esl* in a collision set, the semantic expectation is reported.

Example 1.1 is a straightforward example of colliding *esl*s of the same type. In example 1.2, several *esl*s of different types collide (*N_prep_N*, *V_prep_N*, *Adv_prep_N*). Notice that *esl*s 2 and 5 in the collision set 1.2 have the same expectation. In this case, a simple heuristic is to select the nearest pair of words (e.g. *distribute within April* 15*th*). The first argument of each *esl* is in fact the distance in words between the two co-occurring words.

Finally, example 2 shows the collision sets created by a prepositional phrase including a coordination (*on the scope and ambit*). In this case, the alternative interpretations are represented by pairs of *esl*s.

To run the experiment, we acquired with ARIOSTO_LEX a lexicon of 961 word senses. We did not use the limbo lexicon, but only the selectional restrictions *pre_sem_lex* (18). These data have *not* been supervised by a human analyst. Furthermore, the test set was not incorporated into the learning corpus. (See Table 3.)

The first row shows the information score I_Σ obtained by the two disambiguation methods for the collision sets. Both methods perform rather well, as shown by the good information gain. To make a comparison, in [34], the information gain of rather less complex classification tasks (e.g. with a prior probability of the most probable class not higher than 45%, and with *supervised* learning algorithms) is about 1.

Though this is not shown in the table, I_Σ values are not sensitive to the complexity of the test set. In fact, we did not observe sensitive changes in the I_Σ values for groups of sentences with a number of trees between 2 and 30, from 30 to 90, and over 90. Rather, I_Σ is sensitive to the complexity and the dimension of the learning set. We obtained growing performances for both methods when increasing the learning corpus from 200,000 to 500,000 words. This is intuitive, since all PP disambiguation methods perform better as they gain evidence of linguistic patterns. It is our future objective to observe more closely these dependencies, by experimenting on different linguistic domains, and with fixed domains of growing dimensions (possibly, well over a million words).

Table 3
Experimental results: information score versus accuracy

	SE	*LA*
I_Σ on all *ess*	0.203	0.174
I_Σ on correct *ess*	0.748	0.673
Accuracy over *ess*	68.6%	61.4%
I_Σ averaged on sentences	0.38	0.20

Row 3 provides the accuracy[20] of the two methods at selecting the correct *ess* in each collision set according to a simple maximum likelihood method.

The reader may notice that the accuracy values are relatively low for both methods, if compared with other performance figures reported in the literature. Hindle and Rooths for example report a 78.1% accuracy in their paper. The first reason is that our learning set is of 500,000 words rather than millions of words. Second neither preference score (i.e., *LA* and *SE*) provides any model of many ambiguity phenomena (e.g. anaphora, adjectival or adverbial PP referents). Suitable models of such phenomena should result in higher accuracy values. Another important reason is that the *LA* method has been only tested on *V N prep N* sentences. In our experiment we considered several types of ambiguity, including multiple prepositional phrases and coordination (see the examples of Fig. 3). Our point is that we wish to test the system over *real cases*. *V N prep N* phrases are only a small fragment of the possible ambiguities, and in addition, it would have been necessary to extract by hand these phrases from the sentences in the corpus, in order to build the test set. Rather, we let our DCG grammar run over the corpus, and we retained all the sentences for which the grammar could successfully complete the analysis. Then, we analyzed these sentences by hand, we rejected the sentences for which we judged that the grammar did not produce the complete set of parses, and we marked the (semantically) correct parse for the remaining (232) sentences. This task was facilitated by a graphic interface. When we analyzed the results produced by our disambiguator over the set we noticed that, in many cases, the confidence that the system gained about colliding *esl*s does not allow a reliable choice. Many *esl*s have a very close *SE* value, like (*distribuire*, *entro*, *aprile*) (i.e., (*to_distribute*, *within*, *april*)) and (*presentare*, *entro*, *aprile*) (i.e., (*to_present*, *within*, *april*)) in the second example of Fig. 3. In the example, the two values are the same, therefore the choice is performed on a "nearest best" basis. When the values are close, but not the same, we use the maximum likelihood method. Many methods use a threshold under which the system is prevented from taking unreliable decisions. However, we experimentally observed that this criterion would significantly reduce the number of useful choices. Once again, we think that the accuracy is not an adequate measure. A more "reliable" approach would be to retain all the ambiguous *esl*s for which $P'(esl \in G)$ is higher than $P(G)$, rather than crudely selecting the "most probable" *esl*. We used accuracy in Table 3 only for the sake of uniformity with the preceding literature.

Given the complexity of the task, we are rather happy with our 68.7% accuracy and 0.74 information gain. Though the over-generalized tags in the lexicon may induce the system to accept some erroneous syntactic structure, the combined effect of probabilistic and semantic filters produces very acceptable performances. Consider also that the testing conditions are particularly severe, because the test set was not included into the learning set and the limbo lexicon was not consulted.

[20] Since the methods are forced to make always a choice, accuracy and precision are here the same.

This means that the system might have not acquired the selectional restrictions necessary to accept some syntactic structure (and in fact we manually verified that this was often the cause of errors).

Finally, this performance evaluation does not consider that ARIOSTO_LEX does much more than filtering out wrong parses: it provides a *semantic interpretation* of the accepted patterns. But we already remarked that testing the quality of the semantic information in a lexicon is a very complex matter, for which no methods have been proposed in the literature so far.

4. Related research

In the introduction we provided a general account of corpus-based methods in computational linguistics. For sake of completeness, in this section we summarize the literature on the design of a computational lexicon, that is more closely related to the system presented in this paper.

Several research groups have been recently engaged in the challenging objective of acquiring an unrestricted semantic lexicon for NLP, using machine readable dictionaries (MRDs) and corpora (MRC) as sources. Among these, lexicon acquisition methods based on MRDs are a majority. In [14,21,37,41] taxonomic and structural patterns of words are acquired from MRD (e.g. "*tax*: *a payment imposed upon persons or groups for governmental support . . .*").

Usually, the dictionary used for acquisition is LDOCE, an on-line dictionary that has very desirable features for automatic acquisition, like for example lexical templates of verbs (e.g. *takes NP NP*), or the use of a restricted grammar to describe word senses. The information that can be reliably extracted from LDOCE is mostly syntactic and in part semantic. For example, a word taxonomy can be generated using the genus information included in definitions. However, Sanfilippo and Poznanski [42] remark that the genus of over 20% of verb senses is one of 8 verbs (*cause*, *make*, *be*, *give*, *put*, *take*, *move*, *have*), but many verbs with the same genus belong in fact to different semantic classes. The paper presents an alternative method to correlate word senses across MRDs that is computer assisted, i.e., it requires a constant interaction with a linguist. Methods have also been proposed to process the definition of words, as for example in [13] taking advantage of the simplified grammar used in LDOCE for definitions.

In [24], a system, ULTRA, is described for extracting lexical entries from an MRD, using natural language processing and heuristic techniques. Basically, the semantic information associated with each content word is a list of pragmatic and semantic constraints like:

entity(*bank*4.1, *class*, *countable*, *institution*, *abstract object*, *economics*, *banking*).

This information is extracted in part automatically from the dictionary, in part entered manually. Even in [26] the method for acquiring lexical entries relies largely on human work.

None of the aforementioned papers provide a performance evaluation, since their tools seem to be conceived primarily to help lexicographers.

Very few papers build a semantic lexicon using corpora as a source of word sense information. In [45] verbal case frames are acquired from bilingual corpora. An example is:

(*buy*: (*agent HUMAN*), (*object CONCRETE, ABSTRACT*),
(*for HUMAN*)).

Semantic tags (*HUMAN*, etc.) are selected from an on-line thesaurus, and case labels are the detected syntactic relations (*prepositions*, *agent* = *subject*, *object*). In many cases, syntactic and semantic ambiguities are solved by comparing the two languages. Though promising, the approach produces reliable results with a very limited coverage: bilingual feature descriptions have been obtained only for 16 verbs.

One generally important problem with MRD-based lexical acquisition techniques is that, though the definitions in dictionaries are somehow deeper than simple selectional restrictions (i.e., they include structural and taxonomic information), the latter are in practice a more easy-to-understand, and useful, type of semantic knowledge for the purpose of automatic NLP. Structural and selectional patterns are both relevant types of information for computational lexicons, and we think that at some point research on MRD and MRC should be integrated.

5. Conclusions and future work

The thesis of this paper was to demonstrate that, by combining empirical and rationalist methods, it is also possible to combine the major advantages of these two radically different approaches, that is, scaling up and digging deep.

We presented ARIOSTO_LEX, a two-step algorithm for the acquisition of domain appropriate selectional restrictions from corpora. ARIOSTO_LEX has a utility as a self-standing tool, since it provides in a very readable and compact form linguistic data amenable to a comparative analysis of sublanguages, and to a systematic analysis of complex lexical categories, like verbs. ARIOSTO_LEX however was not conceived as a tool for lexicographic studies, but primarily as a computational tool, that could be used in any NLP system without any strong commitment on the designer of the final application. The information acquired by ARIOSTO_LEX is a dynamic knowledge base of lexical facts with respect to selected corpora. For each lexical entry, ARIOSTO_LEX provides an account of all the situations in which a word can participate in a given domain, expressed by selectional restrictions. This information is clearly only one of the desirable features of a semantic lexicon, yet no existing industrial or research NLP projects could demonstrate an adequate coverage of this type of lexical data, and portability to different languages and domains. In particular, ARIOSTO_LEX

could be profitably used for *tuning an existing lexicon*, by adding domain specific case relations to existing lexical items. The potential interest of one such computational tool needs not to be stressed, since one could say that the grand challenge to natural language processing at the moment is that of systematic and reliable linguistic knowledge acquisition on a large scale, which we take to include lexical acquisition as the major part, since most intelligent applications are now lexicon driven.

In this paper, we performed a rather detailed analysis of the lexical entries produced by ARIOSTO_LEX. In addition to overwhelming linguistic material for a systematic study of sublanguages, we gained experimental evidence that has an impact on the applicability of some popular approaches to automatic language processing:

- The relational structure of verbs is highly variable and poorly overlapping. Finding the common invariants of these structures (i.e., a type hierarchy), is a task that has inherent limitations.
- Most verbs impose weak expectations on their argument structures. This finding has a problematic impact on the validity of expectation driven semantic interpreters.
- The relational structure of words varies significantly across sublanguages. General purpose approaches to lexicon design seem inappropriate, if the lexicon is to be used by automatic language processors.

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. We demonstrated that selectional restrictions do not generalize across sublanguages, and acquiring them by hand is often an unintuitive and very time-consuming task.

The limitation is that the conceptual types used to generalize selectional patterns are very high level, and in some case may not provide adequate selectional power. On the other hand, using more refined conceptual types would beg the question of automatic lexical acquisition. The performance evaluation section, however, demonstrated that in general the discriminating power of the lexicon is good, despite the generality of the semantic tags used.

Another, in our view more serious, limitation is verb polysemy. In fact, while for nouns the high-level categories are "good enough" for discriminating most ambiguous senses, verbs are classified in few over-general categories. These categories are not sufficient to discriminate among the subtle and highly polysemous senses of verbs. Hence, different senses of the same verb are collapsed into a unique lexical entry. This issue was explored in [6,10,11], where we presented CIAULA, a corpus driven conceptual clustering method for word classification, which classifies in different categories polysemous words. The results of this algorithm are, as expected, more problematic for verbs, because of their highly variable, bushy, relational structure. Because of this inherent difficulty, the advantages of a more refined verb classification in ARIOSTO_LEX are not fully clear. We believe that more insight is needed into this complex matter.

Finally, an open issue is an extensive on-field evaluation of ARIOSTO_LEX. We proposed a formal evaluation that was concerned only with syntactic

disambiguation, while a more appropriate evaluation should consider the interpretative power of the acquired selectional restrictions within some NLP-based application. In Section 3, we pointed out the difficulties of a more substantial evaluation of ARIOSTO_LEX, and of computational lexicons in general.

Appendix A. C_i-*synt_rel*-C_j tables

Legenda[21]

CD (*see Table A.1*)
- PA: PHYSICAL_ACT (*to plough*),
- MA: MENTAL_ACT (*to organize*),
- ART: ARTEFACT (*table*),
- HE: HUMAN_ENTITY (INDIVIDUAL + ORGANIZATION) (*customer*),
- V: VEGETABLE (*corn*),
- B: BUILDING (*greenhouse*),
- BP: BY_PRODUCT (*milk*),
- MT: MATTER (*iron*),
- AN: ANIMAL (*cow*),
- MC: MACHINE (INSTRUMENTALITY) (*grindstone*),
- P: PLACE (LOCATION) (*beach*).

LD (*see Table A.2*)
- A: ACT (*to enclose*),
- RE: REAL_ESTATE (*greenhouse*),
- G: GOODS (*table*),

Table A.1
C_i-*per*-C_j table, commercial domain (CD)

per	PA	MA	ART	HE	V	B	BP	MT	AN	MC	P
PA	0.121	0.075	0.031	0.022	0.000	0.039	0.002	0.002	0.001	0.010	0.004
MA	0.041	0.054	0.023	0.018	0.000	0.023	0.001	0.000	–	0.005	0.003
ART	0.094	0.068	0.045	0.026	0.001	0.057	0.005	0.005	0.000	0.011	0.004
HE	0.016	0.024	0.002	0.024	0.000	0.014	0.000	0.001	–	0.001	0.001
V	0.003	0.001	0.002	0.000	–	0.000	–	–	–	0.001	0.000
B	0.061	0.030	0.012	0.010	–	0.013	0.000	0.000	–	0.001	0.001
BP	0.013	0.004	0.006	0.001	0.000	0.004	0.001	0.001	0.001	0.002	0.001
MT	0.027	0.008	0.011	0.001	0.000	0.015	0.002	0.002	–	0.003	0.001
AN	0.000	0.001	–	–	–	0.000	–	–	–	–	0.000
MC	0.032	0.036	0.010	0.002	0.001	0.033	0.001	0.001	–	0.004	0.001
P	0.004	0.004	0.000	0.000	–	0.001	–	0.000	–	–	0.000

[21] ARIOSTO was first implemented on the two Italian domains, for which an automatic tagger was not available. WordNet tags were adopted later for the RSD domain. This explains why we used in the Italian domains tag labels that do not correspond to WordNet labels. When two labels identify exactly the same category, the corresponding WordNet label is between brackets.

Table A.2
C_i-*per*-C_j table, legal domain (LD)

per	A	RE	G	AM	D	ABS	TE	HE	P	S
A	0.211	0.020	0.021	0.071	0.031	0.117	0.055	0.036	0.007	0.019
RE	0.004	0.002	0.001	0.001	0.001	0.003	0.002	0.001	–	–
G	0.012	0.003	0.002	0.002	0.002	0.001	0.003	0.001	0.002	–
AM	0.033	0.004	0.008	0.016	0.004	0.016	0.012	0.006	–	0.001
D	0.032	0.002	0.001	0.007	0.008	0.015	0.007	0.010	0.001	0.004
ABS	0.030	0.002	0.002	0.010	0.005	0.018	0.010	0.011	0.001	0.005
TE	0.011	–	0.001	0.004	0 001	0.005	0.002	0.002	–	–
HE	0.036	0.010	0.005	0.007	0.006	0.015	0.004	0.006	–	0.001
P	0.002	0.001	–	0.001	–	0.001	–	0.001	0.001	0.001
S	0.048	0.004	0.009	0.018	0.006	0.025	0.020	0.010	0.001	0.006

- AM: AMOUNT (*income*),
- D: DOCUMENT (*contract*),
- ABS: ABSTRACTION (*definition*),
- TE: TEMPORAL_ENTITY (*year*),
- HE: HUMAN_ENTITY (INDIVIDUAL + ORGANIZATION) (*company*),
- P: PLACE (LOCATION) (*beach*),
- S: STATUS (*obligation*).

RSD (*see Table A.3*)

- DS: SCIENTIFIC DISCIPLINE (*geography*),
- CO: COGNITIVE PROCESS (*evaluation*),
- ART: ARTIFACT (*image*),
- TE: TEMPORAL ENTITY (*month, day*),
- ORG: ORGANIZATION, SOCIAL GROUP (*university*),
- INST: INSTRUMENTALITY (*satellite*),
- LOC: LOCATION (*Oregon*),

Table A.3
C_i-*for*-C_j table, remote sensing domain (RSD)

for	DS	CO	ART	TE	ORG	INST	LOC	PR	IND	NO	ABS	ATTR	Act	MO	MT	PS
DS	0.002	0.001	0.001	–	–		0.003	0.002	0.001	0.002	0.002	–	0 002	0.002	–	–
CO	0.019	0.060	0.025	0 006	0.004	0.028	0 019	0.016	0.015	0.021	0.042	0.030	0.092	0.041	0.004	0.001
ART	0.028	0.039	0 010	0.001	0 011	0 011	0.010	0.005	0 004	0 011	0.024	0.006	0.052	0.016	0 001	–
TE	–	0 001	0 002		0.001	0.001	0.004	0 002	0.002	0.002	0 002	0.002	0.004	0.002	–	–
ORG	0.002	0 010	0 004		0.001	0 002	0.001	0.001	0 001	0 002	0.007	0.002	0.012	0.003	–	
INST	0.009	0 015	0 007	0 002	0 002	0 011	0.009	0.003	0.006	0.009	0.011	0 007	0 023	0 005	–	–
LOC	0.001	0.003	0.002	0.002	0 001	0 006	0.005	0.002	0.005	0.002	0.003	0.004	0.009	0 002	0.001	–
PR	–	0.007	0 002	0.003		0.002	0.003	0.007	0 001	0.002	0.011	0.006	0.011	0.003	–	–
IND	0 002	0.006	0 001	0.002	-	0.002	0 001	0 001	0 001	0.002	0.002	0.001	0.005	–	–	–
NO	-	0.004	0.001	0 001	-	0.002	-	-	-	0 002	0.005	–	0.003	0.004	–	–
ABS	0.002	0.006	0 008	0.003	0.003	0.010	0.005	0 005	0 005	0.005	0 014	0.007	0.016	0 011	–	0.001
ATTR	0.002	0 009	0 004	0.001	0 002	0 004	0.002	0.005	0.002	0.001	0.010	0 005	0.013	0.002	0.001	–
Act	0 003	0.018	0.005	0.002	0 001	0.007	0 008	0 010	0.005	0.007	0.012	0 008	0.027	0 009	–	–
MO	0.001	0.015	0 006	0.002	0 005	0 004	0.011	0.009	0.003	0.007	0.010	0.006	0 016	0.007	0.001	–
MT	-	–	–		–	-	-	–	–	–	–	–	–	–	–	–
PS	–	0.001	0.002	0.002			0 002	0 002	-	0.001	0.002	–	0.002	–	–	–

- PR: PROPERTY (*emissivity*),
- MT: MATTER (*iron*),
- IND: INDIVIDUALS (*researcher, chief*),
- NO: NATURAL OBJECT (*mountain, sea*),
- ABS: ABSTRACTION (*data, model*),
- ATTR: ATTRIBUTE (*smooth, raw*),
- Act: ACT, HUMAN ACT (*to manage*),
- MO: MENTAL OBJECT (*idea, project*),
- PS: PROCESS, NATURAL EVENT (*earthquake*).

References

[1] H. Alshawi, Qualitative and quantitative designs for speech translation, the balancing act: combining symbolic and statistical approaches to language, in: *Proceedings ACL Workshop*, Las Cruces, NM (1994).

[2] R. Basili, M.H. Candito, M.T. Pazienza and P. Velardi, Evaluating the information gain of probability-based PP-disambiguation methods, in: *Proceedings International Conference on New Methods in Language Processing*, Manchester (1994).

[3] R. Basili, F. Grisoli and M.T. Pazienza, Might a semantic lexicon support hypertextual authoring, in *Proceedings Applied Natural Language Processing*, Stuttgart (1994).

[4] R. Basili, A. Marziali and M.T. Pazienza, Modelling syntax uncertainty in lexical acquisition from texts, *J. Quant. Linguist.* **1** (1) (1994).

[5] R. Basili, M.T. Pazienza and P. Velardi, A shallow syntactic analyzer to extract word associations from corpora, *Literary Linguist. Comput.* **7** (1992) 114–124.

[6] R. Basili, M.T. Pazienza and P. Velardi, Hierarchical clustering of verbs, in: *Proceedings ACL-SIGLEX Workshop on Lexical Acquisition*, Columbus, OH (1993).

[7] R. Basili, M.T. Pazienza and P. Velardi, Semi-automatic extraction of linguistic information for syntactic disambiguation, *Appl. Artif. Intell.* **4** (1993).

[8] R. Basili, M.T. Pazienza and P. Velardi, What can be learned from raw text? . . ., *Mach. Transl.* **8** (1993).

[9] R. Basili, M.T. Pazienza and P. Velardi, A (not-so) shallow parser for collocational analysis, in: *Proceedings COLING'94*, Kyoto (1994).

[10] R. Basili, M.T. Pazienza and P. Velardi, A context driven conceptual clustering method for verb classification, in: B. Boguraev and J. Pustejovsky, eds., *Corpus Processing for Lexical Acquisition* (MIT Press, Cambridge, MA, 1996).

[11] R. Basili, M.T. Pazienza and P. Velardi, Integration of probabilistic and symbolic methods for semantic categorization, in: *Proceedings AAAI Spring Symposium*, Stanford, CA (1995).

[12] R. Beckwith, C. Fellbaum, D. Gross and G. Miller, WordNet: a lexical database organized on psycholinguistic principles, in: U. Zernik, ed., *Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon* (Lawrence Erlbaum, London, 1991).

[13] B. Boguraev, Building a lexicon: the contribution of computers, IBM Report, T.J. Watson Research Center (1991).

[14] B. Boguraev and T. Briscoe, eds., *Computational Lexicography for Natural Language Processing* (Longman, New York, 1989).

[15] P. Brown, V.J. Della Pietra, J. Cocke, S.A. Della Pietra, F. Jelinek, R. Lafferty, R. Mercer and P. Roossin, A statistical approach to machine translation, *Comput. Linguist.* **16** (2) (1990).

[16] P. Brown, V.J. Della Pietra, P.V. deSouza, J.C. Lai and R. Mercer, Class-based n-gram models of natural language, *Comput. Linguist.* **18** (4) (1992).

[17] K. Church, W. Gale, P. Hanks and D. Hindle, Using statistics in: lexical analysis, in: U. Zernik, ed., *Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon* (Lawrence Erlbaum, London, 1991).

[18] K. Church and R.L. Mercer, eds., *Special Issue on Using Large Corpora*, *Comput. Linguist.* **19** (1) (1993).
[19] K.W. Church and P. Hanks, Word association norms, mutual information, and lexicography, *Comput. Linguist.* **16** (1) (1990).
[20] C. Cookson, Why computers need to learn English, *Financial Times* (September 20th, 1989).
[21] A. Copestake, The ACQUILEX LKB: representation issues in semi-automatic acquisition of large lexicons, in: *Proceedings 3rd ANLP* (1992).
[22] I. Dagan, F. Pereira and L. Lee, Similarity-based estimation of word co-occurrences probabilities, in: *Proceedings ACL Workshop*, Las Cruces, NM (1994).
[23] M. Evens, ed., *Relational Models of the Lexicon* (Cambridge University Press, Cambridge, 1988).
[24] D. Farewell, L. Guthrie and Y. Wilks, Automatically creating lexical entries for ULTRA, a multilingual MT system, *Mach. Transl.* **8** (1993) 127–145.
[25] R. Grishman, L. Hirschman and N.T. Nhan, Discovery procedures for sublanguage selectional patterns, *Comput. Linguist.* **12** (1986).
[26] R. Grishman, C. MacLeod and A. Meyers, Complex syntax: building a computational lexicon, in: *Proceedings COLING '94*, Kyoto (1994).
[27] R. Grishman and J. Sterling, Acquisition of selectional patterns, in: *Proceedings COLING '92*, Nantes (1992).
[28] R. Grishman and J. Sterling, Generalizing automatically generated selectional patterns, in: *Proceedings COLING '94*, Kyoto (1994).
[29] M. Hearst and H. Schuetze, Customizing a lexicon to better suite a computational task, in: *Proceedings ACL-SIGLEX Workshop on Lexical Acquisition from Text*, Columbus, OH (1993).
[30] D. Hindle, Noun classification from predicate argument structures, in *Proceedings ACL Workshop* (1990).
[31] D. Hindle and M. Rooths, Structural ambiguity and lexical relations, *Comput. Linguist.* **19** (1) (1993).
[32] L. Hirschman, R. Grishman and N. Sager, Gramatically-based automatic word class formation, *Inform. Process. Manage.* **11** (1975) 39–57.
[33] P. Jacobs, Text based intelligent systems, in: *Proceedings AAAI Spring Symposium*, Stanford, CA (1990).
[34] I. Koronenko and I. Bratko, Information-based evaluation criterion for classifier's performance, *Mach. Learn.* **6** (1991) 67–80.
[35] F.-M. Lang and L. Hirschman, Improved parsing through interactive acquisition of selectional patterns, in: *Proceedings Second Conference on Applied Computational Linguistics*, Austin, TX (1988).
[36] B. Merialdo, Tagging English text with a probabilistic model, *Comput. Linguist.* **20** (2) (1994).
[37] S. Montemagni and L. Vanderwende, Structural patterns vs. string patterns for extracting semantic information from dictionaries, in *Proceedings COLING '92*, Nantes (1992).
[38] F. Pereira, N. Tishby and L. Lee, Distributional clustering of English verbs, in: *Proceedings ACL Workshop*, Columbus, OH (1994).
[39] S. Pinker, *Learnability and Cognition – The Acquisition of Argument Structure* (MIT Press, Cambridge, MA, 1989).
[40] Background and experiments in machine learning of natural language, in: W. Daelemans and D. Powers, eds., *Proceedings First SHOE Workshop*, Tilburg (1992).
[41] A. Sanfilippo, Word knowledge acquisition, lexicon construction and dictionary compilation, in: *Proceedings COLING '94*, Kyoto (1994).
[42] A. Sanfilippo and V. Pozanski, The acquisition of lexical knowledge from combined machine-readable dictionary sources, in: *Proceedings Applied Natural Language Processing*, Povo Trento (1992).
[43] S. Sekine, J.J. Carrol, S. Anadianou and J. Tsujii, Automatic learning of semantic collocation, in: *Proceedings Third ANLP*, Trento (1992).
[44] J.F. Sowa, *Conceptual Structures in Mind and Machine* (Addison-Wesley, Reading, MA, 1984).

[45] T. Utsuro, Y. Matsumoto and M. Nagao, Verbal case frame a acquisition from bilingual corpora, in: *Proceedings IJCAI-93*, Chambery (1993).
[46] P. Velardi, Why human translators still sleep in peace? (Four engineering and linguistic gaps in NLP), in: *Proceeding COLING '90*, Helsinki (1990).
[47] D. Yarowsky, Word-sense disambiguation using statistical models of Roget's categories trained on large corpora, in: *Proceedings COLING '92*, Nantes (1992).
[48] D. Hindle and M. Rooth, Structural ambiguity and lexical relations, in: *Proceedings ACL Workshop*, Berkeley, CA (1991).