Creation of a bottom-up corpus-based ontology for Italian Linguistics[1]

Elisa Bianchi*, Mirko Tavosanis*, Emiliano Giovannetti°
*Dipartimento di Studi Italianistici - University of Pisa, Italy
° Istituto di Linguistica Computazionale "Antonio Zampolli" (ILC-CNR), Pisa, Italy
e.bianchi@italicon.it, tavosanis@ital.unipi.it, giovannetti@ilc.cnr.it

Abstract

This paper describes the steps of construction of a shallow lexical ontology of Italian Linguistics in Italian, set to be used by a meta-search engine for query refinement. The ontology was constructed with the software Protégé 4.0.2 and encoded in OWL format; its construction has been carried out following the steps described in the well-known Ontology Learning From Text (OLFT) layer cake. The starting point was the automatic term extraction from a corpus of web documents concerning the domain of interest (304,000 words); as regards corpus construction, we describe the main criteria of the web documents selection and its critical points, concerning the definition of user profile and of degrees of specialisation. We then describe the process of term validation and construction of a glossary of terms of Italian Linguistics; afterwards, we outline the identification of synonymic chains and the main criteria of ontology design: top classes of ontology are Concept (containing taxonomy of concepts) and Term (containing terms of the glossary as instances), while concepts are linked through part-whole and involved-role relation, both borrowed from Wordnet. Finally, we show some examples of the application of the ontology for query refinement.

Keywords: Ontologies, Italian Linguistics, Query refinement

1. Introduction

This paper aims at outlining the stages of the design of a domain-specific ontology, i.e. an ontology of Italian Linguistics in Italian; it is a shallow lexical ontology, set to be used by a meta-search engine for query refinement related to queries about Italian Language and Linguistics[2].
We will focus mainly on the special features characterizing ontologies in humanities: we will highlight the relevance of the point of view and the human interpretation in selecting relevant concepts, organizing them in a taxonomy and linking them through relations (Aussenac-Gilles & Soergel, 2005: 38). Furthermore, we will address the issue of the adequacy of existing tools (e.g. Protégé) to manage an ontology of language and linguistics, showing that the conceptual model of the software could be modified according to the specificity of represented concepts.

We will show the steps that led to the construction of the corpus, the ontology design and its implementation through concepts and terms. We will conclude showing an application of the ontology for query refinement in a meta-search engine.

2. User profile and web documents selection criteria

The ontology described here was envisioned as a step in the creation of an integrated services platform for semantic and multilingual access to Italian cultural contents in the web. This platform is the main deliverable of the FIRB research project "Panorama" funded by the Italian Ministry of Education, University and Research (RBNE07C4R9 - Decree 190/Ric., 12 march 2009; web site http://www.panoramafirb.it/), coordinated by Marco Santagata and involving many leading institutions in the field of Italian cultural contents. In the course of the project, research units of the universities of Pisa and Roma "La Sapienza" and of the ICCU (Istituto Centrale per il Catalogo Unico) planned and created ontologies for three humanities subfields (Italian Linguistics, Italian literature, Art in Italy). This paper will describe only the Italian Linguistics ontology.

Ontologies have a strict relationship with the profiles of the users whose needs they address (Staab & Studer, 2003). User profiles in this context can be defined by two parameters: first of all, the level of specialisation of domain-knowledge, i.e. Italian Linguistics; secondly, the reason why the user makes use of a meta-search engine. Since the meta-search engine is aimed at a wide range of users, from schoolchildren to professional research, and since it is meant to facilitate a wide range of searches, it was felt that the ontology should include the whole range of linguistic concepts covered by web texts. The desirability of a bottom-up ontology was therefore clear.

Web documents concerning Italian Linguistics have a wide range of variation, according to web genre, content validity and completeness; in ontology design, therefore, it is essential to take account of different points of view and degrees of specialisation. For example, regarding genres, Italian Linguistics texts are included in very different types of web sources: institutional sites, research sites, blogs, forums, non specialised sites etc. However, no established guidelines for web genre classification

[1] Elisa Bianchi wrote §§ 3.3, 3.4, 3.5, 4, 6; Mirko Tavosanis wrote §§ 1, 2, 5; Emiliano Giovannetti wrote §§ 3.1, 3.2.

[2] The used meta-search engine is TagMySearch, developed by the research unit of the Computer Science Department of the University of Pisa, coordinated by prof. Paolo Ferragina. Details about the meta-search engine can be found at http://acube.di.unipi.it/tagme/. The integration of the OWL file to the meta-search engine was carried out by Ugo Scaiella and Daniele Vitale. The Panorama meta-search engine is available at the URL http://toma.di.unipi.it/panoramafirb/search.jsp.

exist at the moment (Rehm et al., 2008; Mehler et al., 2010; Tavosanis, 2011). It is difficult, then, to define selection criteria for the corpus. The selection for this work was therefore based exclusively on the competence of the authors of the texts, taking into account any kind of web document showing at least some degree of confidence with Linguistics as a discipline. It is worth noting that the task was made comparatively hard by a dearth of reliable web texts in this field; at the date of completion of this paper the "Panorama" research project could list in its online archive only 337 sites, institutes or collections providing reliable texts in this field.

Aiming to serve the needs of the highest possible number of user types, we chose then web pages from 53 websites listed by the project. The pages covered a wide range of variation according both to web genres and degree of ation[3].

The cleaning (partly automated, partly manual) of the pages allowed to create a corpus of about 304.000 words. As the first step in the construction of the ontology the corpus was then processed with a knowledge extraction tool (T2K) to extract a list of domain-relevant terms (Bonin et al., 2010).

3. Ontology construction methodology

3.1 Steps of ontology construction

The construction of the ontology has been carried out following the steps described in the well-known Ontology Learning From Text (OLFT) layer cake (Buitelaar et al., 2005). According to this methodological approach, an ontology is built "bottom-up" starting from the very words composing a text. First, domain-relevant terms (single and multi-word) are detected, representing the "domain terminology" (sometimes called "glossary"). Terms are subsequently aggregated into classes of synonyms and then into concepts. The latter are then organised into a hierarchy (or taxonomy) through the relations of hyponymy intercurrent between the terms denoting them, and thereafter placed in relation with each other by means of non-taxonomic semantic relations. The last stage of the process of OLFT can include the definition of a set of rules, by means of which it is possible to establish logical inferences in the form of "if-then" expressions.

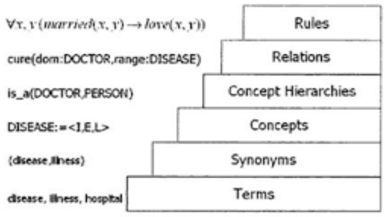


Figure 1: Ontology Learning Layer Cake (Buitelaar et. al., 2005)

To date, we have carried out the first 5 steps of ontology learning, although the results, and the ontology itself, are still partial and need further revision and development.

3.2 Automatic Terminology Extraction

The automatic term extraction process used for the construction consisted of two fundamental steps: 1) identifying term candidates (either single or multi–word terms) from text, and 2) filtering through the candidates to separate terms from non-terms. Concerning the detection of multi-word terms, usually constituting the 85% of the total of domain terms (Nakagawa & Mori, 2003), we used a combination of "termhood" measures, assessing the likelihood that the term was a valid technical term, and contrastive methods. In particular, multi–word term extraction has been carried out by identifying candidate multi–word terms in the corpus, previously automatically POS-tagged and lemmatized. The selected terms have been weighted using the C-NC value (Frantzi et al., 1999), currently considered as the state–of–the–art method for terminology extraction. The ranking of identified multi–word terms has been then revised on the basis of a contrastive score calculated for the same terms with respect to corpora testifying general language usage.

3.3 Term validation

By the word "validation" we mean the process of selection and evaluation of lexical units automatically extracted from the collection of web documents, according to the described approach.

The intended goal was to set up a glossary of Italian Linguistics, to serve as the basis for identifying classes of synonyms and concepts of the ontology. Thus, we validated as terms specific lexical units corresponding to relevant concepts of the domain. Terms can be thus seen as the nodes of a conceptual map of the domain, and at the same time as lexical entries of an encyclopaedia of linguistics. We chose to define terms as "the words that are assigned to concepts used in special languages that occur in subject-field or domain-related texts" (Wright & Budin, 1997).

This terminological approach has some issues: between terms (T) and "non-terms" (NonT, i.e. lexical units not belonging to the domain), we identified lexical units not corresponding to concepts, but nevertheless useful to identify distinctive features of texts about Italian language and linguistics. To represent this intermediate collocation between Terms and Non-Terms, we called this group of lexical units "Near-Terms" (NT).

All terms included in the glossary are nouns in singular form. They are both single-word units (*Parola, Word*) and multiwords (*Prova di valutazione, Evaluation test*). These are the final data about term validation:

[3] Some significant sites texts of the corpus were taken from are: the website of the Accademia della Crusca (www.accademiadellacrusca.it), the website of an Italian academic research project about spoken Italian (www.parlaritaliano.it), the high-quality informational site of the main Italian encyclopedia (www.treccani.it), an institutional site (www.ladante.it).

Terms	1372	40%
Near Terms	554	16%
Non Term	1501	44%
Total	3627	100%

Table 1: Term validation

Here are some examples of terms representing concepts:

Lingua italiana (*Italian language*), Parola (*Word*), Dialetto (*Dialect*), Nome (*Noun*), Lessico (*Lexicon*), Verbo (*Verb*), Grammatica (*Grammatics*), Accento (*Accent*), Vocale (*Vowel*).

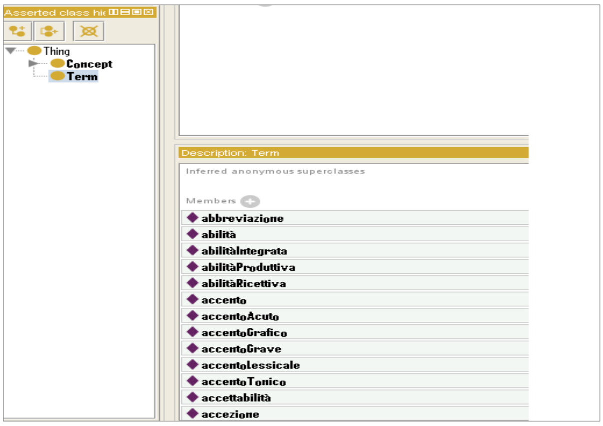
Examples of Near-Terms: Numero di parole (*Word number*), Modello di repertorio (*Repertory model*), Materiale lessicale (*Lexical material*), Insieme di parole (*Set of words*) etc.

Examples of NonTerms: Introduzione (*Introduction*), Ipotesi di partenza (*Starting hypothesis*), Immissione di dati (*Data entry*), Gestione di classe (*Classroom management*) etc.

We then obtained a glossary of terms of Italian language and linguistics extracted from web documents. This allows us to keep a close connection with the scope of application of the ontology: since terms have been extracted from web documents (the Web used as the "source"), the meta-search engine will be able to exploit the obtained glossary for query refinement (the Web used as the "target") in a more efficient way (see § 5).

3.4 The ontology architecture and the construction of taxonomy

Following the term validation and creation of the glossary, we proceeded to set up the ontology, by using Protégé 4.0.2. The adopted format for exporting the ontology was OWL, a knowledge representation language built upon RDF and RDFS, and endorsed by the from World Wide Web Consortium (W3C) as the *de facto* standard for representing ontologies in the Semantic Web.
The architecture of the ontology responds to its application purpose, i.e. query refinement related to queries about Italian language and linguistics: it shall be foremost a tool for matching user queries with concepts of the domain, by suggesting new queries including both possible synonyms of the original query and related concepts.
Top concepts (referred to "Classes" in Protégé) are Term and Concept. The choice to base the ontology on these two top concepts is motivated by the need to represent, in the ontology, the conceptual structure of our domain and to link concepts to the terms by which they are referred to in web documents. The class Term should therefore include linguistic elements of possible queries about Italian Language and Linguistic by a user.
All terms of the glossary are therefore instances (or individuals) of the class Term. All terms are singular nouns, and their length varies from 1 (*Parola*) to 4 (*ProvaStrutturataDiValutazione, Structured evaluation test*).



abbreviazione
abilità
abilitàIntegrata
abilitàProduttiva
abilitàRicettiva
accento
accentoAcuto
accentoGrafico
accentoGrave
accentoLessicale
accentoTonico
accettabilità
accezione

Figure 2: Instances (or individuals) of class Term

The class Concept includes relevant concepts of Italian language and Linguistics, hierarchically ordered in a taxonomy. Our aim was to include, as far as possible, the most important concepts of various branches of Italian Linguistics and different degrees of specialisation of Italian language and linguistics.
Since defining concepts is one of the first validity requirements of a domain ontology, most of the concepts (about 50%) are defined (in "is defined by" of "Annotations" field) by the corresponding URL in Italian version of Wikipedia (http://it.wikipedia.org): we chose to define them in this way because URLs of Wikipedia are used by the meta-search engine to process thematic clustering, which could be enhanced by OWL in future developments of research.

Properties stated about Classes in our ontology are (with corresponding inverse relations):

denotes/is_denoted by
has_holonym/has_meronym (part-whole)
involved/role
refers_to/is_referred_by

We outline now the features of the first three properties (the fourth one was used only in a small number of cases): denotes/is_denoted by, has_meronym/has_holonym and involved/role.

3.4.1. Denotes/is_denoted_by

Through the use of the property "denotes" and its inverse property "is_denoted_by" concepts are related to instances of the class Term: each term must denote at least one concept, and each concept must be denoted at least by one term.
So, for the concept Dizionario (*Dictionary*), the following

information is visualised:

Dizionario (*Dictionary*) is denoted by value dizionario (*Dictionary*)
Dizionario (*Dictionary*) is denoted by value vocabolario (*Vocabulary*)

In this structure, different terms linked to the same concept by the relation "denotes" are considered as synonyms: for example, dizionario is synonym of vocabolario since both denote the same concept. In our structure, a term is synonym of another term if it can "be replaced by another in a specific context" (Alonge et al., 1998: 93). In some cases, to establish synonymic equivalence between two lexical units is not straightforward, since it is essential to take as a point of reference a certain level of competence of the user: for example, in non specialised texts, L2 (*Second Language*) is used as a general term, while in academic and specialised communication there is a clear distinction between L2 (*Second Language*) and LS (*Foreign Language*).
In the taxonomy, as expected, hyponyms are linked to hypernyms by the relation ISA, and taxonomic grouping follows criteria of similarities/differences of semantic features:

Testo descrittivo (*Descriptive Text*) is-a Testo (*Text*)
Lingua Italiana Commerciale (*Commercial Italian Language*) is-a Varietà Della Lingua (*Variety of language*)

Properties stated about the hypernym are inherited by hyponyms, so that the reasoner can infer properties not explicitly stated about each class. For example, the properties stated about class Frase (*Sentence*) are:

has_holonym some Enunciato (*Utterance*)
has_meronym some Complemento (*Complement*)
has_meronym some Parola (*Word*)
role some Accento (*Accent*)
role some Ordine Marcato (*Marked order*)
role some Ordine Naturale (*Natural order*)
role some Struttura (*Structure*)
is_denoted_by value frase (*Sentence*)
is_denoted_by value proposizione (*Clause*)

All these properties are inherited from hyponyms of Frase, for example Frase Affermativa (*Affirmative Sentence*), Frase Esclamativa (*Exclamatory Sentence*) and so on.

3.4.2. Part-whole relation

The inverse relation "has_meronym/has_holonym" is used for part-whole relation, e.g.:

Sillaba (*Syllable*) has_holonym some Parola (*Word*)
Parola (*Word*) has_meronym some Sillaba (*Syllable*)

Unlike other lexical ontologies and semantic nets, such as Wordnet, our ontology encodes only one type of part-whole relation. Different types of meronymy (Winston et al., 1987), indeed, were not considered as relevant for the application of ontology, i.e. for query refinement.
An interesting facet of part-whole relations in our ontology are the so called "multiple relations":

Frase (*Sentence*) has_meronym some Complemento (*Complement*)
Frase (*Sentence*) has_meronym some Parola (*Word*)

As seen in the example, part-whole relations in linguistics vary according to the level of analysis and to the discipline (for example, morphology, syntax and so on).
The different way of analyzing the structure of language and linguistic phenomena in general also leads to a different segmentation of the linguistic units, making it essentially impossible to identify unique part-whole relations, and Protégé is not suitable to represent the variability of analysis levels.
The structure of Protégé, indeed, requires a high degree of formalization of concepts and their relations, starting from a homogeneous view of the domain of interest: it is not possible yet to make Protégé's structure fit to the perspectives of analysis' variety of the same concept (such as Word, Sentence and so on), and the consequent multiplicity of part-whole relations, changing with the perspective of analysis itself. This variety of levels of analysis and of segmentation of linguistic units adds complexity to the issue of "consensual conceptualisation" (Maroto & Alcina, 2005: 232), necessary for the construction of a domain ontology: even if there was full agreement in conceptual modeling of the domain of Italian linguistics, the multiplicity of perspectives would in any case remain, as inherent to the analysis of linguistic phenomena itself.

3.4.3. Involved/Role

This inverse semantic relation is borrowed from Wordnet and is used "for encoding data which better characterize a word meaning" (Alonge et al. 1998: 101).
Involved/Role links two concepts one of which involves the other.

Some examples are:

Aspetto (*Aspect*) involved some Verbo (*Verb*)

Competenza Metalinguistica (*Metalinguistic competence*) involved some Metalinguaggio (*Metalanguage*)

Accento (*Accent*) involved some Sillaba (*Syllable*)

This relation adds relevant semantic information to concepts, and is indeed essential for word disambiguation in query refinement. Combination, in query refinement, of two concepts linked by involved/role relation produces indeed word sequences matching with complex phrases, characterizing specialised texts:

accento - parola > accento di parola (*Word accent*)
aspetto - verbo > aspetto del verbo (*Verbal aspect*)
acquisizione - lingua > acquisizione della lingua (*Language acquisition*)
leggibilità - testo > leggibilità del testo (*Text readability*)

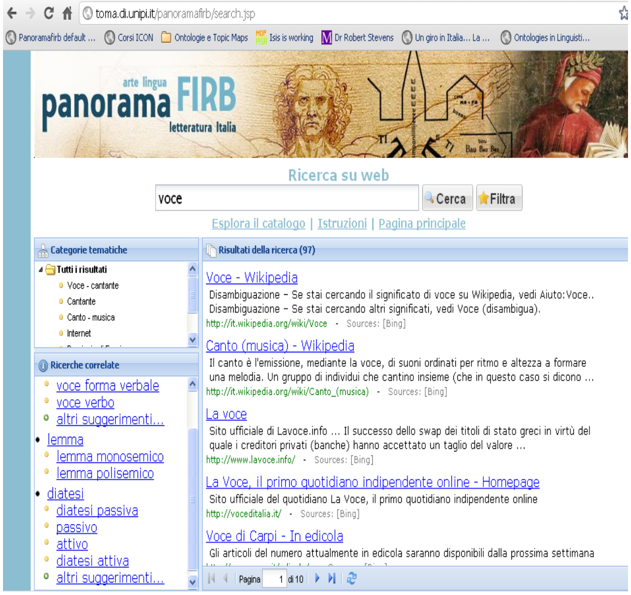
Many of these complex concepts match with multiword units of length 3 in the list of automatically extracted term, with structure N + Prep + N.

3.5 Conceptualisation processes

The process of creating classes (i.e. concepts) from the glossary was not trivial: first of all, we used 516 terms from the list of automatically extracted lexical units, about 15% of the total. We had primarily given priority to the development of taxonomies, i.e. hypernyms/hyponyms relations, rather than the inclusion of unrelated concepts, belonging to different disciplines and taxonomies.
Many automatically extracted terms were part of taxonomies, which we managed to complete as far as possible in our ontology, including hypernyms and hyponyms.
So, we found in the list of automatically extracted terms only partial taxonomies, which we completed by adding new concepts. For example, among automatically extracted terms we did find some terms relating to parts of speech (noun, verb, etc.): in implementing the part-of-speech taxonomy, we completed it by adding missing concepts and stating associative relations between them. Again, in automatically extracted term list there's the term pronome atono (*Unstressed pronoun*), which we included in our ontology with the concept Pronome Atono, hyponym of concept Pronome (*Pronoun*); we decided then to add a term (and corresponding concept) not present in word list, i.e. Pronome Tonico (*Stressed pronoun*), to complete the taxonomy with the antonym of Pronome atono (*Unstressed pronoun*).
Overall, we added 329 terms (38,9%) to the 516 units (61,1%) taken from the list of automatically extracted terms.

4. Application and evaluation

The scope of application of ontology is query refinement: the goal is to enhance precision and recall of retrieved documents by suggesting new queries to the user, starting from lexical relationships of the ontology.
As Bhogal et al. (2007: 875) point out, "ontologies improve the accuracy in fuzzy information search and facilitate mono- and multi-lingual human-computer dialogues by paraphrasing the query of the user through context identification and disambiguation."
If a given query matches with one term of the ontology, the search engine suggests a number of queries related the term; by clicking on one of them, the original query is replaced by the new query.
In figure 3 we show the suggested queries retrieved by the initial query "voce", term which in linguistics is polysemous:



Voce - Wikipedia
Disambiguazione – Se stai cercando il significato di voce su Wikipedia, vedi Aiuto:Voce.. Disambiguazione – Se stai cercando altri significati, vedi Voce (disambigua).
http://it.wikipedia.org/wiki/Voce - Sources: [Bing]

Canto (musica) - Wikipedia
Il canto è l'emissione, mediante la voce, di suoni ordinati per ritmo e altezza a formare una melodia. Un gruppo di individui che cantino insieme (che in questo caso si dicono ...
http://it.wikipedia.org/wiki/Canto_(musica) - Sources: [Bing]

La voce
Sito ufficiale di Lavoce.info ... Il successo dello swap dei titoli di stato greci in virtù del quale i creditori privati (banche) hanno accettato un taglio del valore ...
http://www.lavoce.info/ - Sources: [Bing]

La Voce, il primo quotidiano indipendente online - Homepage
Sito ufficiale del quotidiano La Voce, il primo quotidiano indipendente online
http://voceditalia.it/ - Sources: [Bing]

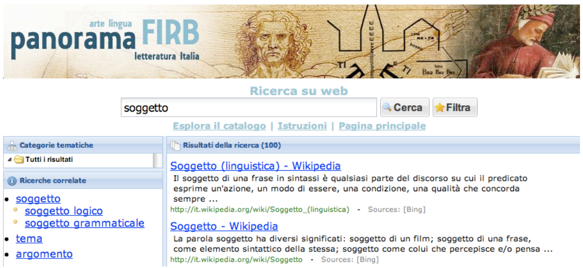
Voce di Carpi - In edicola
Gli articoli del numero attualmente in edicola saranno disponibili dalla prossima settimana

Figure 3: Example of query refinement from query "voce"

As can be seen, the suggested queries guide the user to disambiguate the meaning of "voce" and to refine the query with new words.
Suggested queries are drawn from ontology, according to the following rules:

- there's a first match between the query and concepts of the ontology;
- for each retrieved concept, hyponyms (e.g. accento acuto and accento grave for accento grafico) and involved/role relations (e.g. accento parola and accento proposizione for accento), are displayed as additional terms of the original query.

By selecting one of the suggested queries, a new query replaces the initial query, and is processed for new suggested queries derived from the OWL file.
Initial queries are processed with stemming, to find matches with morphological variations (e.g. singular vs. plural and so on).
Suggested queries add elements of knowledge to the initial query and guide the user to refine the query, first of all by disambiguating polysemous terms and distinguishing different meanings. For example, if the query is soggetto, a polysemous term designating in linguistics both "subject" and "topic", suggested queries are shown in figure 4.
For Soggetto (*Subject*), two relevant hyponyms, i.e. soggetto grammaticale and soggetto logico are displayed, whereas for Soggetto (*Topic*) synonymic terms Tema and Argomento are suggested.



soggetto

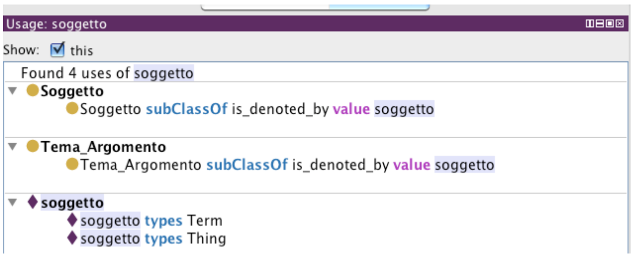
- soggetto
  - soggetto logico
  - soggetto grammaticale
- tema
- argomento

Soggetto (linguistica) - Wikipedia
Il soggetto di una frase in sintassi è qualsiasi parte del discorso su cui il predicato esprime un'azione, un modo di essere, una condizione, una qualità che concorda sempre ...
http://it.wikipedia.org/wiki/Soggetto_(linguistica) - Sources: [Bing]

Soggetto - Wikipedia
La parola soggetto ha diversi significati: soggetto di un film; soggetto di una frase, come elemento sintattico della stessa; soggetto come colui che percepisce e/o pensa ...
http://it.wikipedia.org/wiki/Soggetto - Sources: [Bing]

Figure 4: Suggested queries for query “soggetto”

Underlying data of this query refinement concerning “soggetto” are the properties of the term “soggetto” (and its link with different concepts) stated in the OWL file, and summarised in “Individual usage” window:



Found 4 uses of soggetto

- Soggetto
  - Soggetto subClassOf is_denoted_by value soggetto
- Tema_Argomento
  - Tema_Argomento subClassOf is_denoted_by value soggetto
- soggetto
  - soggetto types Term
  - soggetto types Thing

Figure 5: Individual usage of “soggetto” in the OWL file

This query expansion, as it has been set up, performs two basic functions:

- it improves the user’s knowledge, suggesting relevant terms of Italian Linguistics related to the original query; documents retrieved using suggested queries are more relevant and limited to selected topics.

- it provides meaning disambiguation of polysemous terms, such as soggetto, grammatica and so on.

Actually, this mechanism improves relevance of retrieved documents, although it is very difficult to assess objectively the degree of improvement according to standard criteria of precision and recall (Barathi, 2010): it is indeed impossible to calculate both the total number of documents concerning Italian Linguistics and to define univocally the degree of “relevance”.

According to the analysis of a series of meaningful queries of the domain, we can state that suggested queries allow to narrow the topic and therefore to retrieve more relevant documents, excluding at the same time non-relevant documents.

If the user searches for grammatica italiana (*Italian grammar*), retrieved documents refer to different meanings of grammatica: grammar as a set of rules, grammar as a text book, grammar as a branch of linguistics.

In this case, although all retrieved documents are relevant for the query grammatica italiana, it would be useful to distinguish among different meanings of the term, and to refine the query with more specific terms, such as grammatica prescrittiva, grammatica normativa and testo di grammatica. This allows to narrow and refine the query, and thus to exclude not relevant documents.

In some cases, there may be a mismatch between the concept and the lexical occurrences in web documents: for example, grammatica prescrittiva (*grammar rules*) is a relevant concept of Italian Linguistics but, from the point of view of lexical occurrences, it would be more useful to retrieve the involved terms, i.e. regola or norma (grammaticale), which are much more frequent in web documents. This could be a critical element and in future research it needs to be revised, particularly as regards retrieval of Concepts vs. Terms.

Furthermore, the rules of combination of retrieved concepts should be revised: currently, if the user searches for “grammatica italiana”, “testo di grammatica” is displayed as suggested query; but if this last query is clicked (i.e. “testo di grammatica”), a crucial information, “italiana”, is lost.

5. Conclusions

The match between the ontology and the meta-search engine seems able to generate effective improvements in the search both for experts and non-experts in linguistics. It is worth noting that the final outcome of the queries seems strictly determined by the conceptualisation process described in 4.4. This process supplements from other sources the data taken from the corpus; we feel, however, that this does not change the fundamentally “bottom-up” nature of the ontology. The supplemented terms could instead play a role similar to that of the “highly available” terms in dictionary creation, where lists high-frequency words taken from corpora are merged with short lists of low-frequency words found by other sources and well known to the speakers of a language, even if seldom used.

A further step should bring us to evaluate in a systematic way, with real users and for different kind of queries, the results given by the system. This kind of trial could allow us to understand if this kind of ontology really outperforms more standard (and “abstract”) constructs, and, eventually, by what margin and for which kind of uses.

6. Acknowledgments

The ontology described in this paper is part of the Italian FIRB project “Web service and search engine to support a semantic and plurilingual access to Italian Culture on the Web” (RBNE07C4R9).

Automated terminology extraction was carried out by Francesca Bonin, formerly associate with the ILC.

We thank Amedeo Cappelli (ISTI – CNR/Pisa) for his help in setting out the ontology in its initial stages: his contribution has been essential to deal properly with issues raised by the management of terminology and conceptualisation.

7. References

Alonge A. et al. (1998). The Linguistic Design of the EuroWordnet Database. In *Computers and the Humanities*, 32, pp. 91--115.

Aussenac-Gilles, N., Soergel, D. (2005). Text Analysis for ontology and terminology engeneering. In *Applied Ontology* 1(1), pp. 35--46.

Barathi, M., Valli, S. (2010). Ontology Based Query Expansion Using Word Sense Disambiguation, In *IJCSIS, International Journal of Computer Science and Information Security*, 7(2), pp. 22--27.

Bhogal, J., Macfarlane, A., Smith, P. (2007). A review of ontology based query expansion. In *Information Processing and Management* 43, pp. 866--886.

Bonin, F., Dell'Orletta, F., Venturi, G., Montemagni, S. (2010). A Contrastive Approach to Multi-word Term Extraction from Domain Corpora. In *LREC'10 - Seventh International Conference on Language Resources and Evaluation (Valletta, Malta, 17-23 May 2010). Proceedings*, pp. 3222--3229.

Buitelaar, P.; Cimiano, P. and Magnini, B. (2005). Ontology learning from text: methods, evaluation and applications. Amsterdam: IOS Press.

Cappelli, A. (1991). Intensional Semantics and Relationships between Epistemology and Ontology. In *Description Logics*, pp. 33--38.

Frantzi, K., Ananiadou, S. (1999). The C–value / NC Value domain independent method for multi–word term extraction. In *Journal of Natural Language Processing*, 6(3), pp. 145--179.

Maroto, N., Alcina, A. (2005). Formal description of conceptual relationships with a view to implementing them in the ontology editor *Protégé*. In *Terminology* 15(2), pp. 232--257.

Mehler, A.; Sharoff, S.; Santini, M. (Eds.) (2010). Genres on the Web: Computational Models and Empirical Studies. Amsterdam: Springer.

Nakagawa, H., Mori, T. (2003). Automatic Term Recognition based on Statistics of Compound Nouns and their Components. In *Terminology*, 9(2), pp. 201–219.

Rehm, G., Santini, M., Mehler, A., Braslavski, P., Gleim, R., Stubbe, A. et al. (2008). Towards a Reference Corpus of Web Genres for the Evaluation of Genre Identification Systems. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC '08)*, Marrakech, pp.351--358.

Sartori, F. (2009). A Comparison of Methods and Techniques for Ontological Query Expansion, in Sartori, F.; Sicilia, M.A.; Manouselis, N. (Eds.), MTSR 2009, CCIS 46, pp. 203--214.

Staab, S.; Studer R. (Eds.) (2003). Handbook on ontologies. Dordrecht: Springer.

Tavosanis, M. (2011). L'italiano del web. Roma: Carocci.

Vossen, P. (Ed.) (1998). EuroWordNet: a multilingual database with lexical semantic networks. In *Computer and Humanities* 32 (2/3).

Winston, M., Chaffin, R., Herrmann, D. (1987). A Taxonomy of Part-Whole Relations. In *Cognitive Science*, 11, pp. 417--444.

Wright, S.E., Budin, G. (1997). Handbook of Terminology Management, Vol.2. Amsterdam: John Benjamins Publishing Company.