Al –Khalil: The Arabic Linguistic Ontology Project

Hassina Aliane[1], Zaia Alimazighi[2], Mazari Ahmed Cherif[2]
[1]Semantic web and Arabic Language Team, Research Center on Scientific and technical Information, Algiers.
[2]Computer Science Department, University of Science and Technology, USTHB, Algiers.
[1]haliane@mail.cerist.dz, [2]alimazighi@wissal.dz

Abstract

We present in this paper our project to building an ontology centered infrastructure for Arabic resources and applications. The core of this infrastructure is a linguistic ontology that is founded on Arabic Traditional Grammar. The methodology we have chosen consists in reusing an existing ontology, namely the Gold linguistic ontology. We discuss the development of the ontology and present our vision for the whole project which aims at using this ontology for creating tools and resources for both linguists and NLP researchers.

1. Introduction

Despite Arabic is the language of hundred millions of people over the world, little has been done in terms of computerized linguistic resources, tools or applications. The project we present in this paper is about building an ontology for Arabic linguistics. Our aim is to contribute filling this gap in two ways:

1- Giving some more visibility for Arabic linguistics and give opportunity to people to know about Arabic linguistics and especially Arabic Traditional Grammar.

2- The ontology will serve as an infrastructure to build applications for linguistics or NLP researchers and at the same time it would constitute a mean to develop some useful tools and programs for NLP and IR communities.

To achieve the goal of our project we have opted to work through the following steps:

1. Delimit the intended content and users for our ontology,
2. Choose a motivated approach for the ontology development,
3. Fix the necessary development tools,
4. Develop the interfaces,
5. put the ontology online,
6. Feedback and evaluation,
7. Develop some motivated applications that use the ontology.

The work that has been done up to now concerns the first four steps and will be discussed in section 3.

2. Related Work

There is an increasing interest in linguistic ontologies (e.g. WordNet) for a variety of content-based tasks, including conceptual indexing, word sense disambiguation and cross-language information retrieval. A relevant contribution in this direction is represented by linguistic ontologies with domain specific coverage, which are a crucial topic for the development of concrete applications (Magnini & Speranza, 2002). In the recent years the increasing interest in ontologies for many natural language applications has led to the creation of ontologies for different purposes and with different features systems.

2.1 Ontologies for Linguistics

In computer Science, an ontology is a shared and common understanding of some domain that can be communicated across people and application systems(Fensel, 2000) or enabling knowledge sharing, it is a specification of a conceptualization(Gruber, 2000).

The rise of linguistic ontologies is a result of two concurrent situations. Indeed, it is undeniable that our age is the age of information. There are huge amounts of information everywhere, especially for our concern on the web. This information needs to be structured and represented in a way that facilitates its exploitation by users later: this is the topic of ontologies. In the same time, language is the way to vehicle information and knowledge and furthermore, the need for linguistic data is crucial in all research fields that are concerned by the organisation of information and its retrieval for the end user like information indexing, extraction, retrieval and NLP; linguistics and formal (computational linguistics) are also concerned by the availability of (organized, structured and easily retrievable) linguistic data.

As Farrar noticed in (Farrar & Langendoen, 2003), the World Wide Web has the potential to become a primary source for storing and accessing linguistic data. Today, ontologies are not only central to the vison of the semantic web, but they‘re pervasively used in numerous different domains. Consequently, using an ontology to organize linguistic data is not surprising. As examples for such ontologies, we’ll cite:

- **GOLD** is the first ontology being designed for linguistic description on the semantic web and it is based on the principles of knowledge engineering. Domain knowledge is made maximally explicit in a knowledge representation language (Farrar & Langendoen, 2003).
- **DOLCE** is a descriptive ontology for linguistic and cognitive engineering at applied ontology laboratory, Italy.
- **GUM** The Generalized Upper Model is a general task and domain independent `linguistically motivated ontology' that provides semantics for natural language expressions.
- **Mikrokosmos** has been initially developed in the framework of a machine translation project at the computing research laboratory, New Mexico University.

2.2 Lexical Resources for Arabic

The absence of free usable lexical and syntactic resources and tools for Arabic makes it a "pi- language" (poorly informatised) (Daoud &al, 2009). This constitutes a real difficulty in the process of transferring technology into Arabic. As existing, perhaps most known resources , we cite the DIINAR lexical database (Dichy &al, 2002) and the ARABIC WORDNET (Black& al, 2006).

DIINAR.1 (*DIctionnaire INformatisé de l'Arabe*), is a comprehensive Arabic Language database operating at word-form level (morphological analysis or generation).

Arabic Word Net is a lexical resource in standard Arabic based on the design and contents of the universally accepted Princeton Word NET. We think the semantic web is a real opportunity to boost research and development in Arabic NLP and IR in general.

3. ALKHALIL : an OWL Ontology for Arabic Linguistics

Al-khalil is an OWL ontology under development. We have baptized our project Al-Khalil in the sake of the famous grammarian AL-Khalil Ibn Ahmad Alfarahidi because we consider in some sense he was the first to have built an ontology for the Arabic language trough his "kitab alayn" which means the book of the letter ع. the name came from the fact that the dictionary follows a phonetic order starting from the pharyngeal sound ع

We have chosen to build our ontology on an existing linguistic ontology namely the Gold ontology. The development of our ontology is two steps:

- Bootstrapping manually the ontology by choosing the linguistic concepts from Arabic linguistics and relating them to the concepts in GOLD.
- Using an automatic extraction algorithm to extract new concepts from linguistic texts to enrich the ontology. The algorithm is based on the repeated segments calculus method. The general architecture of the system is depicted in figure 1.

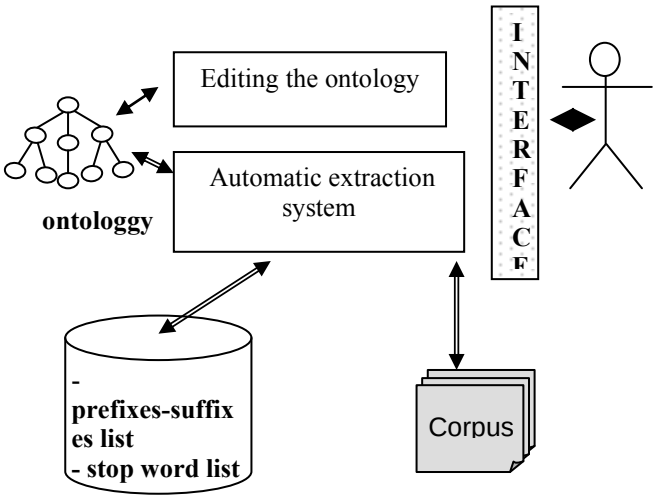

Figure 1 **:** Architecture of the prototype

3.1 Bootstrapping the Ontology

Our ontology is intended to provide a reference for description of Arabic linguistics. We focus on Arabic Traditional Grammar but we are also concerned by all new concepts from linguistics frameworks applied to Arabic. GOLD, General Ontology for Linguistic Description (Farrar & Langendoen, 2003; Farrar, 2003) is the first ontology being designed specifically for linguistic description on the Semantic Web. GOLD is an OWL ontology that aims to capture"the general knowledge of the field that is usually possessed by a well trained linguist. This includes knowledge that potentially forms the basis of any theoretical framework. In particular, Gold captures the fundamentals of descriptive linguistics. Exemples of such knowledge are "a verb is a part of speech", gender can be semantically grounded", "linguistic expressions realize morphemes" (Magnini & al, 2002; Farrar, 2003). We have manually extracted from a chosen corpus the most prominent concepts of Arabic Traditional Grammar (ATG); the chosen corpus concerns the Néo-khalilian description of ATG. Thus we have related those concepts to the concepts of GOLD. GOLD organizes linguistically related concepts into four major domains: expressions, grammar, data constructs, and meta-concepts. Figure2 shows the Gold upper taxonomy. The Protégé-OWL ontology editor has been used to both visual construction and visual editing of the ontology.

Entity
- *Abstract*
  - *FeatureValue*
  - *GrammaticalUnit*
  - *LinguisticDataStructure*
  - *LinguisticFeature*
  - *LinguisticSign*
  - *PhonologicalUnit*
  - *SemanticUnit*
- *Object*
  - *SymbolicString*

Figure2 **:** GOLD upper taxonomy

We have related our hierarchy of concepts types for Arabic linguistics to the Gold concept type *Entity* (root) and we then relate our concepts to either *abstract* or *object*. The concepts manually selected are:

Qiyās, aṣl, far', mawḍi', miṯāl, ta'āqub, binā', waṣl, ḥarf, kalima, Racine (gidr), linear Scheme, lafda, infiṣāl-'ibtidā', tamakkun ,taṣarruf, lafda ismiya, lafda fi'lya, scheme of lafda ismiya, fi'l madi, fi'l mudari', fi'l'amr, scheme of fi'l al-māḍī, scheme fi'l al-muḍāri', scheme fi'l al-'amr, tectonies, 'āmil, ma'mūl, Mubtada', Fā'il, Maf'ul, Mabni t mabni'alayh, 'ittisā', fi'il

muta'addi, fi'il gayr muta'addi, fi'il ya-ta'addā ilā maf'ūlayn, maf'ūl, darf, Determinant, Al-ḥāl, At-tamyīz, Al-maf'ūl ma'ahū, Al-mustaṯnā, Al-maf'ūl lahū, Al-maf'ūl al-mutlaq, AL-darf , Al-maf'ūl fīhi, Al-badal, As-ṣifa, 'itāla., Taṯniya (takrīr.),

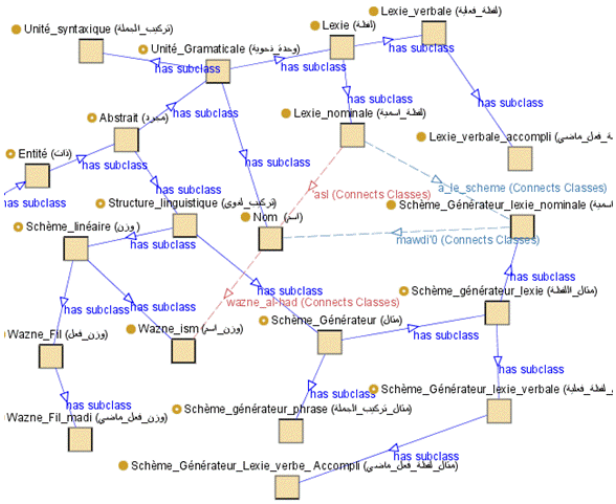

Figure3 an is-a hierarchy

Nevertheless, an ontology is not just a taxonomy of a domain concepts but an ontological theory specifies the entities of interest in a given domain and those entities include not only classes and their instances but the relations that hold among those instances. We have chosen conceptual graphs to formalize our conceptualization of AGT concepts and the relations holding between them. Figure3 depicts a portion of the ontology where concepts are related by the" is- a" relation and figure 4 shows an example of non hierarchical relations between concepts in the ontology.

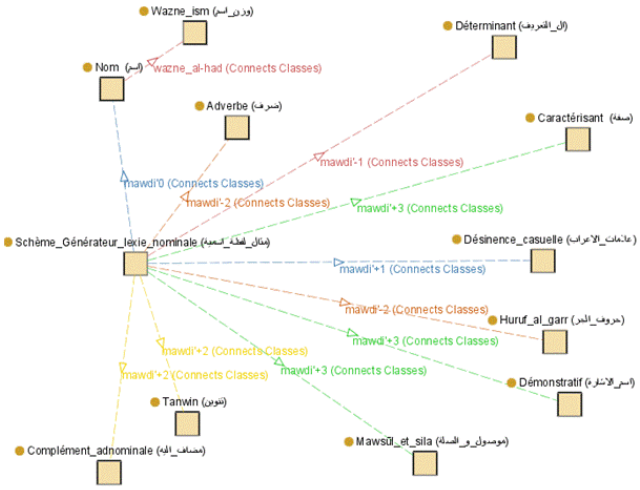

Figure4 non hierarchical relations

3.2 Enriching the ontology

Once the core of the ontology manually created, we have implemented an information extraction algorithm to enrich it. In the absence of free NLP or IE tools for Arabic, we have opted for a statistical approach, namely the method of repeated segments calculation combined with some prior processing of the texts that comprise: segmentation, light stemming, stop words elimination(Aliane, 2006). Python language is used for implementing the extraction system. The candidate concepts and relations are proposed to the expert before inserting them in the ontology. Browsing and editing interfaces are provided. We are about testing and validating the first version of the ontology before putting it online.

4. Conclusion and future work

In constructing the first prototype of our ontology we have focused on the concepts of Arabic Traditional Grammar that don't appear in other linguistic theories such as mital, qiyas, lafda, … and other concepts pertaining to the Neo-khalilean framework which is a modern interpretation of Arabic Traditional Grammar . We make this difference because in the future we aim at:

- Building a community of practice (cope) (Wilcok, 2007) for the Neo-khalilean school of Arabic traditional grammar.

A cope is a subontology that inherits from and extends the overall gold ontology. Subontology classes are distinguished from each other by different name space prefixes, for example gold:noun, hpsg: noun, ATG: noun, ism.

- Extending the content of the ontology. Indeed, as the ontology is intended to be a reference for linguists and NLP researchers in different areas of the field, we aim the ontology to contain exhaustive knowledge about standard Arabic, formal and NLP works on Arabic, dialects and linguistic phenomena relating to Arabic,
- Linking our ontology to projects on Arabic corpus for instance the Algerian Arabic treasury project an building significant applications that use the ontology. The overall project looks like:

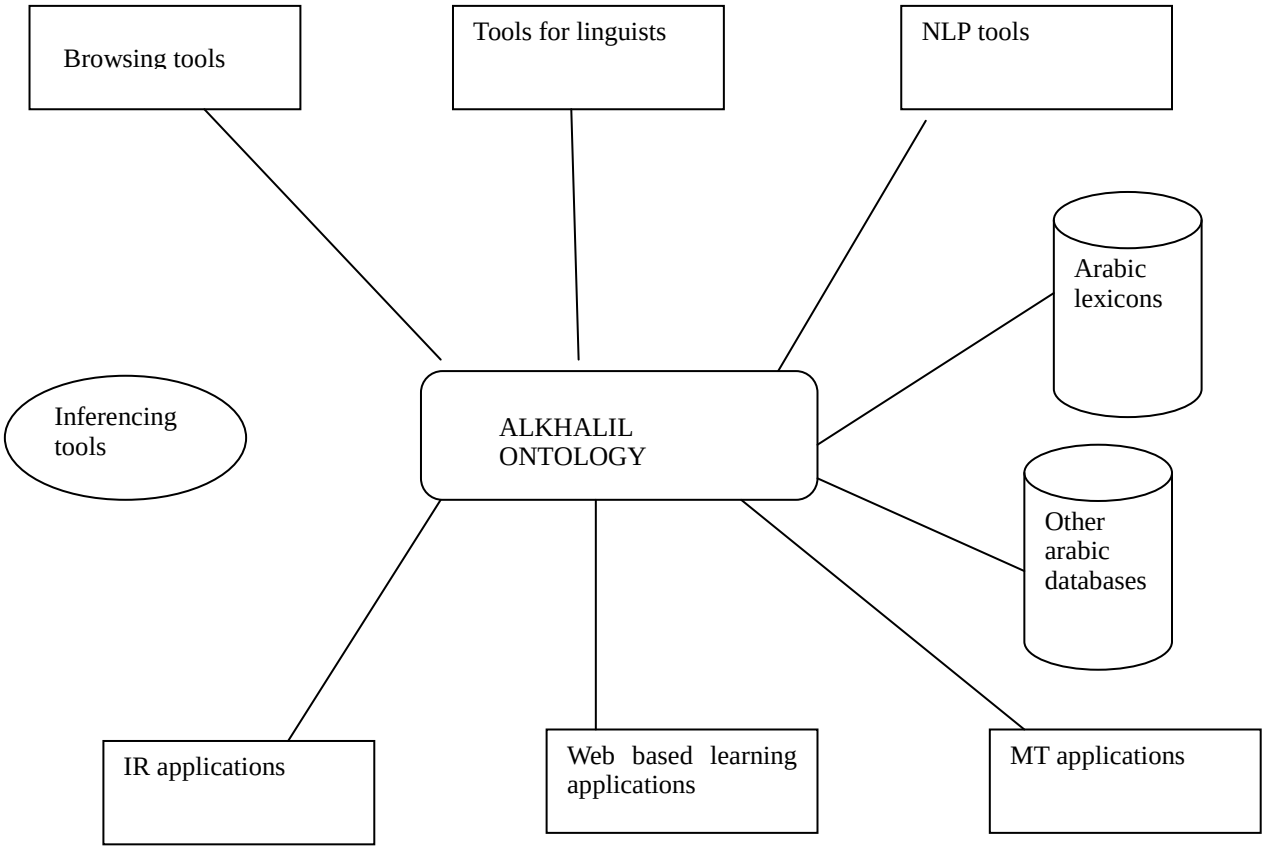

Figure5 AL-KHALIL Ontology Proiect

5. References

Aliane H. (2006) "An ontology based approach for multilingual information retrieval" proceedings of ICTTA 2006.

Black, W. and S. Elkateb, H. Rodriguez, H, Alkhalifa, M., Vossen, P., Pease, A. and Fellbaum C." Introducing the Arabic WordNet Project", in *Proceedings of the Third InternationalWordNet Conference*, Sojka, Choi, Fellbaum and Vossen eds.

Daoud M &al (2009)"Collaborative construction of Arabic lexical resources" *2nd international conference on Arabic language resources and tools, cairo*.

Dichy, J. and A. Braham, S. Ghazali, M. Hassoun (2002) " La base de connaissances linguistiques DIINAR.1*"* , *Proceedings of the International Symposium on The Processing of Arabic.*

Farrar, S. and T. Langendoen (2003) "A linguistic ontology for the semantic web" *Glot International Vol 7, N° 3, March 2003", (P 97-100)*

Farrar S. (2003) «An ontological account of linguistics: Extending SUMO with GOLD." *Proceedings of the 2003 IEEE International Conference on Natural Language Processing and Knowledge Engineering.*

Magnini, B. and S. Speranza (2002) "Merging global and Specialized Linguistic Ontologies". *Proceedings of Ontolex 2002 pp. 43-48.*

Fensel D. (2000): Ontologies: *Silver Bullet for Knowledge Management and Electronic Commerce*, *URL:http://www.cs.vu.nl/%7Edieter/ftp/paper/silverbullet.pdf*

Graham W. (2007) "an OWL Ontology for HPSG" *proceeding of the ACL 2007 demo and poster sessions, 169-172.*

Gruber T. (2000), "What is an Ontology?" *URL: Http:// www-ksl.stanford.edu/kst/what-is-an-ontology.html*