A new method for interoperability between lexical resources using MDA approach

Malek Lhioui, Kais Haddar, Laurent Romary

A new method for interoperability between lexical resources using MDA approach

Abstract. Lexical resources are increasingly multiplatform due to the diverse needs of linguists. Merging, comparing, finding correspondences and deducing differences between these lexical resources remain difficult tasks. Thus, interoperability between these resources is hard even impossible to achieve. In this context, we establish a new method based on MDA approach to resolve interoperability between lexical resources. The proposed method consists of building common structure (OWL-DL ontology) for involved resources. This common structure has the ability to communicate involved resources. Hence, we may create a complex grid between involved resources allowing transformation from one format to another. We experiment our new built method on an LMF lexicon.

Keywords: lexical resources, interoperability, MDA approach, OWL ontology.

Introduction

NLP Applications typically require interoperability because they generally need the same linguistic resources. Exchanging information between lexical resources, having different representation formalisms, is difficult. Thus, ancient resources may need to change over time. In addition to that, the transformation process from one format to another is not guaranteed all over the time. The built method allows constructing a pivot format for involved lexical resources with no prior restriction. A challenge that NLP communities confronts is the disappearing of several old representation formalisms after long periods of development. This method will easily protect several resources from disappearing. So, several formalisms will continue to persist. Thus projects, which require merging several formalisms in the same application, will prefer to use our method. In fact, it allows using the number of formalisms one wants. We use MDA (Model Driven Architecture) transformation approach because of its great interest in areas handling heterogeneous knowledge. In fact, if even a current version of a used standard (LMF for example) is not yet stable or a new version is born, MDA approach ensures enrichment and not destruction of the current version.

Building a new method for interoperability between lexical resources may face large problems. The first difficulty resides on how to choose an optimal strategy for interoperability between these resources: algebraic specifications, alignment ontology techniques, Meta modeling, etc. In addition to that, the choice of the ontology representation language (RDF(S), OWL-Lite and OWL-DL) is also a crucial dilemma. In addition to that, the construction of meta-models and models in MDA approach requires a big cognition of the involved lexical resources. Moreover, transformation rules have to be so definite.

The paper presents a new method strictly founded to resolve interoperability between lexical resources (LRs) using the well-known MDA Transformation approach. Indeed, we attempt to find compulsory techniques in order to establish a method for interoperability between lexical resources whatever their formats (LMF, TEI, HPSG…). The method consists on building a pivot format by making an automatic mapping process between involved lexical resources. The target building format plays the role of the pivot. In order to build this pivot format, we have to succeed to fulfill a set of steps. We construct OWL-DL ontology for lexical resources: construction of meta-model associated to lexical resource, construction of the ATL transformation and deduction of OWL-DL model. Thus, applying these steps, we build a new format for involving lexical resources.

The originality of this method is that there are no previous works aiming to make interoperability between lexical resources operable. Moreover, the use of MDA approach for resolving interoperability between lexical resources is in itself an innovation. Projects and NLP applications today must rely on interoperability; otherwise they are out of progress. In this context, an article named TAUS (TAU, 2011) declares that: "The lack of interoperability costs the translation industry a fortune". As a matter of fact, this fortune is compensated mostly in order to adjust data formats. In addition, our method is operable whatever the language.

In the following sections, we introduce a brief state of the art in order to give a global idea about existing works related to our topic. Then, we explain precisely our proposed method for resolving the interoperability issue between lexical resources. We apply, in the next section, the new proposed method to LMF lexicon. Finally, we conclude with a small discussion for the obtained results.

State of the art

The state of the art provides an important idea about existing works regardless of the language. There have been several works dealing with the use of MDA transformation approach for the processing of several applications. However, there is no use of the MDA approach for the processing of interoperability between lexical resources. Yet, this approach ensures interoperability according to the OMG "Portability and interoperability are built into the architecture"[1]. Since there are many related topics, we classify the state of the art into three main parts: lexical resources, MDA Transformation and interoperability issue. The first part gives an idea about existing lexical resources regardless the language. We give examples of lexical resources in several languages such as Arabic. In the second part, we talk about MDA as a great method for transformation models. The last part deals with interoperability issue, and since there are no serious attempts to resolve this notion in NLP area, we will discuss the bidirectional mapping from one format to another.

---

[1] http://www.omg.org/mda/specs.htm

2.1 Lexical resources

Lexical resources vary in accordance with the need of linguistics and this requirement varies with the NLP community development progress. This process makes the resources more complex and heterogeneous. In the literature, existing lexical resources are innumerable. We can concentrate on some of them. In the 1980ies SGML markup language was created as the first formalism representation of linguistic data. Early in the following century, several markup languages have been invented by the Text Encoding Initiative (TEI) (Wörner et al., 2006). After years and exactly in 2003, a new standard named LMF was born due to efforts provided by the community of NLP (Francopolou, 2013). Speech is one of the several areas of NLP domain. This area includes several representation formalisms as well as the other areas. For example, EXMARaLDA is one of these formalisms. It represents spoken interaction with an annotation graph (Bird & Lieberman, 1999). Other formalisms in this context was born such as ELAN, TASX, Praat and ANVIL. They are efficient for multimodal annotation. In the same context, there are formalisms that include several heterogeneous resource structures. The well known example for that is Tusnelda. It is inspired typically from the work of TEI (Wagner and Zeisler, 2004). There are other formalisms which take care of various linguistic levels (phonology, morphology, syntax, semantic, etc.) (Ide and Romary, 2001). Thus, from an historical point of view, there is a large number of heterogeneous resources which inducing the question of transformation. This notion is the subject of the next subsection.

2.2 MDA Transformation

MDA Transformation is an approach proposed by OMG (Poole, 2001) in 2001. It is increasingly used in several applications and projects whatever their kind. It consists on using different models phases. It allows interoperability between applications by connecting their models (Accord, 2002). It supported evaluation and decreased manually implementation of hundred of codes for a specific domain by separating conception from implementation (Miller and Mukerji, 2001). The implementation of MDA requires three main levels: MOF (Meta-Object Facility) defines the platform for implanting all models (OMG/MOF, 1997). PIM (Plateform Independant Model) which serves as a basis for the business part specification of an application, PSM (Plateform Specific Model) which participates in the specification model creation of the application after projection on a platform. The major advantage of this approach apart from time saving is preoccupations separation and the transformation process. This transformation allows mapping from PIMs to PSMs using modules described in specific languages such as ATL. ATL (Atlas Transformation Language) is a language providing rules allowing transformation from source to target models. Since this approach allows interoperability between applications, it leads us to think about making evident interoperability between models. Thus, we introduce in the following subsection interoperability notion in general.

2.3 Interoperability issue

Interoperability is the substitution, merging and sharing knowledge between different entities whatever their kind. NLP community replaces these terms by only one term which is communication. Thus, interoperability allows communication between involved entities. Interoperability is a general notion that can be projected to many domains. In this paper, we interested to interoperability between lexical resources. Lexical resources are more and more multiplatform, multi-providers… and these characteristics are increased by the time, so that, interoperability becomes hard even impractical to achieve between lexical resources. These last suffer from several interoperability issues. For example the definition of procedures to implement a set of services in NLP applications (machine translation, named entity recognition, part of speech tagging) shall be made through LMF by ISO, TEI by TEI Consortium and HPSG by linguistics… This leads to interoperability problems when experts have to collaborate. Thus, information technology professionals consider that interoperability is an important criterion as well as security and reliability in their applications.

From an historical point of view, there are no significant efforts resolving interoperability between lexical resources. Yet, there are several challenges consisting on mapping from one format to another. The first mapping attempt is done by (Wilcock, 2007) consisting on converting HPSG lexicons to an OWL ontology. In 2010, Loukil has expanded these processes by inventing a rule-based system opting to translate LMF syntactic lexicon into TDL within the LKB platform (Loukil et al., 2010). (Haddar et al, 2012) have developed a prototype for projection HPSG syntactic lexica towards LMF. In the same context, there is a mapping process already done by (Lhioui et al., 2015) aiming to convert LMF lexicons to ontologies described on OWL-DL language. Bidirectional processes are usually limited to involved formats. Whatever we desire to involve more than two formalisms, processing became hard and impossible to achieve even if we use several properties such as transitivity. For these reasons and in order to attenuate task complexity of mapping process, several organizations such as ISO give a quick solution but not efficient for interoperability using normalization. In fact, Lexical Markup Framework (LMF) is one of these solutions proposed by the ISO in 2003 (Francpolou, 2013). It involves several packages aiming to cover the maximum of the large domains: phonology, morphology, syntactic, semantic, pragmatic, etc. Other researchers have used another tool for resolving interoperability which is ontologies. A famous example of these works is the General Ontology for Linguistic Description (GOLD)[2]. GOLD is an OWL ontology having specific knowledge related to linguistic domain. The GOLD ontology contains the basis linguistic knowledge of any theoretical framework. According to (Farrar and Lewis, 2005), GOLD defines linguistic knowledge as axioms, for example "a verb is a part of speech", and uses at the same time language neutral, for example "parts of speech are subclasses of gold: GrammaticalUnit". The classes are presented in the protégée editor and then expressed as concepts in the GOLD ontology (Farrar and Langendoen, 2003). Thus, GOLD is an abstract model and representation formalisms such as

[2] Gold is accessible and free downloadable from (http://www.linguistics-ontology.org/)

HPSG are the instantiation of this abstract model. (Farrar and Lewis, 2005) consider these instantiations as sub-communities of practice noted Communities Of Practice Extension (COPEs). COPEs, sub-communities or sub-ontologies designed the same nomenclature and extend the overall GOLD ontology (Wilcock, 2007). The integration of these COPEs in the GOLD ontology is a hard process and necessitates different mechanisms of ontology alignment. In the next subsection, we try to give an idea for techniques of ontologies alignment.
All these notions will be strongly correlated to introduce our approach. In the following section, we define a new approach for interoperability between lexical resources using MDA Transformation.

Proposed method

The new build method is based on MDA Transformation approach. This approach is well-known and has proved its importance in guaranteeing reusability. This characteristic is crucial since it makes projects up to date. The proposed method is characterized by the ability to allow involved lexical resources to operate together. The new introduced method has as input a set of lexical resources. Lexical resources are composed of a set of lexicons such as LMF lexicon. It consists of three main steps. The first one is the achievement of the two independent models PIM (source and target) and the source PSM of each LR. The second is the achievement of the transformation module in ATL and finally, the generation of the specific model PSM (OWL-DL in our case). The output of the proposed method is a set of ontologies which can operate together using several algorithms or free tools of alignment. In fact, the use of ontologies as an output is the keystone of our method. Ontology structures allow merging, comparing, finding correspondences, finding correspondences and deducing differences between lexical resources due to the tools of ontology alignment. Fig. 1 describes the whole process of the proposed method.

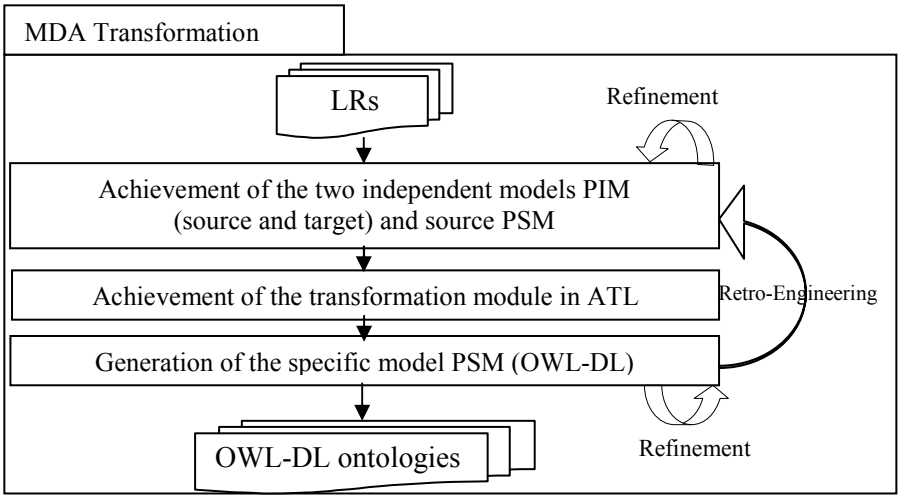

Fig. 1. Steps of the proposed method

The full schema of the proposed method will be explained carefully by examination of each step separately. In fact, MDA Transformation of the LRs to OWL ontologies is a crucial step in our method. The main idea of this step is to distinguish functional specifications from specifications of implementation related to a given platform in order to prepare structures able to operate together (in our case ontologies able to be aligned and then interacted). Thus, using MDA as an approach will make us able to elaborate independent specifications from the implementation in a specific platform using models. The first model to build is the PIM. The PIM is the model conceived to specify involved structures independently from any specific platform. This characteristic allows us abstracting functionalities of the involved lexical resource and to compare it to other resources. If the lexical resource is updated, the associated PIM will never be destroyed, but, it will be refined as many times as possible; this makes one of the most advantages of the MDA approach when resolving interoperability issue. Fig. 2 summarized the MDA Transformation in general:

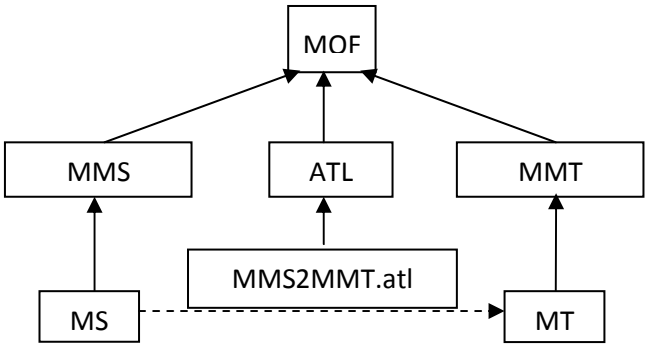

Fig. 2. : MDA Transformation of the LRs to OWL ontologies

Fig. 2 describes the ATL transformation in the MDA approach. MOF is the meta-meta model. MMS and MMT designate respectively the meta-model source and the meta-model target. MS and MT denote respectively model source and model target. MMS2MMT.atl includes the set of transformation rules. This method is composed of three sub-steps as we have mentioned below: Achievement of the two independent models PIM (source and target) of each LR, the achievement of the transformation module in ATL and the generation of the specific model PSM (OWL-DL in our case).

3.1 Achievement of the two independent models PIM (source and target) of each LR and source PSM

The achievement of the first independent model PIM of the source is concluded from the lexical resource. PIM is a model independent to any plateformes or technologies and describes the heart of the method. It is represented in UML (Unified Modeling Language) with OCL (Object Constraint Language) constraints if exist. This model defines all functionalities of the given lexical resource described in an abstract manner. The PIM model ensures analysis and design of applications. At this step, the de-

sign phase of the process involves the application of design pattern, partition into modules and sub-modules, etc. This PIM allows making available a structural and dynamic vision of the application without recourse to the technical design of the application. Therefore, a model (in our case the PIM) is essentially defined by a set of concepts and their relationships presented in a class diagram.

3.2 Achievement of the transformation module in ATL

The achievement of the transformation module in ATL ensures transition from one model (source) to another (target). Modules transformations based on meta-models constitute the main step of the MDA. In fact, a transformation model corresponds to a function taking a set of input models and finding a set of output models. The models, in and out, respect their meta-models previously built. The transformation uses the model manipulation API. In order to carry out the transformation between the two involved models, we define a set of transformation rules which are expressed in ATL language allowing the passage from the source PIM to the target. There are three different manners to model transformation in general: programming approach, template approach and modeling approach. The first one is based on object-oriented languages. It is to program a transformation model as well as a computer application. The second consists to define templates models and then replace them with their equivalent values in source models. The last one models transformation rules using MDA approach.

3.3 Generation of the specific model PSM (OWL-DL).

After achievement of the PIM model (source and target) and elaborating rules allowing the passage from the source PIM to the target, we project the source PIM to a specific model PSM (Platform Specific Model). In order to generate the target PSM, we execute the ATL rules, then, we obtain automatically the target PSM. In fact, PSM is closest to the final code. It is related to a particular platform.

Implantation: Transformation of LMF lexicon to OWL-DL ontology using MDA approach

In this section, we present the steps of the cited method applied to LMF lexicon: the two PIMs (source: LMF, target: OWL), the transformation rules and the two PSMs. Fig. 3 represents the source PIM of the core model of the following extract of LMF lexicon developed using Eclipse Galileo:

```
<?xml version="1.0" encoding="UTF-8"?>
<LexicalResource dtdVersion="16">
 <GlobalInformation>
  <feat att="languageCoding" val="ISO 639-3" />
  <feat att="scriptcoding" val="ISO 15 924" />
 </GlobalInformation>
 <Lexicon>
  <feat att="language" val="arabic" />
  <LexicalEntry morphologiquePatterns=" فعل
تام غير متعدّي">
   <feat att="partOfSpeech" val="verb" />
   <feat att="root" val="ن_ق_ل" />
   <feat att="scheme" val="تَفَعَّلَ" />
  </LexicalEntry>
 </Lexicon>
</LexicalResource>
```

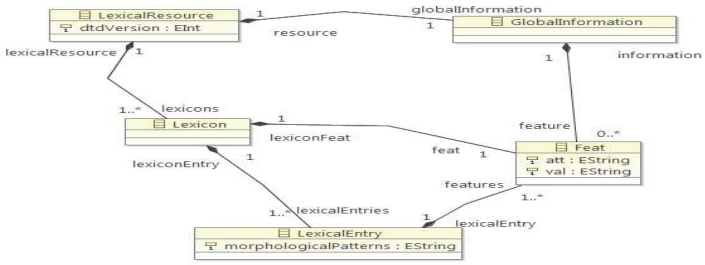

Fig. 3. The associated source PIM of the below extract of LMF lexicon

The build PIM can be refined as well as possible if the lexical resource (LMF lexicon) is updated. After building the source model, we have now obliged to build the target PIM of this lexical resource. Since we need to construct OWL-DL ontologies, we build a PIM for OWL-DL ontologies. Fig. 4 represents the target PIM for the previous lexical resource (LMF lexicon):

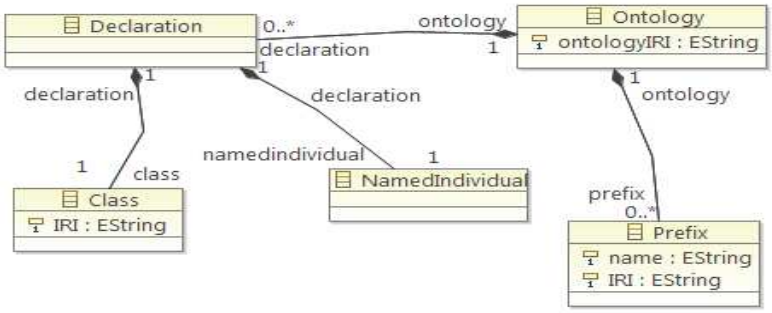

Fig. 4. The associated target PIM of the below extract of LMF lexicon

Fig. 4 defines the independent model of OWL ontologies which is related to the given lexical resource. It represents the class "Ontology" which presents the root, the prefixes which are used to abbreviate and minimize scripture of the namespaces in the entire ontology. Then, we define the set of transformation rules:

```
module LMF2OWL;
create OUT : OWL from IN : LMF;
-----------------------------Ontology-----------------------------------
rule LexicalResource2Ontology{
        from s:LMF!LexicalResource
        to
        t:OWL!Ontology(ontologyIRI <-),d:OWL!Prefix(name<-'rdf',
                        IRI<-'http://www.w3.org/1999/02/22-rdf-syntax-ns#'),
                        u:OWL!Prefix(name<-'rdfs',
                        IRI<-'http://www.w3.org/2000/01/rdf-schema#'),
                h:OWL!Declaration(),
g:OWL!Class(IRI <- '#LexicalResource',declaration <- h)}
rule GlobalInformation2DeclarationClass{
        from s:LMF!GlobalInformation
```

```
        to t:OWL!Declaration(),
        g:OWL!Class(IRI <- '#GlobalInformation', declaration <- t)
}
rule Lexicon2DeclarationClass{
        from o:LMF!Lexicon
        to p:OWL!Declaration(),
        i:OWL!Class(IRI <- '#Lexicon', declaration <- p)
}
rule LexicalEntry2DeclarationClass{
        from k:LMF!LexicalEntry
        to n:OWL!Declaration(),
        j:OWL!Class(IRI <- '#LexicalEntry', declaration <- n)
}
```

These transformation rules create an OWL PSM for the LMF lexicon of Fig. 2 which is an ontology described in OWL language. These rules are stored in an ATL file. Finally, fig. 4 represents this target PSM created automatically when executing the ATL file of the LMF Lexicon presented in Fig. 2:

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<xmi:XMI xmi:version="2.0" xmlns:xmi="http://www.omg.org/XMI" xmlns="owl"> <Ontology ontologyI-
RI="http://www.semanticweb.org/asus/ontologies/2016/2/"/> <Prefix name="rdf" IRI="http://www.w3.org/1999/02/22-
rdf-syntax-ns#"/> <Prefix name="rdfs" IRI="http://www.w3.org/2000/01/rdf-schema#"/>
 <Declaration><class IRI="#LexicalResource"/> </Declaration>
 <Declaration> <class IRI="#GlobalInformation"/></Declaration>
 <Declaration><class IRI="#Lexicon"/></Declaration>
 <Declaration><class IRI="#LexicalEntry"/></Declaration></xmi:XMI>
```

This output describes an ontology which is created automatically by a quick processing of the output of PSM. The processing consists of the removal of the “xmi” prefix since it is an automatic output of the ATL transformation.

Discussions

The new built method has proved its interest in handling heterogeneous resources. The evaluation process has been successfully led by fixing three criterions: sustainability of expertise, productivity gains and inclusion of execution platforms. The first criterion (sustainability of expertise) affects two characteristics. The first one supervises lifetime of the built models (PIM and PSM). The models must have a lifetime greater than the code. This is guaranteed by the unrestricted refinement of models. The second characteristic provides modeling languages supporting different levels of abstraction. This point is guaranteed by the fact that UML and OCL support abstraction. The second criterion concerns productivity gains. In fact, the automation operations of models guarantee the productivity gain. Moreover, the built method facilitates the creation of operations of production on the models. The last criterion concerns the taking into account of the execution platforms. This stage is explicit in the life cycle of applications. MDA approach guarantees this characteristic as platforms are related to models. These aspects make the method very robust. The other important aspect in the built method is that the transformation process is done automatically.

Conclusion

In this paper, we have proposed a new method for interoperability between interoperability using MDA approach. This new method allows merging, comparing, finding correspondences, finding correspondences and deducing differences between lexical resources. Then, we implement the method by projection on LMF. Our method is reusable and generic, and operable on all lexical resources whatever the language. Our method is generated automatically. In future works, we have to extend our method using the alignment of the building ontologies. In fact, if we combine MDA Transformation and ontology alignment, interoperability appears to be quite suitable. Therefore, combining these two approach MDA Transformation and ontology alignment for this study seems to have promising results.

References

1. CNAM, EDF R&D, ENST, ENST-Bretagne, France Telecom R&D, INRIA, LIFL et Softeam, Projet ACCORD (Assemblage de composants par contrats), Livrable 1.1-5, Date : Mai 2002.
2. Bird, S., and Liberman, M. 1999. Annotation graphs as a framework for multidimensional linguistic data analysis. Towards Standards and Tools for Discourse Tagging, Proceedings of the Workshop. Association for Computational Linguistics.
3. Scott Farrar and D. Terence Langendoen. 2003. A linguistic ontology for the semantic web. GLOT International, 7.3:97–100.
4. Scott Farrar and William D. Lewis. 2005. The GOLD Community of Practice: An infrastructure for linguistic data on the web. http://www.u.arizona.edu/~farrar/.
5. Francopoulo, G., 2013, *Lexical Markup Framework*, US, Great Britain and the United States: ISTE Ltd and John Wiley & Sons, Inc.
6. Haddar, K., Fehri, H., Romary, L., 2012, A prototype for projecting HPSG syntactic lexica towards LMF, JLCL.
7. Ide, N., and Romary, L. 2001. Standards for Language Resources. Proceedings of the IRCS Workshop on Linguistic Database, 141-149.
8. Lhioui M, Haddar K, Romary L. 2015. A prototype for projecting LMF lexica towards OWL.
9. Loukil, N., Ktari, R., Haddar, K., Benhamadou, A., 2010, A normalized syntactic lexicon for arabic verbs and its evaluation within the LKB platform, ACSE, Egypt.
10. Miller, J. and Mukerji, J. Model Driven Architecture (MDA) http://cgi.omg.org/docs/ormsc/01-07-01.pdf, July 2001. Architecture Board ORMSC.
11. OMG/MOF Meta Object Facility (MOF) Specification, OMG Document AD/97-08-14, Septembre 1997 (www.omg.org).
12. John D. Poole. "Model-Driven Architecture : Vision, Standards And Emerging Technologies". ECOOP 2001, Workshop on Metamodeling and Adaptive Object Models, April 2001.
13. (Tau, 2011) (TAU, 2011) TAUS, Report on a TAUS research about translation interoperability, 25 February, 2011.
14. (Wagner and Zeisler, 2004) Wagner, A., and Zeisler, B. 2004. A syntactically annotated corpus of Tibetan. In: Proc. of LREC, p. 1141–1144, Lisboa.
15. (Wilcock, 2007) Wilcock, G., 2007, *An OWL ontology for HPSG*, ACL, Finland.
16. (Wörner et al., 2006) Wörner, K., Witt, A., Rehm, G., and Dipper, S. eds. 2006. Modelling Linguistic Data Structures, Extreme Markup Languages, Montréal, Québec.