

\*Multilinguality and Interoperability in Language Processing with  
Emphasis on Romanian\*

Dan Tufiş and Corina Forăscu (eds.)

Romanian Academy Publishing House, Bucharest, 2010

ISBN: 978-973-27-1972-5

-----  
\*Bootstrapping and Collaboratively Enriching  
the Italian Domain WordNet through  
the WiKyoto Knowledge Editor\*

Francesco RONZANO\*\* , Monica MONACHINI\* , Andrea MARCHETTI\*\* ,  
Maurizio TESCONI\*\*\* , and Nicoletta CALZOLARI\*  
\* Consiglio Nazionale delle Ricerche (CNR), Institute of Computational  
Linguistics (ILC)  
\*\* Consiglio Nazionale delle Ricerche (CNR), Institute of Informatics  
and Telematics (IIT)  
Via Moruzzi, 1 - CNR, Research Area  
Pisa, Italy  
E-mail: {monica.monachini,nicoletta.calzolari}@ilc.cnr.it,  
{francesco.ronzano, andrea.marchetti, maurizio.tesconi}@iit.cnr.it

\*Abstract\*

Enhancing the development of multilingual resources is of utmost importance for use in computer applications. The need of ever growing resources for effective multilingual content processing has given impulse to a radical change in the perspective of language resource (LR) creation, structuring, exploitation and maintenance. The Web has played a key role in this process: indeed the possibility to access growing amounts of structured and unstructured data as well as the ease of creating and sharing contents between distributed communities of users have strongly affected the methodologies and techniques to bootstrap, enrich and access LRs. From static knowledge bases usually created and maintained by groups of experts and tailored to the specific exploitation contexts, LRs have turned into dynamic repositories of linguistic knowledge. Their content is usually easily accessible over the Web and often exploited, aggregated and optimized on-the-fly by on-line information mining services. In this context, the adoption of standardized data formats to facilitate interoperability and data exchange is essential. Moreover, the creation and maintenance of these resources has taken great advantage from the possibility to harvest Web data in order to bootstrap or enrich them. Several new frameworks have been proposed to support access, search, integration and interoperability of "new generation" LRs. Wide distributed communities of Web users are more and more directly or indirectly involved in keeping language resources updated or in extending them. After a brief description of modern LRs, we focus our attention on two essential issues involving them: the need for standard formats that support interoperability in a distributed Web context and the possibility for the Web communities to collaboratively maintain and enrich these resources. In particular, we present the Italian WordNet (IWN) and its exploitation in the context of the KYOTO Project, as a real-world scenario where standardization, interlinking, enrichment as well as collaborative editing are put into practice. The KYOTO Project is a complex knowledge-driven environment built with the aim of enabling communities of users to mine information from textual documents, sharing

the collected facts across cultures, languages and domains. The semantic ground supporting all the information mining tasks of KYOTO is constituted by the Multilingual Knowledge Base, composed by a collection of WordNets encoding language-specific lexical patterns for each language covered by KYOTO. All of them are mapped to the language-independent entities of the KYOTO Central Ontology. In the context of KYOTO, we describe the process followed to define WordNet-LMF (WN-LMF), the standard format tailored to represent lexical resources adhering to the lexical knowledge WordNet model, useful to easily integrate general and domain lexicons in KYOTO. We present the conversion of the IWN to the WN-LMF standard, as a necessary pre-requisite for IWN to be integrated in the Multilingual Knowledge Base. We expose a (semi)-automatic procedure which allows IWN to upgrade ILI connections to the last version available of the Princeton English WordNet, 3.0. We also consider the Species2000 SKOS thesaurus, a knowledge resource with a data structure different from WordNet: we present its conversion to WN-LMF. To enable the multilingual and multicultural community of KYOTO users to maintain and extend KYOTO knowledge resources, we introduce the Wikyoto Knowledge Editor: it is the Web-based wiki environment useful to navigate, collaboratively enrich the Multilingual Knowledge Base. We describe its Web interface by a practical use case concerning the extension of the Italian Domain WordNet.