A Framework for Mapping Refinement Specification

F. Hamdi[1], **C. Reynaud**[1], **B. Safar**[1]

Abstract. Ontology alignment is an important task for information integration. There are many ongoing efforts to develop matching systems but these systems are not efficient for all application domains or ontologies. Very often the quality of the results can be improved by considering the specificities of the domain ontologies. In this paper, we propose an approach implemented in TaxoMap Framework, based on the alignment tool TaxoMap, which helps an expert to specify treatments exploiting mappings. The aim is to refine an alignment or to merge, restructure or enrich ontologies. We apply our approach to mapping refinement in the topographic field within the ANR (The French National Research Agency) project, GéOnto.

1 INTRODUCTION

The explosion of the number of data sources available in the web increases the need for techniques which allow the integration of these sources. The ontologies are essential elements in integration systems. However, they have to be aligned to achieve interoperability. There are many ongoing efforts to develop matching systems [4] but they are not efficient in all domains neither in all ontologies. They are very good in some cases, worse in others. The quality of their results is not always guaranteed and could often be improved if the alignment process took more into account the specificities of the aligned ontologies.

Considering these specific aspects can be done in different ways: (1) during the alignment process itself or (2) by refining the results generated by the alignment. In the first case, the adaptation of the handled ontologies is possible by the modification of the alignment process parameters or by the definition of a particular combination of alignment systems. In the second case, the refinement of mappings (the alignment results) extends and completes the alignment process, applied in the same way to all ontologies. This solution allows a finer adaptation of the alignment to the specificities of the handled ontologies. It allows also performing differentiated refinements according to the generated results. We retain this solution and extend it to consider, not only the improvement of the quality of an alignment, but also other tasks such as merging, restructuring or enriching ontologies. All these tasks are based on mappings and on the characteristics of the concerned ontologies. They must be made in interaction with the expert. Currently, there is no tool which allows to specify easily particular treatments to be applied to an alignment. The environment, called TaxoMap Framework, based on the alignment tool TaxoMap [14] [7], allows such specifications.

Our contributions, in this paper, focus on the conception of this environment, on the definition of a first set of primitives to support the specification of treatments exploiting mappings, and on the presentation of a use of the environment for mapping refinement in the topographic field.

The paper is organized as follows. In the next section, we present the context of this work, in particular the ontology alignment tool TaxoMap and the objectives aimed by the conception of TaxoMap Framework. In Section 3 we presents the approach adopted in our framework and in Section 4 we describe the use of this approach for mapping refinement applied in the topographic field within the ANR project GéOnto [5]. In Section 5 we present some related works. Finally we conclude and give some perspectives in Section 6.

2 CONTEXT

TaxoMap Framework is based on the alignment tool TaxoMap [14] [7]. We describe the tool in Section 2.1 and the objectives of the approach in Section 2.2.

2.1 TAXOMAP

TaxoMap has been designed to align *owl* ontologies $O = (C, H)$ (C is a set of concepts and H is a subsumption hierarchy). The alignment process is an oriented process which tries to connect the concepts of the source ontology O_S to the concepts of the target ontology O_T. The correspondences found are equivalence relations (*isEq*), subsumption relations (*isA*) or proximity relations (*isClose*).

To identify these correspondences, TaxoMap implements techniques which exploit the labels of the concepts and the subsumption links that connect the concepts in the hierarchy [6]. The morpho-syntactic analysis tool, *TreeTagger* [16], is used to classify the words of the labels of the concepts and to divide them into two classes, *full words* and *complementary words*, according to their category and their position in the labels. This repartition between *full* and *complementary words* is then used first, by a similarity measure that compares the tri-grams of the labels of the concepts [12] and gives more weight to the common *full words*, and second, by the alignment techniques. For example, one technique named "*isAStrictInclusion*" generates an "*isA*" mapping between X and Y if (1) the concept Y is the concept of O_T having the highest similarity value with the concept X of O_S, (2) one of the labels of Y is included in one of the labels of X, (3) all the words of the included label of Y are classified as *full words* by *TreeTagger*.

Mappings identified by TaxoMap are generated in the Alignment format [3] used as a standard in the OAEI campaign (Ontology Alignment Evaluation Initiative) [9]. We added to this standard the information about the names of the techniques

---

[1] LRI – University of Paris-Sud 11, CNRS & INRIA Saclay Île-de-France
Parc Orsay Université – 4 rue Jacques Monod – 91893 Orsay (France)
{Faycal.Hamdi, Chantal.Reynaud, Brigitte.Safar}@lri.fr

used in the alignment process. The aim is to facilitate the specification of treatments exploiting mappings which have been generated by these techniques.

2.2 TAXOMAP FRAMEWORK OBJECTIVES

Many ontology alignment tools have been developed in these last years but as shown in the results of the OAEI campaigns [1], no tool reaches 100% of precision and recall, even if the results obtained by some of them are very good. We observed TaxoMap through its results in this competition in the two last years [7][6] and also through our participation in the ANR project GéOnto [5]. The aim of this project is the construction of a topographic ontology and its enrichment with elements coming from other geographic ontologies using alignment techniques. In this setting, tests performed on taxonomies provided by the COGIT-IGN (project partner) have shown that TaxoMap gives very good results (precision 92.3%) but these results could still be improved.

A more closely study showed that the improvements desired by the experts are often specific to the aligned ontologies. Our aim was not that TaxoMap becomes a tool dedicated to the alignment of such topographical taxonomies (and thus the quality of results would not be guaranteed when TaxoMap would be used to align other ontologies coming from other domains). Therefore, we proposed to the experts of the project an environment allowing them to specify and perform different treatments. This environment will be used to improve the quality of an alignment provided by TaxoMap, but also for any other treatment based on the results of an alignment between ontologies, such as merging, restructuring or enriching ontologies.

The expert will be able to select all primitives through an appropriate GUI. Note that the approach is based on the use of TaxoMap as alignment tool, but it could be based on another tool if the primitives associated with the tool have been defined.

3 TAXOMAP FRAMEWORK APPROACH

The approach TaxoMap Framework has been designed to meet the objectives described in section 2.2. We describe the approach and a diagram representing the architecture of this environment respectively in Section 3.1 and 3.2. This environment allows the specification of treatments from predefined primitives. These are presented in Section 3.3.

3.1 PRESENTATION OF THE APPROACH

An important feature of the approach is to allow a declarative specification of treatments based on particular alignment results and concerning particular ontologies, using a set of generic and predefined primitives.

Treatments which can be specified depend on the characteristics of the concerned ontologies and the aimed task (mapping refinement, ontology merging, restructuring, enriching). These treatments are thus associated to independent specifications modules, one for each task, having each their own set of primitives of specification. The approach is extensible and a priori applicable to any treatment based on the alignment results.

This approach should help to refine the alignment results produced by TaxoMap. It must be possible, for example, to specify that the subsumption mapping "*isA*" generated between "Way and coastal path" and "Path", as shown in Figure 1 must be replaced by a mapping of the same type but between "Way and coastal path" and "Way". Indeed, "Path" is defined in O_T as a kind of "Way" and the term "Way" itself is used in the label "Way and coastal path". The expert would thus prefer to establish a mapping directly between "Way and coastal path" and "Way".
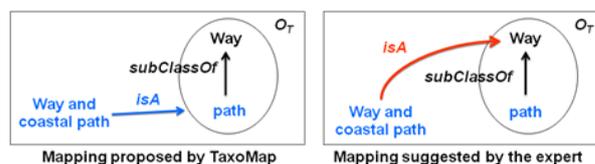

Figure1. Example of treatment to specify

The specification of treatments should be as generic as possible. Thus, the treatment shown in Figure1 should not refer directly to the concepts denoted by "Way", "Path" and "Way and coastal path". To help the expert to clarify the conditions for applying the treatment he wishes to implement, we propose a set of predefined generic primitives. These primitive allow a representation of various conditions which can be tested on the mapping concepts identified by TaxoMap. By analyzing alignment results and by leaning on proposed primitives, the expert will be able to identify a "group" of mappings requiring the same refinement and to specify the appropriate treatment to apply to each identified group. The specification will be so declared in a generic way then instantiated on the alignment results and the concerned ontologies to perform the corresponding treatments.

The approach should also allow other treatments such as the restructuration of an ontology O' built from O_S and O_T, and the alignments generated by TaxoMap between these two ontologies. Thus, it should help to choose, for example, what mappings "*isA*" must be transformed into "*subClassOf*" relations and what dependant concepts have to be imported in O'.

3.2 ARCHITECTURE OF TAXOMAP FRAMEWORK

Figure 2 presents the specification environment implemented in TaxoMap Framework. This environment has three modules: a "controller", a "knowledge" and a "treatment" module.

The "knowledge" module includes all pieces of knowledge that may be used in the specification process. It includes the ontologies aligned by TaxoMap, O_S and O_T, and the generated alignment (stored in a mappings database). According to the treatment performed, we may also have other ontologies corresponding to the result of a merging or a restructuring process.

The "treatment" module includes the alignment tool TaxoMap and all the modules associated to the different tasks to perform. TaxoMap executes sequentially 9 techniques, which may or may not be chosen during a particular session. In addition, the execution order of these techniques is customizable. The modules associated to the tasks allow to specify particular treatments that an expert wishes to implement

on particular ontologies or alignments, and also to execute these treatments. Additional modules can easily be added if they combine the appropriate primitives (which may be taken from the primitives proposed in the other modules).

The "controller" module manages all the treatments that can be performed using this environment, i.e. the specification of treatments and their execution, the access to relevant data and the storage of the obtained results.
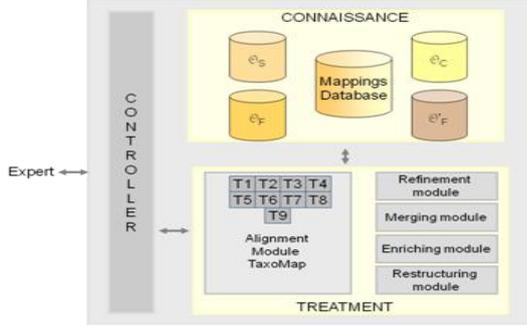

Figure 2. Architecture of TaxoMap Framework

3.3 PRIMITIVES OF TAXOMAP FRAMEWORK

The vocabulary used in the mapping refinement specification language contains:

- a set of predicate constants: {*isEquivalence*, *isAStrictInclusion*, *isCloseStrictInclusion*, …}
- a set of individual constants: {a, b, c,...}
- a set of variables: {x, y, z, …}
- a set of procedures: {*Add_Mapping, Delete_Mapping, …*}
- a set of logical symbols: {∃, ∧, ¬}

This vocabulary corresponds to the primitives that can be selected by the expert to specify treatments. The sets of primitive differ according to the aimed task (mapping refinement, ontology merging, restructuring, enriching).

The specification of a treatment has two parts: a "condition" part which must be satisfied to make the execution of the treatment possible, and an "action" part which expresses the process to achieve when the "condition" is satisfied.

The **condition part** is expressed through a set of primitives which corresponds mainly to the names of predicates. These predicates are designed to test (1) the technique used to identify the considered mapping, (2) the structural constraints on mapped elements, for example, the fact that they are related by a subsumption relation to concepts verifying or not some properties, or (3) the terminological constraints, for example, the fact that the labels of a concept are included in the labels of other concepts. These conditions are represented using three kinds of predicates:

The predicates relating to the type of techniques applied in the identification of a mapping by TaxoMap. By testing the existence in the mappings database of a particular relation generated by a given technique, these predicates test implicitly the conditions for the application of this technique. The expert neither needs to know precisely the techniques used nor to re-specify them. For example the primitive "*isAStrictInclusion(X, Y)*" tests the existence of a mapping "*isA*" generated between

two concepts X and Y using the technique t_2. It validates implicitly at the same time the conditions for the application of t_2, i.e. (1) one of the labels of Y is included in one of the labels of X, (2) all the words of the labels of Y are classified as *full words* by *TreeTagger*, and (3) the concept Y is the concept of O_T having the highest similarity value with the concept X.

TaxoMap including several alignment techniques and thus, several predicates will be defined. More formally, let:

R_M = {*isEq*, *isA*, *isClose*}, the set of correspondence relations used by TaxoMap,

T = {t_1, t_2, t_3, t_4, t_5, t_6, t_7, t_8, t_9}, the set of techniques.

T_M, the table storing generated mappings in the form of 4-tuple (x, y, r, t) where x ∈ C_S, y ∈ C_T, r ∈ R_M, t ∈ T.

The pairs of variables *(X, Y)* which can instantiate these primitives will take their values in the set {(x, y) | (x, y, r, t) ∈ T_M}.

The primitives corresponding to predicate symbols necessary to illustrate the task of refinement in this paper are *isEquivalence*, *isAStrictInclusion* and *isCloseStrictInclusion*. They are defined as follows:

- *isEquivalence (X,Y)* is true iff ∃ (X, Y, isEq, t_1) ∈ T_M
- *isAStrictInclusion (X,Y)* is true iff ∃ (X,Y, isA,t_2) ∈ T_M
- *isCloseStrictInclusion(X,Y)* is true iff ∃ (X,Y,isClose, t_3) ∈ T_M

These primitive will be presented to the expert via an interface with comments clarifying their conditions of validity as well as examples and counter-examples of use.

The predicates expressing structural relations between concepts X and Y of the same ontology $O = (C,H)$. Note that the instances of variables in these predicates will be constrained, either directly because they instantiate the previous predicates, i.e. concerning the type of the applied techniques, or indirectly by having to be in relation with other instances.

- *isSubClassOf(X,Y,O)* is true ⇔ subClassOf(X,Y) ∈ H
- *isSuperClassOf(X,Y,O)* is true ⇔ subClassOf(Y,X) ∈ H

The predicates expressing terminological relations between the labels of the concepts:

- *strictInclusionLabel*(X,Y) is defined as follows:

    For each label L_1 of X

      For each label L_2 of Y

        If $L_1 \subseteq FullWords(L_2, L_1)$ then return true

    End

where X and Y ∈ $C_S \cup C_C$ and *FullWords*(L_2, L_1) is a function, which calculates all the terms of L_2 considered as full words in its comparison with L_1.

- *conceptsDifferent*(X,Y) is true ⇔ $ID(X) \neq ID(Y)$ with $ID(X)$ is the identifier of the concept X.

- *inclusionInLabel*(X,Y) is true iff ∃ a label L_1 of Y / X ⊂ L_1, where X ∈ {"and", "or"} and Y ∈ $C_S \cup C_T$.

The **action part** describes the procedures to be performed. We identified an initial set of actions. They are represented using the following three procedures:

- *Add_Mapping*(X, Y, R) has the effect of adding a tuple to the table T_M which becomes $T_M \cup$ {(X, Y, R, t)} where R and t are fixed in the treatment condition, by instantiating the predicate identifying the technical question.

- *Delete_Mapping*(X, Y, _) has the effect of removing a tuple from the table T_M which becomes $T_M -$ {(X, Y, _ , _)}

- *Add_Relation*(X, Y, R) corresponds to the addition of a relation R between X and Y with R ∈ R_M ∪ {subClassOf}.

4 APPLICATION TO THE MAPPING REFINEMENT

The mapping refinement module is the first module of TaxoMap Framework, realized within the ANR project, GéOnto [5]. We describe the application setting and then we present the specifications of the treatment of mapping refinement required by the experts of the COGIT-IGN (project partner).

4.1 APPLICATION DOMAIN

One of the goals of the GéOnto project is to build an ontology of topographic concepts, as complete as possible, by enriching an initial taxonomy of terms. Topo-Cogit is an ontology already achieved by the COGIT. The enrichment is carried out by the alignment of this ontology with other ones of the same domain. Thus, within the project, other partners are developing an ontology based on the topographic specifications of the IGN databases and on travel books using NLP techniques [8, 10]. The enrichment process must be automated and reusable in the future on other domain ontologies. As this process is based on alignment results, these results must be as accurate as possible to minimize the contributions of the experts.

The first tests have been performed using a second ontology, Carto-Cogit, built manually from the specification of the IGN database. We aligned the 495 concepts of the source ontology Carto-Cogit with the 600 concepts of the target ontology Topo-Cogit in order to enrich Topo-Cogit. 326 mappings have been identified by TaxoMap and presented to the experts. 25 mappings (precision 92.33%) have been deemed as invalid. For other mappings, the expert proposed alternative mappings. The treatments of mapping refinement proposed below intended to obtain these alternative mappings.

4.2 THE MAPPING REFINEMENT SPECIFICATION

We present in this section two expected changes and in each change the specification of treatments such as they can be expressed in TaxoMap Framework.

- **Case 1:** The first improvement is presented in Section 3.1 (see Figure 1). Generally, it concerns mappings connecting by a subsumption relation "*isA*" a concept c_S of the source ontology O_S to a concept c_{TMax} of the target ontology O_T, such as one of the labels of c_{TMax} is included in the labels of c_S. If one of labels of the concept a that subsumes c_{TMax} in C_T is also included in the label c_S, (see Figure 3), the expert prefers to attach c_S with a, the most general concept of O_T.
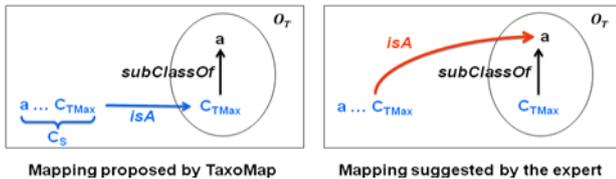

Figure 3. Illustration of the treatment1

The specification of the correspondent treatment is:
Conditions of application:
$\exists$X $\exists$Y *isAStrictInclusion* (X,Y)
$\wedge$ $\exists$Z *isSubClassOf* (Y,Z,O_T)
$\wedge$ *strictInclusionLabel* (Z,X))
Actions: *Delete_Mapping* (X, Y, _)
$\wedge$ *Add_Mapping* (X,Z, isA)

The application of this treatment on the example presented in Figure 1 allows first to select from the mappings database the mapping (*ID*("Way and coastal path"), *ID*("Path"), isA, t_2) satisfying *isAStrictInclusion*(*ID*("Way and coastal path"), *ID*("Path")). The variables X and Y are instantiated respectively by X/*ID*("Way and coastal path") and Y/ *ID*("Path"). The use of the structural predicate *isSubClassOf*(*ID*("Path"), Z, O_T) allows the instantiation of the variable Z, Z/ID("Way") and the verification of the terminological predicate *strictInclusionLabel*(*ID*("Way"), *ID*("Way and coastal path")). The mapping (ID ("Way and coastal path"), ID ("Path"), isA, t_2) is removed from the database and replaced by the mapping (ID ("Way and coastal path"), ID ("Way"), isA, t_2).

- **Case 2:** This case concerns the mapping connecting by a relation of proximity "*isClose*" a concepts c_S of the source ontology O_S to a concept c_{TMax} of the target ontology O_T, such as one of the labels of c_S is included in c_{TMax} labels. If another label of a concept in O_T contains also c_S labels and if this concept has the same father p in O_T that c_{TMax}, the expert prefers to connect to c_S to p. An illustration is given in Figure 4.

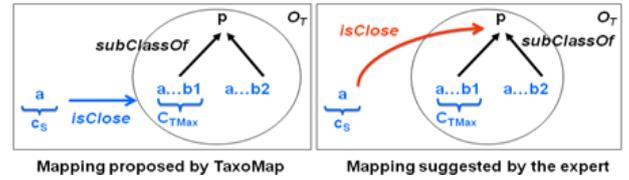The specification of the treatment associated to the case 2 is the following:

Figure 4. Illustration of the treatment2

Conditions of application:
$\exists$X $\exists$Y (*isCloseStrictInclusion* (X,Y)
$\wedge$ $\exists$Z $\exists$P (*isSuperClassOf* (P,Y,O_T)
$\wedge$ *isSuperClassOf* (P,Z,O_T)
$\wedge$ *conceptsDifferent*(Y,P)
$\wedge$ *strictInclusionLabel* (X,Z)))
Actions: *Delete_Mapping* (X,Y,_)
$\wedge$ *Add_Mapping* (X,P, isClose)

5 RELATED WORKS

The closest work we know is the COMA++ system [2]. It aims to build powerful alignment tools by the combination of existing matchers then to refine the obtained alignment results considered as preliminary. The refinement process is here totally automatic. The COMA++ alignment process is re-applied on groups of elements whose proximity has been established by a first treatment applied to ontologies. Other works deal with alignment refinement or transformation which are close

activities but not similar. In [15], correspondence patterns are used to assist the design of precise and complex ontology alignments when parts of both ontologies represent the same conceptualizations but modeled in two different ways. This approach can be seen as a way to refine one to one correspondences which can then be used to transform an ontology into another. Other works propose services to transform alignments. The Alignment API [3] generates transformations which are implementations for rendering the alignments, but the alignments are not modified.

Regarding our environment, another related work is PROMPT-Suite integrating the ontology merging tool IPROMPT [13], the alignment tool Anchor-PROMPT, versioning, comparison, translation functionalities. All these tools are interactive and semi-automatic. For example, in the fusion process the system makes suggestions. The expert can hold one of them or specify an operation to perform. The system executes then the operation, calculates the resulting changes, makes other suggestions and detects any inconsistencies.

All systems combining several alignment systems are very modular. The possibility of defining the strategy of combination makes them adaptable to a new field of application. This modularity and adaptability are strong points which also characterize our approach. However, TaxoMap Framework differs from existing tools, such as COMA++ or PROMPT-Suite, by considering that performance of an alignment tool implementing general alignment algorithms is necessarily limited. Some improvements can be obtained only after taking into account the particularities of the aligned ontologies, which involves improvements depending on the ontologies. The definition of such improvements needs to be familiar with the aligned ontologies. This process cannot thus be automatic. Only an expert of the domain is able to make it. As in PROMPT-Suite, we offer an interactive environment to help the expert to carry out this task, but we do it differently. We allow him to define particular generic treatments. In PROMPT-Suite, this is not possible. The treatments are all pre-defined.

6 CONCLUSION AND FUTURE WORK

TaxoMap Framework is an environment for the specification of treatments based on the alignment results generated by TaxoMap. We presented the implemented approach in this system, its architecture, then a first set of pre-defined primitives allowing an expert of the domain to specify easily the treatments to apply to an alignment. We presented the module for assistance to the mapping refinement that we conceived basing on the results of experiments realized within the ANR project, GéOnto.

The conception of TaxoMap Framework is adapted to the specification of other treatments such as merging, restructuring and enriching of ontologies based on alignment results. These modules are being implemented as well as the graphical interface that will allow to experts to easily select the appropriate primitives.

REFERENCES

[1] C. Caraciolo, J. Euzenat, L. Hollink, R. Ichise, A. Isaac, V. Malaisé, C. Meilike, J. Pane, P. Shvaiko, H. Stuckenschmidt, O. Svab, V. Svatek, *Results of the ontology alignment initiative 2008*, in P. Shvaiko, J. Euzenat, F. Giunchiglia, H. Stuckenschmidt (Eds), Proceedings 3rd ISWC workshop on Ontology Matching, Karlsruhe (DE), pp 73-119, 2008.

[2] H.-H. Do, E. Rahm, *Matching large schemas: Approaches and Evaluation*, Information Systems 32 (2007), 857-885.

[3] J. Euzenat, An API for ontology alignment. *An API for ontology alignment*, ISWC, Hiroshima (JP), LNCS 3298:698-712, 2004.

[4] J. Euzenat, P. Shvaiko, *Ontology Matching*, Springer-Verlag, Heidelberg (DE), 2007.

[5] GéOnto : http://geonto.lri.fr/

[6] F. Hamdi, B. Safar, N. Niraula, C. Reynaud, *TaxoMap in the OAEI 2009 alignment contest*. – Ontology Matching, 2009.

[7] F. Hamdi, H. Zargayouna, B. Safar, C. Reynaud, *TaxoMap in the OAEI 2008 alignment contest*. – Ontology Matching 2008.

[8] M. Kamel, N. Aussenac-Gilles, *Ontology Learning by Analysing XML Document Structure and Content*, KEOD, Madère, Portugal, 2009.

[9] http://www.ontologymatching.org/

[10] E. Kergosien, M. Kamel, C. Sallaberry, M.-N. Bessagnet, N. Aussenac, M. Gaio, *Construction automatique d'ontologie et enrichissement à partir de ressources externes*, JFO, Poitiers, 2009.

[12] D. Lin, *An Information-Theoretic definition of Similarity*, ICML, Madison, pp. 296-304, 1998.

[13] N. F. Noy, M.A. Musen, *The PROMPT Suite: Interactive Tools For Ontology Merging And Mapping*, IJHCS, 59(6), pp. 983-1024, 2003.

[14] C. Reynaud, B. Safar, *Techniques structurelles d'alignement pour portails Web*, Revue RNTI W-3, Fouille du Web, Cépaduès, 2007.

[15] F. Scharffe, J. Euzenat, D. Fensel, *Towards Design Patterns for Ontology Alignment*, SAC, pp. 2321-2325, 2008.

[16] H. Schmid, *Probabilistic Part-of-Speech Tagging Using Decision Trees*, Int. Conf. on New Methods in Language Processing, 1994.