Content Interoperability of Lexical Resources: Open Issues and “MILE” Perspectives

Francesca Bertagna[1], Alessandro Lenci[2], Monica Monachini[1], Nicoletta Calzolari[1]

[1]Istituto di Linguistica Computazionale (ILC) – Consiglio Nazionale delle Ricerche
Via Moruzzi 1, 56100 Pisa, Italy
(francesca.bertagna, monica.monachini, nicoletta.calzolari)@ilc.cnr.it
[2]Dipartimento di Linguistica, Università degli Studi di Pisa,
Via S. Maria 36, 56100 Pisa, Italy
alessandro.lenci@ilc.cnr.it

Abstract

The paper tackles the issue of content interoperability among lexical resources, by presenting an experiment of mapping differently conceived lexicons, FrameNet and NOMLEX, onto MILE (Multilingual ISLE Lexical Entry), a meta-entry for the encoding of multilingual lexical information, acting as a general schema of shared and common lexical objects. The aim is to (i) raise problems and (ii) test the expressive potentialities of MILE as a standard environment for Computational Lexicons.

1 Introduction

The exchange and integration of information between systems is known as "Interoperability" (Vckovski, 1999). While HTML and XML allow the access and interchange of data at the formal and structural level, a metadata representation language like RDF/S (further extended with ontology formalization capabilities, e.g. in OWL or DAML+OIL) is expected to enable a new and unprecedented progress towards content interoperability among resources. Such is the main vision of the Semantic Web: a wealth of new possibilities stemming from representing documents and data semantics with metadata defined within ontologies, which will be easy for a machine to interpret and make use of in an intelligent way (Lassila, 1998). Computational lexicons are repositories of syntactic and semantic information. Recently, there have been various efforts to translate existing lexical resources in RDF/S or in DAML+OIL, in the attempt to make their content available in the Semantic Web for various future applications (see Narayanan et al., 2002; Melnik&Decker at www.semanticweb.org/library). However, there is a concrete risk for these experiments to become mere conversion exercises, unless they are backed by an additional framework providing a common/shared compatible representation of lexical objects. Actually, in order to reach a truly content interoperability, intelligent agents must be provided with the possibility to manipulate the objects available in different lexical repositories understanding their deep semantics. This would entail, for instance, that applications should be enabled to understand whether two lexical objects are of the same type so that the same operations can be applied to them. In the paper we will tackle the issue of content interoperability among lexical resources by presenting an experiment of mapping differently conceived lexicons (in the particular case FrameNet and NOMLEX) to a general schema of shared and common lexical objects. The schema adopted in this experiment is MILE (Multilingual ISLE Lexical Entry), a meta-entry for the encoding of multilingual lexical information (Calzolari et al., 2003) developed within ISLE[1] (International Standards for Language Engineering). The aim of the experiment is to evidence problems and collect hints that may emerge while mapping lexicons against an abstract model, while testing the expressive potentialities of the MILE as a standard for computational lexicons.

2 MILE

The MILE Lexical Model (MLM) is described with Entity-Relationship (E-R) diagrams defining the entities of the lexical model and the way they can be combined to design an actual lexical entry. MLM defines a first repertory of "MILE Lexical Classes" (MLCs), which formalize the main building blocks of lexical entries. The MLCs are defined on the basis of an extensive survey of major existing practices in lexicon development. MLCs form a "top ontology of lexical objects", as an abstraction over different lexical models and architectures. The MLM defines each class by specifying its attributes and the relations among them. Classes represent basic lexical notions. Instances of MLCs are the "MILE Data Categories" (MDCs), each of them identified by a URI. MDCs can be either user-defined or reside in a shared repository. Part of the class structures in the MLM has been formalized as a RDF Schema, and data categories have been created using RDF and OWL (Ide et al., 2003).

3 The Mapping Experiment

Two main methodological scenarios concerning the mapping may be envisaged.

(1) The first implies to resort to a high level mapping of the elements in a lexicon onto the MILE lexical objects. This is similar to the proposal in (Peters et al 1998), i.e. a common object model, sitting on top of the resource-specific models, which allows a uniform access procedure for all the resources. In this approach, the expert of the specific lexicon takes a number of decisions concerning the mapping between the linguistic information in the

[1] ISLE was an initiative under the FP5 within the EU-US International Research Co-operation, with the aim to develop and promote widely agreed on Human Language Technology standards and best practice recommendations for infrastructural LRs.

lexicon and the set of available lexical objects in the abstract model. One of the main advantages of such a solution is that resources would retain their native structure, without being submitted to format conversion.

(2) In the second approach, the possibility provided by MILE of creating instances of the lexical classes can be exploited to create lexical entries directly in MILE, which thereby acts as a true interchange format.

The most appropriate mapping strategy clearly depends on the possible applicative scenarios in a distributed and open environment requiring lexical resources content interoperability. The first approach is actually most promising for a "smart" access to lexical repositories. In this sense, mapping the resource data model onto a common schema provide with an explicit formal characterization of object semantics would easy the off-line processing of extracting the required information.

On the contrary, the second approach would be more suitable for the purpose of managing, integrating and merging lexical information residing in different repositories. Creating lexical entries in an MILE-like schema would be a way to make available the semantics of each lexical entry in a fully explicit way, allowing intelligent computational agents to exploit it in inferential systems and knowledge-intensive applications. In what follows, we will present some preliminary results of the experiment we have undertaken to map FrameNet and NomLex onto MILE. We preferred to perform the mapping at lexical object level (following strategy (1), since it is expected that, once the mapping conditions are formally and totally explicitly defined, the conversion at the entry level would follow naturally.

3.1 FrameNet to MILE

Our first experiment concerns the possibility to map the FrameNet (FN) architecture to MILE.

In this paper, we have preferred not to involve in the mapping experiment two other important lexicon models: the WordNet "family" and the PAROLE/SIMPLE lexicons. From the beginning, one of the requirements for the standard was to perfectly represent WordNet notions of *synset* and semantic relations. In this sense, mapping WordNet to MILE is more straightforward and the interested reader can have an exemplification of it in (Lenci, 2003). At the same time, being the MILE architecture grounded on the GENELEX model, it perfectly adheres to SIMPLE. Representing FrameNet with the expressive modalities of MILE is a more difficult task.

FrameNet (Baker et al., 1998) is an important reality in the lexicon scenario and its linguistic design offers original features the standard has to deal with. The notion of Frame as such doesn't belong to the classes provided by MILE. Moreover, Narayan et al. (2002) offer us a ready set of DAML+OIL classes representing the FrameNet notions to work on. We will try to map the Frame, the Frame Element (FE) and the Lexical Unit (LU) on the correspondent MILE classes. The following picture shows how a certain degree of correspondence is possible.

The Frame can be represented by the MLC Predicate, the FrameElement by the Argument and the Lexical Unit by the SemU.

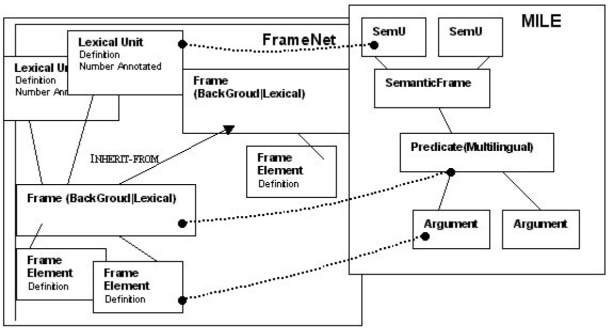

Fig. 1: Mapping FrameNet Lexical Objects to MILE

The Frame is an extended and complex structure of knowledge evoking what we may call "actantial scenarios" and playing the central role in the design of the resource. The Frame Elements are the "actants", the entities playing a part in the *scenario* evoked by the Frame. So, the Frame "Getting" represents a situation where "a *Recipient starts off without the Theme in their possession, and then comes to possess it* etc..". In this situation, the Frame Elements are the Recipient, the Theme and others. In MILE the notion that better expresses this same information is the Predicate. It can be *lexical* or *primitive* and it is linked to Arguments by means the *hasArgument* relation. If we want to use the Predicate to represent the Frame we have to choose its Primitive (non-lexical) modality. The MILE notions of Predicate, Argument and SemU are flexible enough to be interpreted in a narrower or in a wider way: the Predicate can be more close in size to the subcategorization frame or more extended and close to the notion of Frame and of *scenario*. In FN, the set of FE types is open in order to better fit the specific needs of the different Frame. Even though providing a recommended set of possible values for the Thematic Roles of the Arguments (derived from SIMPLE), MILE allows the user to independently choose the most appropriate values; in this way, the MLM allows the representation of the open set of Frame Element names. In mapping FrameNet onto MILE, some mismatches of formal nature emerge. For example, while in FrameNet the Lexical Unit is directly linked to the Frame, in MILE the Predicate is inserted in a more general class, the SemanticFrame, which is in between the SemU and the Predicate. It specifies the predicative argument structure of the lexical entry and de facto contains the Predicate (with its arguments) and the type of link between the SemU and the Predicate, expressed by means of the attribute *TypeOfLink*. As a matter of facts, different words belonging to different POSs may share the same predicate in the predicative representation[2]. The problem is that while in MILE the specification of the *TypeOfLink* is not optional, in FrameNet the nature of the link between the Lexical Unit and the Frame is underspecified (so we find in the same group the Lexical Units *to acquire*, *to gain*, *acquisition_act* etc., all sharing a membership to the same Frame GETTING). A possible solution is to add a new value (Underspecified) for the *TypeOfLink* attribute in MILE. A more serious problem consists in the lack of any inheritance or embedding mechanisms for the MILE

[2] For instance, the verb *destroy* and the nouns *destruction* and *destroyer* may share the same predicate DESTROY respectively with a MASTER, VERBNOM, and AGENTNOM type of link

Predicate. In FrameNet two types of relations among frames are possible: first of all, the various frames can be organized in a hierarchical way, exploiting a sort of IS-A relation among the frames: "*if frame B inherits from frame A, then B elaborates A, and is a subtype of A.*" (Narayanan et al., 2002). Moreover, a kind of sub-type relation can be established among a complex frame and several simpler frames (the so-called subframes). These important features of FrameNet cannot be represented using MILE: under this point of view, we can state that a complete "translation" from FrameNet to the standard cannot be successfully achieved. The *modularity* of the MILE, however, may be an answer to this problem: it would allow the addition, for instance, of a new object PredicateRelation to the LexicalModel. Even without the availability of a specific class SubPredicate, MILE would be able to represent the semantics of a predicate considered a part/sub-type of a more complex and articulated Frame. By envisaging specific relations among predicates, it would also be possible to express the temporal ordering among the frames (another information we can find in FN). In the next future, we would like to verify if also the FN strong correlation between lexical entry and corpus evidences (by means of annotation) is representable using MILE devices. We will discuss later of this aspect but surely the *flexibility* of the model (i.e. its being open to adaptations and improvements without changing the existent) is an important feature a standard should have in order to represent new linguistic notions and different lexicon "vision".

3.2 NOMLEX to MILE

The second experiment proposes a mapping between the MLCs and NOMLEX (Reeves *et al.* 1999), a syntactic lexicon for English nominalizations. NOMLEX has been designed similarly to COMLEX, a syntactic subcategorization lexicon for English verbs. Basically, the strong reason underlying the choice of such a lexicon for the mapping, is that NOMLEX has an architecture very far form the MILE E-R model: lexicon entries take the form of parenthesized, nested feature-value structures, allowing to express lexical information in a very synthetic and compact way. NOMLEX, basically, describes syntactic frames of nominalization and also relates the noun complements to the verb arguments. All this information, once mapped against the MILE basic notions, proves to be covered by their corresponding MILE Lexical Classes (MLCs). The immediate main divergence consists, hence, in the adopted expressive means. Whereas in the previous experiment, two lexicons both based on an E-R model but not with perfectly overlapping notions have been confronted, viceversa, here, the mapping has to deal with the same linguistic notions, expressed with two conceptually opposite lexicon structures. Another important diverging point characterizes the two lexicons: the definition of the clear cut between the levels of linguistic representation. In a NOMLEX lexical entry, not only purely syntactic properties are provided, but some semantic pieces of information enter into the description. In a same feature value, no clear boundaries between the syntactic and semantic parts are defined: as a consequence, the level of *interface* between syntax and semantics as well is partly hidden in the syntactic description of the lexicon. Conversely, in MILE the representation of lexical information is highly modular, flexible and layered, with notions distinctly distributed over different levels of linguistic representation. These differences make the experiment particularly challenging, thus giving the opportunity to better test the MILE model, in terms of adequacy, expressiveness and potentialities. By way of an example of the mechanism and efforts that two differently conceived lexical organizations involve, notwithstanding the mappability of the linguistic notions, one object class, shared by all NOMLEX entries, is mapped onto the MLCs. This is the class expressed by the feature :nom-type in which the type of nominalization is declared, i.e. if it expresses the event/state of the verb, if includes incorporations of a verb argument. Mutually exclusive values can be specified, depending on the different expressions and possible incorporations of the argument. Expressing that in the MILE model means to *decompress* the information and spread it over different MLCs, belonging to different lexical layers. According to the MILE architecture, indeed, the type of relation between a nominalization and its verb base is more properly of a semantic nature. It involves many MLCs, and, moreover, implies the level of interface between syntax and semantics. Next to an MLC:SynU, a corresponding MLC:SemU is needed, with the object CorrespSynUSemU to state a link between the two. From the SemU, the MLC:SemanticFrame branch out, dominating the MLC:Predicate and its connected MLC:Argument(s)[3]. Two attributes of the class SemanticFrame, the 'typeOfLink' and 'includedArg', respectively, are in charge of specifying the relation between the SemU and the Predicate and the incorporation of the argument. In Fig. 2, the values 'AGENTNOM' and '0' instantiate the agent nominalization[4]. The object CorrespSynUSemU, at this level of conceptual mapping, remains empty: if the mapping is pushed at the level of lexical entries it will be instantiated to specify the way the Syntactic and Semantic Frames correspond each other and, particularly, how semantic Arguments are projected on to the syntactic Slots.

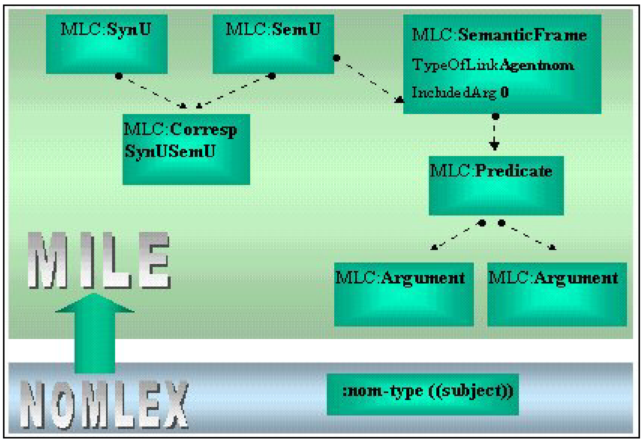

Fig. 2: A NOMLEX class mapped onto MILE MLCs.

[3] It should be noted that a verb and its nominalizations are supposed to share the same semantic frame.

[4] The mechanism applies to all the values: changing the value in NOMLEX means to change the value in one of the MILE MLCs.

The mapping between such a models is highly costly, since information expressed in a very compact and dense way should be explicitly decompressed and distributed over pertaining levels. This operation is due to the high level of granularity in MILE, which, however, has been thought up exactly to allow the compatibility with differently packaged linguistic objects.

Experiment Results and Open Issues

In this paper, we presented an experiment aiming at testing the expressive potentialities of the MILE as a standard for computational lexicons. The fundamental idea is that, by providing an efficient standard for the representation of notions at different level of linguistic description, we can obtain the key element for content interoperability among lexical resources. FrameNet and NOMLEX, two important, representative yet differently conceived lexicons, were chosen for the mapping experiment. The results of both experiments are promising, yet some reflections need to be made.

In the FrameNet to MILE experiment, we see that, even with some limits and approximations, all the FN basic notions can be, in some ways, represented using the MILE Lexical Classes. The possibility to work on a lexicon whose design follows a relational model allows an easier recognition of the lexical objects playing central roles at architectural level. MILE adheres to a relational model of the lexicon, where the *semantics* of each object is made explicit by the many relations the object has with the other objects available in the data structure. FrameNet is a lexicon of this type: the meaning of the Frame is not given by a description, a label or a code, but rather by the relations the Frame has with the Lexical Units, the Frame Elements etc.. When trying to map the FN structures on MILE, we have to verify if:

i) among the MLCs there is a valid correspondent for each FN lexical object,

ii) the internal coherence of FN is preserved when passing to MILE (i.e. if the reciprocal relations between the Frame, the Frame Element and the Lexical Unit are mirrored by the relations between the Predicate, the Argument and the Semantic Unit),

iii) there is no loss of information (and we saw that the danger of losing the important inheritance and embedding mechanisms among the Frames can be averted adding new specific modules to the MLCs).

The underlying models of NOMLEX and MILE are instead deeply different and the mapping is much more difficult. While the MILE pushes at the extreme the E-R model, NOMLEX adopts a type feature structure formalism to represent syntactic phenomena. The difference between the two is extremely evident when we observe how what in MILE belongs to distinct layers of representation (usually the semantic and syntactic layers) is represented in NOMLEX simply by juxtaposed labels within the same description code. Performing the mapping of a non-E-R lexicon onto MILE presents more difficulties and it is much more costly in terms of human intervention in the definition of the mapping conditions. It seems, however, an unavoidable price that we have to pay if we want to open the semantics and make the data structure more explicit, comparable with other lexical architectures and repositories. All in all, it can be a very useful enterprise when wanting to share and make interoperable the lexicon content in a distributed environment.

The two experiments are promising in showing how the highly expressive MILE can be used to represent both FN and NOMLEX. The modular, granular and flexible framework of the MILE model seems well suited for acting as a true interface between differently conceived lexical architectures, since it provides well recognizable, atomic, primitive notions that can be combined, nested and inherited to obtain more complex ones.

The described experiments are a first small-scale attempt to establish mapping conditions from some existing lexicons and the MILE. If we want MILE to become a really used standard, we should work intensively in the next future to provide mapping conditions between the most important lexicon models and architectures and MILE. It is obvious that this can be achieved only with the participation and help of the lexicon community, in order to benefit by the competence of each lexicon developer.

Acknowledgements

We want to dedicate this contribution to the memory of Antonio Zampolli, who has been the pioneer of standardization initiatives in Europe.

References

Baker C.F., Fillmore C.J., Lowe J.B. (1998). The Berkeley FrameNet Project. In Coling-ACL 1998: Proceedings of the Conference (pp. 86-90).

Calzolari, N., Bertagna, F., Lenci, A., Monachini, M. (2003). Standards and best Practice for Multilingual Computational Lexicons and MILE (Multilingual ISLE Lexical Entry).

Ide, N., Lenci, A., Calzolari N.: RDF Instantiation of MILE/ISLE Lexical Entries. Proceedings of the ACL 2003 Workshop on Linguistic Annotation: Getting the Model Right, 11 July 2003 Sapporo, Japan (2003).

Lenci A. (2003). Lexicon Design in the Age of the Semantic Web. Eurolan-2003 Summer School Tutorial, Bucharest, Romania.

Narayanan S., Fillmore C. J., Baker C.F, Petruck R.L. (2002). FrameNet Meets the Semantic Web: a DAML+OIL Frame Representation. In AAAI Workshop Proceedings.

Peters W., Cunningham H., McCauley C., Bontcheva K, Wilks Y. (1998). An Uniform Language Rosource Access and Distribution. In LREC-1998 Proceedings.

Reeves R., Macleod C., Meyers A. (1999). Manual of NOMLEX: The Regularized Version. Computer Science Department. New York University.

Vckovski A. (1999). Interoperability and Spatial Information Theory. Kluwer.