

Building an ontological support for multilingual legislative drafting

T. AGNOLONI, L. BACCI, E. FRANCESCONI, P. SPINOSA, D. TISCORNIA ^{a,1},
S. MONTEMAGNI, G. VENTURI ^b

^a *ITTIG-CNR, via de' Barucci 20, Florence, Italy*

^b *ILC-CNR, Via Moruzzi 1, Pisa, Italy*

Abstract. The quality of legislative drafting process at European and national levels is highly influenced by the legal drafters control over the multilingual complexity of European legislation and over the linguistic and conceptual issues involved in its transposition into national legislations. The DALOS project aims at ensuring coherence and alignment in the legislative language, providing law-makers with an ontological-linguistic resource and knowledge management tools to support the multilingual legislative drafting process. This paper outlines the current activity within DALOS, aiming at the construction of a two-level knowledge (ontological and linguistic) resource to be used as support for multilingual legislative drafting.

Keywords. Multilingual legal drafting, Legal ontologies, NLP techniques

1. Introduction

Quality in European and national legislative texts is one of the main purposes of the current initiatives of the European Commission. In particular the problem of harmonizing EU legal terminology is considered a precondition for improving the quality of legislative language and for facilitating access to legislation by legal experts and citizens. In a multilingual environment and, in particular, in EU regulations, only the awareness of the subtleties of legal lexicon, in the different languages, can enable drafters to maintain coherence among the different linguistic version of the same text. This is as much important for the EU Member State legal orders, strongly influenced by the obligation to implement EU directives.

To face this problem the DALOS² project has been recently launched within the “eParticipation” framework, the EU Commission initiative aimed at promoting the development and use of Information and Communication Technologies in the legislative decision-making processes. The aim of such initiative is to foster the quality of the legislative production, to enhance accessibility and alignment of legislation at European level, as well as to promote awareness and democratic participation of citizens to the legislative process.

¹Corresponding Author: E. Francesconi, via de' Barucci 20, Florence, Italy; E-mail: francesconi@ittig.cnr.it

²DrAfting Legislation with Ontology-based Support

In particular DALOS aims at ensuring that legal drafters and decision-makers have control over the legal language at national and European level, by providing law-makers with linguistic and knowledge management tools to be used in the legislative processes, in particular within the phase of legislative drafting.

To this aim DALOS use an ontological characterisation of legal language, giving conceptual meaning to the lexical units and providing connection with other terms, in order to provide law-makers with linguistic and knowledge management tools supporting legislative drafting in a multilingual environment.

In this paper the development of the DALOS ontological-linguistic resource is described. In particular in Section 2 and 3 the complexity of the multilingual legal scenario and a possible available resources to face it are addressed; in Section 4 the characteristics of the DALOS knowledge and the specification of its Knowledge Organization System (KOS) are presented; in Section 5 the phases to implement the ontological-linguistic resource are shown; in Section 6 the methodologies to implement the linguistic part of the resource are illustrated; finally in Section 7 some conclusions are reported.

2. The multilingual legal scenario

In legal language every term collection belonging to a language and law system is an autonomous vocabulary resource and should be mapped through relationships of equivalence with the others. The best approach to do it consists in developing parallel alignment with the same methodology and the same conceptual model.

Different methods may be applied, depending on the characteristic of the domain, the data structure and on the result to achieve. Among structured data different degrees of formalization can be distinguished: controlled vocabularies (such as thesauri, classification trees, directories, keywords lists), semantic lexicons as well as foundational, core, and domain ontologies.

The integration of lexical resources (heterogeneous because belonging to different law systems, or expressed in different languages, or pertaining to different domains) leads to different solutions depending on the desired results:

- generate a single resources covering both (merging);
- compare and define correspondences and differences (mapping);
- combining different levels of knowledge representation, basically interfacing lexical resources and ontologies.

The methodological approach chosen in the DALOS project is the third one: it requires the definition of mapping procedures between semantic lexicons, driven by the reference to an ontological level where the basic entities which populate the legal domain are described. Such an approach has been followed to obtain a correspondence between terms of different languages as well to align corresponding terms towards a common conceptualization at a higher knowledge level.

3. A legal semantic lexicon: the LOIS database

The DALOS resource is based on one of the wider lexical resource currently available in the legal field: the LOIS database³ composed by about 35000 concepts in five European languages (English, German, Portuguese, Czech, and Italian, linked by English).

In LOIS a concept is expressed by a synset. A synset is a set of one or more uninflected word forms (lemmas) with the same part-of-speech (noun, verb, adjective, and adverb) that can be interchanged in a certain context. For example *action, trial, proceedings, law suit* form a noun synset because they can be used to refer to the same concept. A synset is often further described by a gloss, explaining the meaning of the concept. English glosses drive cross-lingual linking.

In monolingual lexicons terms are linked by lexical relations: synonymy, near-synonym, antonym, derivation. Synsets are linked by semantic relations of which the most important are hypernymy/hyponymy (between specific and more general concepts), meronymy (between parts or wholes), thematic roles, instance-of.

Cross-lingual linking is based on equivalence relations (complete or near equivalence, as well as hyponym or hyperonym). The network of equivalence relations, the Inter-Lingual-Index (ILI), determines the interconnectivity of the indigenous wordnets. The LOIS approach is not completely language-independent, since the equivalence setting passes throughout the English wordnet and the English translation of glosses support the localization process.

The lesson learned from the LOIS experience is that a limited language independence could be enough for cross-lingual retrieval tasks, but it could be a weak point when considering re-using, extending, updating the semantic connections or when integrating external lexical resources (for instance multilingual thesauri) within the framework. What is needed is “the distinction between conceptual modeling at a language-independent level and a language and culture specific analysis and description of discourse-related units of understanding” [1].

These considerations led us to make clear distinction, when designing the overall model of DALOS and the system architecture, among types of knowledge, layers of knowledge representation as well as semantic relationships between knowledge elements.

4. DALOS knowledge organization and features

DALOS aims at providing a knowledge and linguistic resource for legislative drafting on the basis of the LOIS experience. The two projects however address two different scenarios: while the LOIS knowledge resource is addressed to multilingual legal information retrieval, the DALOS knowledge resource is expected to support legislative drafting.

This distinction of the addressed scenario is particularly important because it contributes to identify the type of knowledge to be described in DALOS, so to avoid the common attitude to indiscriminately mixing domain knowledge and knowledge on the process for which it is used (drafting, reasoning, searching, etc.). Such a mixing pre-

³created within the European project LOIS Legal Ontologies for Knowledge Sharing, EDC 22161, 2003-2006)

vents knowledge representations from being automatically reusable outside the specific context for which the knowledge representation was originally developed [2].

The DALOS case addresses the legislative drafting process, namely a process that creates norms on specific domains to be regulated. What is needed therefore is a knowledge and linguistic support giving a description of concepts, as well as their lexical manifestations in different languages, in specific domains. In particular, for the DALOS knowledge resource we want to avoid that the knowledge to be used as support for legislative drafting on a specific matter (*domain knowledge*) is mixed with the knowledge on the general process of drafting which, obviously, is matter independent (see also [3]). For the aim of developing a project pilot, the “consumer protection” domain has been chosen.

In the first phase of the project the most part of the activities have been addressed to provide the specification for the DALOS resource. After having chosen the domain of interest (“consumer protection”), currently the activities for domain knowledge specification are oriented to: 1) the standards to be used for knowledge representation, 2) the Knowledge Organization System (KOS). As regards standards, the RDF/OWL standard for WordNet representation as approved by the W3C standards has been used for the linguistic resource, thus guaranteeing interoperability as well as scalability of the solution. As regards KOS, on the basis of the arguments expressed in Section 3, the DALOS resource is expected to be organized in two layers of abstraction (see Fig. 1):

- the *ontological layer* containing the conceptual modeling at a language-independent level;
- the *lexical layer* containing lexical manifestations in different languages of the concepts at the ontological layer.

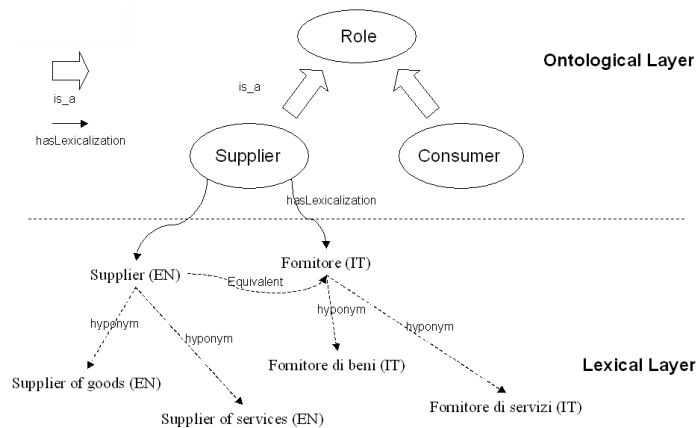


Figure 1. Knowledge Organization System (KOS) of the DALOS resource.

Basically the ontological layer acts as a knowledge layer where to align concepts at European level independently from the language and the legal order, where possible. Moreover the ontological layer allows to reduce the computational complexity of the problem of multilingual term mapping (N-to-N mapping). Concepts at the ontological layer act as a “pivot” meta-language in a N-language environment, allowing the reduction of the number of bilingual mapping relationships from a factor N^2 to a factor $2N$.

Concepts at the ontological layer are linked by taxonomical (*is_a*) as well as object property relationships.

On the contrary the lexical layer aims at describing language-dependent lexical manifestations of the concepts of the ontological layer. At this level terms are linked by linguistic relationships as those ones used for the LOIS database (*hyperonymy*, *hyponymy*, *meronymy*, etc.). In particular, to implement the lexical layer, the subset of the LOIS database pertaining to the “consumer protection” lexicon is used. Moreover this database is upgraded by using further texts where to extract pertaining terms from.

The connection between these two layers is aimed at representing the relationships between concepts and their lexical manifestations: a) within a single-language context (different lexical variations (lemmas) of the same meaning (concept)); b) in a cross-language context (multilingual variations of the same concept).

In the DALOS KOS such link is represented by the *hasLexicalization* (and its inverse *hasConceptualization*) relationship.

5. Implementation of the DALOS resource

The DALOS ontological-linguistic resource is implemented through three main activities:

1. Term extraction on the domain of “consumer protection” law from a set of selected texts by using NLP tools; this activity is aimed at upgrading the LOIS database (Lexical layer);
2. Construction of a Domain Ontology on the “consumer protection” domain (Ontological layer);
3. Semi-automatic connection between the Lexical layer (the LOIS database) and the Ontological layer by the *hasLexicalization* property implementation and its inverse *hasConceptualization* [Lexical layer ↔ Ontological layer]. This activity will be supported by semi-automatic tools and validated by humans.

The first activity (implementation of the Lexical layer) is being carried out using different NLP tools specifically addressed to process English and other EU language texts as GATE⁴ as well as Italian texts as T2K⁵. GATE is a tool to support advanced language analysis, data visualisation, and information sharing in many languages, owned/provided and maintained by the Department of Computer Science of the University of Sheffield. T2K is a terminology extractor and ontology learning tool jointly developed by CNR-ILC and University of Pisa.

The second activity (construction of a domain knowledge at the Ontological layer) is an intellectual one which aims at describing the scenario to be regulated. In this context the use of an ontology is of primary importance. Laws in fact usually contain provisions [4] which deal with entities (arguments) but they do not provide any general information on them: for example the Italian privacy law regulates the behaviour of the entity “Data controller” who is the owner of a set of personal data, but such law does not give any additional information on this role in the real domain-life. Therefore a formalized description in terms of an ontology of the domain to be regulated will allow to obtain such

⁴General Architecture for Text Engineering

⁵Text-to-Knowledge

additional general information on the entities a new act will deal with. Moreover, the use of an ontology, and particularly of the associated lexicon, allows to obtain a normalized form of the terms with which entities are expressed, enhancing quality and accessibility of legislative texts.

The third activity will deal with the connection between the two levels of abstraction (the *Ontological layer* and the *Lexical layer*). This activity is expected to be time consuming, since it will implement the legal concept alignment on the basis of their lexical manifestations in a multilingual environment. A tool to support such semi-automatic mapping is expected to be implemented within the project.

One of the current activities in DALOS is the first one, related to the bottom-up construction of the Lexical layer. Hereinafter such activity as regards the construction of the Italian part of the DALOS lexicon is described.

6. DALOS knowledge lexical layer implementation: the Italian case

As discussed in the previous sections, the Lexical layer of the DALOS knowledge is mainly based on the lexical database of the LOIS project. However, hand-crafted lexical and ontological resources need to be continuously extended and refined in order to incorporate up-to-date knowledge: “ontology-learning” [5] from texts can be of some help in this direction. To our knowledge, relatively few attempts have been made so far to automatically induce legal domain lexicons and ontologies from texts: this is the case, for instance, of [6], [7] and [8]. In the DALOS project, we decided to semi-automatically extend and customize pre-existing lexical and ontological resources on the basis of the terminological and ontological knowledge automatically acquired from texts belonging to the “consumer protection” domain with an ontology learning system. To this end, we used T2K (*Text-to-Knowledge*), a hybrid ontology learning system combining linguistic technologies and statistical techniques jointly developed by CNR-ILC and the Linguistics Department of the Pisa University [9].

At this stage, we evaluated how and to what extent the current version of the T2K system meets the project needs and what types of customization are needed for it to be used to extend and tune pre-existing resources. We started from the results of a case study carried out with T2K on a corpus of Italian legislative texts belonging to the environmental domain [10]. Currently, experiments are being carried out with the DALOS corpus of selected EU legislative texts on “consumer protection”.

6.1. T2K architecture

T2K is a hybrid ontology learning system combining linguistic technologies and statistical techniques. T2K does its job into two basic steps: 1) extraction of domain terminology, both single and multi-word terms, from a document base; 2) organization and structuring of the set of acquired terms into proto-conceptual structures, namely fragments of taxonomical chains and clusters of semantically related terms.

The approach to ontology learning adopted by T2K differentially exploits different levels of linguistic annotation of texts in an incremental fashion. Term extraction operates on texts annotated with basic syntactic structures: we use “chunking” technology to attain this level of basic syntactic structuring. NLP requirements become more demanding

when identified terms need be organised into conceptual structures. For this purpose syntactic information must include identification of dependencies among lexical heads (e.g. subject, object, modifier, etc.). The Italian parsing system underlying T2K is AnIta [11], a suite of linguistic tools in charge of the tokenisation of the input text, its morphological analysis (including lemmatisation), and syntactic parsing, which is in turn articulated in two different steps: “chunking”, carried out simultaneously with morpho-syntactic disambiguation, and dependency analysis.

In what follows we will illustrate the ontology learning process and on how its results could be exploited in the different activities foreseen for the implementation of the DALOS resource (see section 5).

6.2. Term extraction

Term extraction is the first and most-established step in ontology learning from texts. For our present purposes, a term can be a common noun as well as a complex nominal structure with modifiers (typically, adjectival and prepositional modifiers).

T2K looks for terms in shallow parsed texts, i.e. texts segmented into an unstructured (non-recursive) sequence of syntactically organized text units called “chunks” (e.g. nominal, verbal, prepositional chunks). Candidate terms may be one word terms (“single terms”) or multi-word terms (“complex terms”). The acquisition strategy differs in the two cases. Single terms are identified on the basis of frequency counts in the shallow parsed texts, after discounting stop-words. The acquisition of multi-word terms, on the other hand, follows a two-stage strategy. First, the chunked text is searched for on the basis of a set of chunk patterns. Chunk patterns encode syntactic templates of candidate complex terms: for instance, adjectival modification (e.g. *organizzazione internazionale* ‘international organisation’), prepositional modification (e.g. *commercializzazione di autovetture* ‘marketing of cars’), including more complex cases where different modification types are compounded (e.g. *commercio di prodotti fitosanitari* ‘trade of fitosanitary products’). Secondly, the list of acquired potential complex terms is ranked according to their log-likelihood ratio [12], an association measure that quantifies how likely the constituents of a complex term are to occur together in a corpus if they were (in)dependently distributed, where the (in)dependence hypothesis is estimated with the binomial distribution of their joint and disjoint frequencies. It should be noted that in T2K the log-likelihood ratio is applied in a somewhat atypical way: instead of measuring the association strength between adjacent words, T2K measures it between the lexico-semantic heads of adjacent chunks.

In T2K recognition of longer terms is carried out by iterating the extraction process on the results of the previous acquisition step. This means that acquired complex terms are projected back onto the original text and the acquisition procedure is iterated on the newly annotated text. The method proves helpful in reducing the number of false positives consisting of more than two chunks [13]. Interestingly, the chunk patterns used for recognition of multi-word terms need not necessarily be the same across different iteration stages. In fact, it is advisable to introduce potentially noisy patterns only at later stages. This is the case, for instance, of coordination patterns.

The iterative process of term acquisition yields a list of candidate single terms ranked by decreasing frequencies, and a list of candidate complex terms ranked by decreasing scores of association strength. The selection of a final set of terms to be included in the

ID	Term	Freq	Lemmatised headwords
1442	diritti d'autore	4	diritto autore
1600	diritti della difesa	4	diritto difesa
772	diritti di proprietà	10	diritto proprietà
1069	diritti fondamentali	6	diritto fondamentale
16	diritto	601	diritto
462	diritto d'uso	18	diritto uso
953	diritto di godimento	8	diritto godimento
120	diritto di recesso	148	diritto recesso
732	diritto di rescissione	11	diritto rescissione
674	diritto di revoca	12	diritto revoca
1211	diritto di utilizzazione	6	diritto utilizzazione

Table 1. An excerpt of the automatically acquired TermBank

TermBank requires some threshold tuning, depending on the size of the document collection and the typology and reliability of expected results. Thresholds define *a)* the minimum frequency for a candidate term to enter the lexicon, and *b)* the overall percentage of terms that are promoted from the ranked lists.

In what follows we illustrate what discussed so far with preliminary results obtained using the DALOS consumer law corpus, including Directives, Regulations and case law on protection of consumers' economic and legal interests, for a total of 284,795 word tokens. Different acquisition experiments have been carried out, by changing the minimum frequency threshold. Given the relatively restricted size of the acquisition corpus, better results were achieved setting the minimum frequency threshold to be equal to 3 for both single and multi-word terms. Concerning the percentage of terms to be selected from the candidate lists, we used standard thresholds: i.e. selected single terms are the topmost 10% in the ranked list, and selected multi-word terms are the topmost 70% in the ranked list of potential complex terms. We obtained a TermBank of 2.166 terms (both single and multi-word terms), which is currently being evaluated by domain experts.

Table 1 contains a fragment of the automatically acquired TermBank. For each selected term, the TermBank reports its prototypical form (in the column headed "Term"), its frequency of occurrence in the whole document collection, and the lemma of the lexical head(s) of the chunk(s) covering the term. It should be noted, however, that reported frequencies are not limited to the prototypical form, but refer to all occurrences of the abstract term.

The outcome of the term extraction step will be exploited to extend the lexical coverage of the LOIS database (Lexical layer).

6.3. Term organization and structuring

In the second extraction step, proto-conceptual structures involving acquired terms are identified. Two levels of conceptual organization are envisaged. Terms in the TermBank are first organized into fragments of head-sharing taxonomical chains, whereby *commercio dei medicinali* 'trade of medicines' and *commercio elettronico* 'electronic trade' are classified as co-hyponyms of the general single term *commercio* 'trade'. In this way, single and multi-word terms are structured in vertical relationships providing fragments of taxonomical chains such as the one reported below:

```

applicazione
  applicazione della legge
  applicazione delle disposizioni
  applicazione delle sanzioni
    applicazione delle sanzioni amministrative
    applicazione delle sanzioni previste
  ...

```

where the acquired direct and indirect hyponyms of the term *applicazione* ‘enforcement’ are reported. In this example, it can be noticed that terms sharing the head only are the direct hyponyms of the root term. Further hyponymy levels can be detected when two or more multi-word terms share not only the head but also modifiers. With minimum frequency threshold set to 3, the number of extracted hyponymic relations from the DALOS corpus is 911 referring to 172 hyperonym terms.

The second structuring step performed by T2K consists in the identification of clusters of semantically related terms which is carried out on the basis of distributionally-based similarity measures. This is done by using CLASS, a distributionally-based algorithm for building classes of semantically related terms [14]. According to CLASS, two terms are semantically related if they can be used interchangeably in a statistically significant number of syntactic contexts.

For all terms (both single and complex) in the TermBank, we extracted from the dependency-annotated text all relations involving these terms in the text. For each term, we identified (after discarding auxiliary and commonest verbs) the most meaningful (i.e. selective) verbs as resulting from the log-likelihood ratio association measure. The clusters of related terms were computed with respect to the most salient verbs associated with each target term; the terms similarity chains resulting from context-sensitive similarity measures were then merged and ranked according to decreasing similarity weights.

In what follows, clusters of semantically related terms are exemplified:

```

disposizioni ‘provision’
  norme, disposizioni relative, decisione, atto, prescrizioni
legge ‘law’
  regolamento, protocollo, accordo, statuto, amministrazioni comunali
cmv (acronym for the comitee for veterinary medicines)
  comitato, cpmp (Comitee for Proprietary Medicinal Products),
  commissione, membri, consiglio

```

For each target term, the set of the first 5 most similar terms is returned, ranked for decreasing values of semantic similarity. With the minimum frequency threshold set to 3, the number of identified related terms is 1,071 referring to 238 terminological headwords. It should be appreciated that in these clusters of semantically related words different classificatory dimensions are inevitably collapsed; they include not only quasi-synonyms (as in the case of *disposizioni* ‘provision’ and *norme* ‘regulations’), hyperonyms and hyponyms (e.g. *comitato* ‘comitee’ and *cmv* (*comitato per i medicinali veterinari*) ‘comitee for veterinary medicines’), but also looser word associations. As an example of the latter we mention the relation holding between *legge* ‘law’ and *amministrazione comunale* ‘municipal administration’.

The proto-conceptual structures, i.e. the fragments of taxonomical chains of terms together with the clusters of semantically related terms, acquired during the term structuring step will provide useful input for both the construction of the DALOS domain ontology and the definition of the mapping between the lexical and the ontological layers.

7. Conclusions

The main purpose of the DALOS project is to provide law-makers with linguistic and knowledge management tools to be used in the legislative processes, in particular within the phase of legislative drafting. The aim is to keep control over the legal language, especially in a multilingual environment, as the EU legislation one, enhancing the quality of the legislative production, the accessibility and alignment of legislation at European level, as well as to promote awareness and democratic participation of citizens. In this paper the characteristics of the ontological-linguistic resource as well as the methodologies for its construction have been presented.

References

- [1] K. Kerremans and R. Temmerman, "Towards multilingual, termontological support in ontology engineering," in *Proceeding of Termino 2004, Workshop on Terminology*, 2004.
- [2] J. Breuker and R. Hoekstra, "Epistemology and ontology in core ontologies: Folaw and Iricore, two core ontologies for law.," in *In Proceedings of EKAW Workshop on Core ontologies. CEUR 2004.*, 2004.
- [3] C. Biagioli and E. Francesconi, "A visual framework for planning a new bill," *Quaderni CNIPA (Proceedings of the 3rd Workshop on Legislative XML)*, no. 18, pp. 83–95, 2005.
- [4] C. Biagioli, "Towards a legal rules functional micro-ontology," in *Proceedings of workshop LEGONT '97*, 1997.
- [5] P. Buitelaar, P. Cimiano, and B. Magnini, "Ontology learning from text: an overview," in *Ontology Learning from Text: Methods, Evaluation and Applications (Volume 123 Frontiers in Artificial Intelligence and Applications)* (B. et al., ed.), pp. 3–12, 2005.
- [6] G. Lame, "Using nlp techniques to identify legal ontology components: concepts and relations," *Lecture Notes in Computer Science*, vol. 3369, pp. 169–184, 2005.
- [7] J. Sais and P. Quaresma, "A methodology to create legal ontologies in a logic programming based web information retrieval system," *Lecture Notes in Computer Science*, vol. 3369, pp. 185–200, 2005.
- [8] S. Walter and M. Pinkal, "Automatic extraction of definitions from german court decisions," in *Proceedings of the COLING-2006 Workshop on Information Extraction Beyond The Document, Sidney*, pp. 20–28, 2006.
- [9] D. F., A. Lenci, S. Marchi, S. Montemagni, and V. Pirrelli, "Text-2-knowledge: una piattaforma linguistico-computazionale per l'estrazione di conoscenza da testi," in *Proceedings of the SLI-2006 Conference, Vercelli*, pp. 20–28, 2006.
- [10] A. Lenci, S. Montemagni, V. Pirrelli, and G. Venturi, "Nlp-based ontology learning from legal texts. a case study," in *P. Casanovas, M.A. Biasiotti, E. Francesconi and M.T. Sagri (Eds.), Proceedings of LOAIT 07 - II Workshop on Legal Ontologies and Artificial Intelligence Techniques*, pp. 113–129, 2007.
- [11] R. Bartolini, A. Lenci, S. Montemagni, and V. Pirrelli, "Hybrid constrains for robust parsing: First experiments and evaluation," in *Proceedings of LREC 2004, Lisbon*, 2004.
- [12] T. Dunning, "Accurate methods for the statistics of surprise and coincidence," *Computational Linguistics*, vol. 19, no. 1, 1993.
- [13] R. Bartolini, D. Giorgetti, A. Lenci, S. Montemagni, and V. Pirrelli, "Automatic incremental term acquisition from domain corpora," in *Proceedings of the 7th International conference on "Terminology and Knowledge Engineering" (TKE2005), Copenhagen, Denmark*, 2005.
- [14] M. S. Allegrini, P. and V. Pirrelli, "Example-based automatic induction of semantic classes through entropic scores," *Linguistica Computazionale*, 2003.