Approaches towards a "Lexical Web": the role of Interoperability

Nicoletta Calzolari
Istituto di Linguistica Computazionale del
CNR / 56100 Pisa, Italy
glottolo@ilc.cnr.it

Abstract

After highlighting some of the major dimensions that are relevant for Language Resources (LR) and contribute to their infrastructural role, I underline some priority areas of concern today with respect to implementing an open Language Infrastructure, and specifically what we could call a "Lexical Web". My objective is to show that it is imperative to define an underlying global strategy behind the set of initiatives which are/can be launched in Europe and world-wide, and that it is necessary an all-embracing vision and a cooperation among different communities to achieve more coherent and useful results. I end up mentioning two new European initiatives that go in this direction and promise to be influential in shaping the future of the LR area.

1 Language Resources: major dimensions

Only in the '90s LRs started to be considered as the necessary platform on which technologies and applications are built, a recognition which is nowadays widely accepted for the takeoff of our field. The following types of initiatives were then considered the major building blocks to set up a LR infrastructure (Calzolari and Zampolli, 1999):

i) *Standards for LRs*: the concept of reusability – directly related to the importance of "large scale" LRs within the increasingly dominant data-driven approach – has contributed significantly to the structure of many R&D efforts, such as EAGLES, ISLE, the recent LIRICS (e-Content), the ISO-TC37/SC4 committee.
ii) *LR construction*: projects such as WordNet, PAROLE, SIMPLE, LC-Star, EuroWordNet.
iii) *LR distribution*: LDC (Linguistic Data Consortium) in US, ELRA (European Language Resources Association) in Europe.

Other dimensions were soon added as necessary complement to achieve the required robustness and data coverage and to assess results obtained with current methodologies and techniques, i.e.:

iv) *Automatic acquisition of LRs* or of linguistic information: projects such as ACQUILEX, SPARKLE, ECRAN.
v) *Use of LRs for evaluation* campaigns, such as MUC, TREC, CLEF, Senseval, ACE.

1.1 Success of the Field

The very large body of initiatives of the last two decades (Calzolari, 1998 for an overview) was instrumental for the formation of a "LR community", and gave rise to a set of international initiatives of a global nature, encompassing many various perspectives on LRs or dealing with policy and meta-level issues related to LRs, such as:

- The Thematic Network ENABLER, grouping European National projects on LRs;
- The LREC Conference (about 900 participants in Lisbon-2004 and Genova-2006);
- The Asian Federation of Natural Language Processing (AFNLP);
- Bodies such as COCOSDA (International Committee for the Coordination and Standardisation of Speech Databases and Assessment Techniques) and WRITE (Written Resources Infrastructure, Technology and Evaluation);
- The new journal *Language Resources and Evaluation* (Ide and Calzolari, 2005).

Not to mention the ever-increasing role of LRs in statistical and empirical methods, and the growing industrial interest in using LRs and standards, specially for multilingual applications.

The flourishing of international projects and activities contributed to substantially advance knowledge and capability of how to represent, create, acquire, access, tune, maintain, standardise, etc. large lexical and textual repositories. There are today countless initiatives in the LR field, but we must admit that they are somehow scattered, opportunistic, often unconnected, with no real ability to build on each other and to form a unified space of LRs. We thus recognise that the LR infrastructure is still a virtual one. There is no real global coordination of efforts, and no body able to create the needed synergies among the various initiatives.

On the other side, the success itself of the field, its vitality and richness, coupled with the lack of coordination and of strategic thinking about future orientations, show that it is time to reflect again on the field as a whole, and ask ourselves which are/will be the major driving forces of today and of tomorrow to give the field the necessary cohesion.

1.2 Need of a Change

The wealth of LRs, in comparison with few years ago, but coupled with the shortage, even now, of a) new types of LRs, b) multilingual LRs, c) LRs of much larger size, d) LRs with richer annotations, and so on, points towards the need to consider whether those mentioned above are still the major driving forces. Which new building blocks do emerge today? I believe that those dimensions are still relevant, even if with an obvious evolution. Emerging pillars in current HLT are:

i) *Interoperability*, and even more *content interoperability*: language is the key mediator to access content, knowledge, ontologies;
ii) *Collaborative creation and management of LRs*, even on the model of wiki initiatives;
iii) *Sharing of LRs*, as a new dimension of the distribution notion;
iv) *Dynamic LRs*, able to auto-enrich themselves;

and finally the more comprehensive notion of:

v) *Distributed architectures and infrastructures for LRs*, encompassing and exploiting the realisation of the previous notions.

I will mention in the last section two new European initiatives where such notions will play a prominent role, could be at the basis of a new paradigm for LRs and language technology (LT) and influence the setting up of a "real" infrastructure.

2 Some Tendencies and Driving Forces in the Lexical Domain

Mixing considerations on what is needed for a broad language infrastructure and for a "lexical web" – undoubtedly a key part of it –, I touch here issues relevant to establishing a lexical web. I do that by pointing at research activities carried out at ILC in Pisa[1] showing a variety of approaches to lexical resources, involving: i) procedures for linking and integrating existing lexicons, ii) standardisation, iii) rel ation between lexical and terminological or ontological resources, iv) "ontologisation" of lexicons, v) architectures for managing, merging, integrating lexical resources.

2.1 Integration/Unification of Existing Lexicons

The market is increasingly calling for new types of lexicons that can be built rapidly – tailored to specific requirements – possibly by combining certain types of information from available lexicons while discarding others. This need could be satisfied exploiting the richness of existing lexicons, aiming at attaining their integration or virtual unification.

ELRA Unified Lexicon. An initiative in this direction, the *Unified Lexicon* project, has been carried out at ELRA by its Production Committee (Monachini et al, 2006). This experiment consisted in linking the LC-Star and PAROLE lexicons to set up a methodology to connect Spoken and Written LRs, thus establishing new models of LR distribution. In the envisaged scenario the same lexicons may be made available to different users, who can select different portions of the same lexicon or combine information coming from different lexicons. In this scenario lexical resources can be shared, are reusable and openly customisable, instead of being static and closed.

Linking ItalWordNet and SIMPLE Semantic Lexicons. The two largest and extensively encoded Italian lexicons, ItalWordNet (IWN) and PAROLE-SIMPLE-CLIPS (PSC), although developed according to two different lexical models, present many compatible aspects. Linking – and eventually merging – these lexicons in a common representation framework means to offer the end-user more exhaustive lexical information combining poten-

[1] Many passages in this section are taken from various papers, listed in the References, of ILC colleagues.

tialities and outstanding features offered by the two lexical models (Roventini et al, 2007). Not only reciprocal enhancements are obtained, but also a validation of the two resources. Their semantic integration is all the more desirable considering their multilingual vocation: IWN is linked to wordnets for many other languages, and PSC shares with 11 European lexicons theoretical model, representation language, building methodology and a core of entries.

Mapping the Ontologies and the Lexicons. Due to a different organisational structure of the two ontology-based lexicons, the linking process involves elements having a different status, i.e. autonomous semantic units in PSC and synsets in IWN. Mapping is performed on a semantic type-driven basis: comparing their ontological framework and establishing correspondences between the conceptual classes of both ontologies, with a view to further matching their respective instances, using also 'isa' relations and semantic features. The result of the first phase, linking concrete entities, sounds promising since 72.32% of the word-senses have been successfully linked.

The linking process makes it possible to enrich each resource by complementary information types peculiar to the other's theoretical model. In IWN, the richness of sense distinctions and the consistency of hierarchical links are remarkable. SIMPLE focuses on richly describing the meaning and semantic context of a word and on linking its syntactic and semantic representation, crucial for most NLP applications. Moreover, the mapping lets inconsistencies emerge, allowing to amend them. The linking process implies a de facto reciprocal assessment of both coverage and accuracy, particularly relevant to hand-built lexicons.

Differences regarding the nature of linking units, the granularity of sense distinction and the ontological typing are complex issues that are also being addressed during the linking process.

2.2 Interoperability: at the Heart of the Field

We have made big steps forward with respect to interoperability. Work started in EAGLES and ISLE (www.ilc.cnr.it/EAGLES96/isle/) (Calzolari et al, 2003) is being recently consolidated in true international ISO standards.

ISO. The Working Group ISO TC37/SC4/WG4 dedicated to NLP lexicons is in charge of defining lexical standards. The result is the LMF (Lexical Markup Framework) standard (Francopoulo et al, 2006). To cope with the challenge that actual lexicons differ very much both in complexity and in type of encoded information, a modular organization was adopted. As a consequence, LMF (http://lirics.loria.fr/documents.html) is made up of a core model, a sort of simple skeleton, and various semi-independent packages of notions, used for the various linguistic layers that make up a lexicon.

Lexical specifications are split in separate object types: LMF defines the lexical structure and is kept simple, while the huge amount of attributes (e.g. Part-of-Speech) are recorded in a data category registry where the peculiarities of languages and linguistic schools can be recorded. This registry, common to all TC37/SC4 standards, guarantees interoperability between lexicon and corpus annotation. An XML DTD is based on the UML modelling. Moreover, an OWL format has been defined that can be smoothly integrated into Semantic Web applications.

NEDO. While EAGLES and ISLE dealt with European languages, the Japanese NEDO project (Tokunaga et al, 2006), that develops international standards for Semantic Web applications, is specifically geared to Asian languages: Chinese, Japanese, Thai. It applies and refines ISO standards so that they are adapted to Asian languages.

But true content interoperability is still far away. We may have solved the issue of formats, of inventories of linguistic categories for the various linguistic layers, but have not solved the problem of relating senses, that only would allow automatic integration of semantic resources. This is a challenge for the next years, and a prerequisite for both a true Lexical Web and a credible Semantic Web.

2.3 Lexicons vs. Terminologies: a Continuum

Due to the strategic relevance of the biomedical field, intensive research is being carried out worldwide to develop LTs to access its large body of literature and extract knowledge from it. Access to and interoperability of biological databases, however, is still hampered by lack of uniformity and harmonisation of both formats and information encoded. A current demand in bioinformatics is to construct a comprehensive and incremental resource which integrates bio-terms encoded in existing different databases. A challenge is to encode all relevant properties of bio-terms according to the most accredited standards for the representation of

lexical, terminological and conceptual information.

BioLexicon. Working in the bio-domain in the European BOOTStrep project[2], we assume that the linguistic side of terminologies is partially informed by the knowledge of the domain and we claim that semantic relations, especially those accounting for the syntagmatic relations of words in context, are crucial for the representation of this kind of information. We also argue (Monachini et al, 2007) that a privileged representational device for encoding these relations is the set of Qualia Relations, as encoded in the SIMPLE general lexicon. These assumptions are made operational in the design of the *BioLexicon*: building a comprehensive terminological resource for the biomedical domain – with morphological, syntactic, semantic descriptions of the terms – which adheres to lexical and ontological standards and links concepts to lexical items is a huge scientific challenge. The BioLexicon (Quochi et al, 2007) is a large-scale resource that combines terminological data coming from bio-databases (mostly UniProt, Swiss-Prot, ChEBI, BioThesaurus and NCBI taxonomy) enriched with lexical information extracted from texts. The lexicon model is designed so as to integrate both typical information provided by domain ontologies and linguistic information available in open-domain computational lexicons: terms and variants are encoded with their semantic information as well as with typical linguistic information such as Part-of-Speech, subcategorisation frames and qualia relations that can be further augmented and tuned to cope with domain specific semantic information.

The model – flexible enough to adapt to different application needs, e.g. text-mining, information extraction, information retrieval, multilingual access – builds on previous experience in the standardisation and construction of lexical resources. Both the conceptual model and its physical implementation are tailored to the automatic population of the resource, independently of the various native data formats. The DB is modular and can automatically upload new data and provide (XML) outputs by means of web services. An *XML interchange format* (XIF) has been designed with the purpose of automatically populating the BioLexicon with data provided by domain experts and by lexical acquisition systems, therefore allowing for a standardisation of the data extracted from the different terminological resources and from texts.

The goal is to propose a standard for the representation of lexicons in the bio-domain, which could eventually be also interoperable with other domain lexicons. For this reason the ISO LMF was chosen as the reference meta-model and the ISO Data Categories as the main building blocks for the representation of the entries. A reusable BioLexicon with sophisticated linguistic information, linked to a bio-ontology, should enable the bio-informatics community to develop information extraction tools of higher quality.

2.4 Lexicons and Ontologies : a Dilemma

Ontologies are recognised as an important component in NLP systems that deal with the semantic level of language (Huang et al, to appear). Most semantic lexical resources (e.g. WordNet, CYC, SIMPLE), have in common the presence of an ontology as a core module. Besides, there is a lot of research in progress on applying ontologies to semantic NLP. The fact that OWL is the ontology language for the Semantic Web and that it provides a formal semantic representation as well as reasoning capabilities has encouraged the NLP community to convert existing resources to this language.

RDF/OWL representation of WordNet. WordNet has recently received a growing attention by the Semantic Web community. Within W3C, WordNet has been translated (Van Assen et al, 2006) in the standard semantic languages RDF/ OWL, which can describe collections of resources on the Web and are convenient data models to represent highly interconnected information and their semantic relations. Moreover, the RDF/OWL representation of WordNet is easily extensible, allows for interoperability and makes no assumptions about particular applications. The availability of WordNet Web Services can be an important step for its integration and effective use into the Semantic Web, and for future multilingual semantic interoperability in the Web (Marchetti et al, 2006).

"Ontologisation" of lexicons. A new initiative at ILC (Toral and Monachini, 2007) is the conversion into OWL of the ontology of SIMPLE, the lexico-semantic resource based on the Generative Lexicon (GL). The elements of SIMPLE modelled in OWL are those of the original ontology, i.e. se-

[2] BOOTStrep (Bootstrapping Of Ontologies and Terminologies STrategic Project) is an IST European project under the 6th Framework Programme (www.bootstrep.eu).

mantic types, qualia relations, semantic features. A challenge in the ontology design is that its nodes are not only defined by their formal dimension (taxonomic hierarchy), but also by the GL qualia dimensions: constitutive, telic, agentive. The OWL ontology is also enriched in a bottom-up approach that extracts further semantic information (e.g. selected constraints on relations and features extracted from the lexicon) by exploring the word-senses that belong to each semantic type and by using the qualia structure as a generative device.

This research aims at the representation of a lexicon based in the GL theory into the Semantic Web ontology language, with reasoning capabilities interfaced to a lexicon. This allows the ontology to be processed and checked by standard reasoners. This is useful for Semantic Web applications, semantic NLP tasks, and for enhancing the quality of the lexicon by validating it (through reasoning one can look for inconsistencies). The ontology is also a key element of a broader forthcoming research aimed at automatic lexico-semantic-driven text mining and knowledge acquisition procedures, which, in their turn, have the goal of gathering knowledge to enrich the lexicon, thus creating a virtuous circle between lexicon/ontology and corpus-based information acquisition.

2.5 Architectures for Integration of Lexicons

Enhancing the development of multilingual lexicons is of foremost importance for intercultural collaboration (Bertagna et al, 2007), as they are the cornerstone of several multilingual applications. Nevertheless, large-scale multilingual lexicons are not yet as widely available as needed. A new trend tries to exploit the richness of existing lexicons, in addition to creating new ones. At the same time, as the history of the web teaches, it would be a mistake to create a central repository of all the shared lexicons, while distribution of resources is a crucial concept. A solution emerging in the LR community consists in moving towards distributed language services, based on open content interoperability standards, and made accessible to users via web-service technologies. There is another deeper argument in favour of distributed lexical resources: LRs are inherently distributed because of the diversity of languages over the world. It is natural that LRs are developed and maintained in their native environment. It is not possible to describe the current state of a language, evolving over time, away from where it is spoken.

Web services for LRs or LRs as web services. Having lexicons available as web services would allow to create new resources on the basis of existing ones, to exchange and integrate information across repositories and to compose new services on demand: an approach towards the development of an infrastructure built on top of the Internet in the form of distributed language services is presented in Ishida (2006). This new type of LRs can still be stored locally, but their maintenance and exploitation can be a matter of agents choreographed to act over them. Admittedly, this is a long-term scenario requiring the contribution of many actors and initiatives (among which we mention standardisation, distribution, international cooperation). A first prerequisite for this scenario to take place is to ensure true interoperability among lexicons, a goal that would be now mature for many aspects. Although the paradigm of distributed and interoperable lexicons has largely been discussed and invoked, little has been made for its practical realisation. Some initial steps to design frameworks enabling interlexica access, search, integration, operability are: the Lexus tool (Kemps-Snijders et al, 2006), based on LMF, managing the exchange of data among large-scale lexical resources, and SHAWEL (Gulrajani and Harrison, 2002), tailored to the collaborative creation of lexicons for endangered language. However, the impression is that little has been made towards the development of new methods, techniques and tools for attaining a real interoperability among lexical resources.

LeXFlow. The design of an architecture able to turn into reality the vision of shared and distributed lexical repositories is a very challenging task. To meet these needs, we have designed and built a distributed architecture, *LeXFlow*, enabling a rapid prototyping of cooperative applications for integrating lexical resources (Soria et al, 2006). It is based on a web-service architecture, fostering integration and interoperability of computational lexicons, focusing on mutual linking and cross-lingual enrichment of distributed monolingual lexicons. As case–studies, we have chosen to work with:

i) two Italian lexicons based on different models, SIMPLE and ItalWordNet, and
ii) two lexicons belonging to the WordNet family, ItalWordNet and the Chinese Sinica BOW.

These represent different opportunities of adopting a bottom–up approach to exploring interopera-

bility for lexicon augmentation and mutual enrichment of lexical resources, either i) in a cross-model or ii) in a cross-lingual enrichment/ fertilisation of monolingual lexicons.

Multilingual WordNet Service. This module is responsible for the automatic cross-lingual fertilisation of lexicons with a wordnet-like structure. Put it very simply, the idea behind it is that a monolingual WordNet can be enriched by accessing the semantic information encoded in corresponding entries of other monolingual WordNets. The various WordNet-lexicons reside over distributed servers and can be queried through web service interfaces. The entire mechanism is based on the exploitation of the Interlingual Index (ILI). The proposal to make distributed WordNets interoperable allows applications such as:

– *Enriching existing resources*. Information is not complete in any WordNet: by making WordNets interoperable we can bootstrap semantic relations and other information from other WordNets.
– *Creation of new resources*. Multilingual lexicons can be bootstrapped by linking different language WordNets through the ILI.
– *Validation of resources*. Semantic relations and synset assignments can be validated if reinforced by data coming from other WordNets.

This work can be a prototype of a web application to support the *Global WordNet Grid* initiative (www.globalwordnet.org/) (Fellbaum and Vossen, 2007), whose success depends on whether there will be tools to access and manipulate the rich internal semantic structure of distributed multilingual WordNets. *LeXFlow* offers such a tool, providing interoperable web-services to access distributed WordNets on the grid. This allows to exploit in a cross-lingual framework the wealth of monolingual lexical information built in the last decade. As an example of use, a multilingual query given in Italian but intended for querying English, Chinese, French, German, and Czech texts, can be sent to 5 different nodes on the Grid for query expansion, as well as performing the query itself. This way, language-specific query techniques can be applied in parallel to achieve results that can be then integrated. As multilingualism clearly becomes one of the major challenges of the future of web-based knowledge engineering, WordNet emerges as a leading candidate for a shared platform, representing a simple and clear lexical knowledge model for different languages. This is true even if it has to be recognised that the WordNet model is lacking some important semantic information (like a way to represent semantic predicates). In LeXFlow we presuppose a de-facto standard, i.e. a shared and conventionalised architecture. Since the WordNet framework is both conventionalised and widely followed, our system is able to rely on it without resorting to a more substantial and comprehensive standard. In the case, however, of integration of lexicons with different underlying linguistic models, the availability of MILE (Calzolari et al, 2003), now of LMF, is an essential prerequisite.

From a more general viewpoint, we must note that the realisation of the new vision of distributed and interoperable LRs is strictly intertwined with at least two prerequisites. On the one side, LRs need to be available over the web; on the other, the LR community will have to reconsider current distribution policies, and investigate the possibility of developing an "Open Source" concept for LRs.

UIMA. Finally, we have started an initiative, at ILC, to integrate both various LRs (lexicons, ontologies, corpora, etc.) and different NLP tools into a common framework of shared and distributed resources, the IBM UIMA middleware (Ferrucci and Lally, 2004). As case study, a first prototype for a UIMA Type System has been built to manage TimeML categories and integrate an Italian Treebank and the SIMPLE lexicon (Caselli et al, 2007). Both a web interface for human access and a series of web services for machine use are being developed. This research intends to contribute both to a UIMA type systems standardisation and to a common framework for resource and tool sharing and interoperability definition. This initiative is linked with the NICT Language Grid project (Ishida, 2006), from which our prototype inherits the service ontology environment.

3 First steps for a LR Infrastructure

Finally, new conditions are emerging, in Europe, that could turn what is so far a virtual LR infrastructure into a real one (Calzolari, 2007). This tendency is helped not only by new technical conditions, but also by the recognition that any organisation has limited resources, and will never be able to create all the necessary infrastructural resources – in adequate quality – as needed. These may instead be spread across several organisational

units.

Sensitivity of LRs: political, economic, social, strategic factors. Behind the notion of "distributed" resources there are also political (very sensitive) factors, behind resources that can be "shared/ and reused" economic factors. Moreover, many today start bringing into focus also the social value of a common infrastructure, and strongly advocate – contrary to current practice – the benefits of open access (vs. the social costs of restricted access). In addition to its scientific implications, the large intellectual, cultural, economic movement behind LRs entails "strategic" thinking, and urges to reflect on field of LRs from a very broad angle. It is perceived as essential to define a general plan for research, development and cooperation in the LR area, to avoid duplication of efforts and provide for a systematic distribution and sharing of knowledge. To ensure reusability, the creation of standards is still the first priority. Another tenet is the recognition of the need of a global strategic vision, encompassing different types of – and different methodologies of building – LRs, for an articulated and coherent development of this field.

Two new European initiatives are linked to these ideas.

3.1 CLARIN

CLARIN (*Common Language Resource and Technology Infrastructure*) (http://www.mpi.nl/clarin/) is an ESFRI project whose mission is to create an infrastructure that makes LRs and LTs easily usable to scholars of all disciplines, in particular of the humanities and social sciences, to prepare an eScience scenario. The purpose is to offer persistent and secure services and provide easy access to LRs and LTs. CLARIN proposes to make this vision a reality: the user will have access to repositories of data with standardised descriptions, processing tools ready to operate on standardised data, and guidance from distributed knowledge centres. All this will be available on the web using a service oriented architecture based on secure grid technologies. CLARIN will turn existing, fragmented LRs and LTs into accessible, stable services that any user can share, adapt and repurpose, building upon the rich history of European and national initiatives. The preparatory phase aims at bringing the project to the required level of legal, organisational and financial maturity. This necessitates an approach along various dimensions in order to pave the way for implementation. Infrastructure building is a time-consuming activity and only robustness and persistency of the offered solutions will convince researchers and users.

3.2 FLaReNet

International cooperation and re-creation of the LR community are among the most important drivers for a coherent evolution of the LR area in the next years. The Thematic Network *FLaReNet (Fostering Language Resources Network),* proposed in the context of an eContent*plus* call, will act as a European forum to facilitate interaction among LR stakeholders. Its structure considers that LRs present various dimensions and must be approached from many angles: technical, but also organisational, economic, legal, political, addressing also multicultural and multilingual aspects, essential when facing access and use of digital content in today's Europe. FLaReNet, organised into working groups focusing on specific objectives, will bring together leading experts (academic and industrial) to ensure, in cooperation with CLARIN, coherence of LR-related efforts in Europe. FLaReNet will consolidate existing knowledge, presenting it analytically and visibly, and will contribute to structuring the area of LRs of the future by discussing new strategies to: convert existing and experimental technologies related to LRs into useful economic and societal benefits; integrate so far partial solutions into broader infrastructures; consolidate areas mature enough for recommendation of best practices; anticipate the needs of new types of LRs.

The outcome of FLaReNet will be of a directive nature, to shape the future of the LR area, and help the EC, and national funding agencies, to identify the priority areas of LRs that need public funding to develop and improve. A blueprint of actions will give input to policy development both at EU and national level for identifying new language policies that support linguistic diversity in Europe, in combination with strengthening the language product market and introducing innovative services, especially for less technologically advanced languages.

References

Bertagna, F., Monachini, M., Soria, C., Calzolari, N., Huang, C., Hsieh, S., Marchetti, A., Tesconi, M., Fostering Intercultural Collaboration: a Web Service

Architecture for Cross-Fertilization of Distributed Wordnets. In Ishida, T., Fussell, S. R., Vossen, P. (eds.) *Proceedings of the First International Workshop on Intercultural Collaboration (IWIC 2007)*, Kyoto, pp. 185-198. Also in: LNCS, vol. 4568, pp. 146-158. Springer, 2007.

Calzolari, N., An overview of Written Language Resources in Europe: a few reflections, facts, and a vision. In *Proceedings of the First LREC*, pp. 217-224. Granada, 1998.

Calzolari N., Towards a new generation of Language Resources in the Semantic Web vision. In Ahmad, K., Brewster, C., Stevenson, M. (eds.), *Words and Intelligence II: Essays in honour of Yorick Wilks*, pp. 63-105. Springer, 2007.

Calzolari, N., Bertagna, F., Lenci, A., Monachini, M. (eds.), *Standards and Best Practice for Multilingual Computational Lexicons. MILE (the Multilingual ISLE Lexical Entry).* ISLE, Pisa, 194 pp., 2003.

Calzolari, N., Zampolli, A., Harmonised large-scale syntactic/semantic lexicons: a European multilingual infrastructure. In *MT Summit Proceedings*, pp. 358-365. Singapore, 1999.

Caselli, T., Prodanof, I., Ruimy, N., Calzolari, N., Mapping SIMPLE and TimeML: improving event identification and classification using a semantic lexicon. In *GL2007: Fourth International Workshop on Generative Approaches to the Lexicon*. Paris, 2007.

Fellbaum, C., Vossen. P., Connecting the Universal to the Specific: Towards the Global Grid. In Ishida, T., Fussell, S. R., Vossen, P. (eds.) *Proceedings of IWIC 2007*. Also in: LNCS, 2007.

Ferrucci, D., Lally, A., UIMA: an architectural approach to unstructured information processing in the corporate research environment. *Natural Language Engineering*. 10(3-4) 2004.

Francopoulo, G., George, M., Calzolari, N., Monachini, M., Bel, N., Pet, M., Soria, C., Lexical Markup Framework (LMF). In *Proceedings of LREC2006*, Genova, pp. 233-236. ELRA, Paris, 2006.

Gulrajani, G., Harrison, D., SHAWEL: Sharable and Interactive Web-Lexicons. In *Proceedings of the LREC2002 Workshop on Tools and Resources in Field Linguistics*, Las Palmas, pp. 1-4, 2002.

Huang, C.R., Calzolari, N., Gangemi, A., Lenci, A., Oltramari, A., Prévot, L. (eds.), *Ontologies and the Lexicon*. Cambridge Studies in NLP. Cambridge University Press, Cambridge, to appear.

Ide, N., Calzolari, N., Introduction to the Special Inaugural Issue. *Language Resources and Evaluation*. Springer, 39(1), pp. 1-7, 2005.

Ishida, T., Language Grid: An Infrastructure for Intercultural Collaboration. In *IEEE/IPSJ Symposium on Applications and the Internet*, pp. 96-100, 2006.

Kemps–Snijders, M., Nederhof, M., Wittenburg, P., LEXUS, a web-based tool for manipulating lexical resources. In *Proceedings of LREC2006*, Genova, pp. 1862-1865. ELRA, Paris, 2006.

Marchetti, A., Tesconi, M., Ronzano, F., Rosella, M., Bertagna, F., Monachini, M., Soria, C., Calzolari, N., Huang, C.R., Hsieh, S.K., Towards an Architecture for the Global-WordNet Initiative. In *Proceedings of SWAP-06, 3rd Semantic Web Workshop*. 2006.

Monachini, M., Calzolari, N., Choukri, K., Friedrich, J., Maltese, G., Mammini, M., Odijk, J., Ulivieri, M., Unified Lexicon and Unified Morphosyntactic Specifications for Written and Spoken Italian. In *Proceedings of LREC2006*, Genova, pp. 1852-1857. ELRA, Paris, 2006.

Monachini, M., Quochi, V., Ruimy, N., Calzolari, N., Lexical Relations and Domain Knowledge: The BioLexicon Meets the Qualia Structure. In *GL2007: Fourth International Workshop on Generative Approaches to the Lexicon*. Paris, 2007.

Quochi, V., Del Gratta, R., Sassolini, E., Monachini, M., Calzolari, N., Toward a Standard Lexical Resource in the Bio Domain. In Vetulani, Z. (ed.), *Proceedings of 3rd Language and Technology Conference*, Poznań, pp. 295-299, 2007.

Roventini A., Ruimy N., Marinelli R., Ulivieri M., Mammini M., Mapping Concrete Entities from PA-ROLE-SIMPLE-CLIPS to ItalWordNet: Methodology and Results. In *Proceedings of the 45th Annual Meeting of the ACL*, pp. 161-164. Prague, 2007.

Soria, C., Tesconi, M., Marchetti, A., Bertagna, F., Monachini, M., Huang, C., Calzolari, N., Towards agent-based cross-lingual interoperability of distributed lexical resources. In *Proceedings of COLING-ACL Workshop on Multilingual Lexical Resources and Interoperability*, Sydney, 2006.

Tokunaga, T., Sornlertlamvanich, V., Charoenporn, T., Calzolari, N., Monachini, M., Soria, C., Huang, C., Prevot, L., Xia, Y., Yu, H., Kiyoaki, S., Infrastructure for standardization of Asian language resources. In *Proceedings of COLING/ACL 2006 Main Conference Poster Sessions*, pp. 827-834. Sydney, 2006.

Toral, A., Monachini, M., Formalising and bottom-up enriching the ontology of a Generative Lexicon. In *Proceedings of RANLP07 - Recent Advances in Natural Language Processing*. Borovets, Bulgaria, 2007.

Van Assem, M., Gangemi, A., Schreiber, G., Conversion of WordNet to a standard RDF/OWL representation. In *Proceedings of LREC2006*, Genova. ELRA, Paris, 2006.