# Towards Building Lexical Ontology via Cross-Language Matching

**Mamoun Abu Helou**
Birzeit University
Birzeit, Palestine
mabuhelou@birzeit.edu

**Matteo Palmonari**
Milano Biccoca University
Milano, Italy
palmonari@disco.unimib.it

**Mustafa Jarrar**
Birzeit University
Birzeit, Palestine
mjarrar@birzeit.edu

**Christiane Fellbaum**
Princeton University
Princeton, United State
fellbaum@princeton.edu

## Abstract

In this paper, we introduce a methodology for mapping linguistic ontologies lexicalized across different languages. We present a classification-based semantics for mappings of lexicalized concepts across different languages. We propose an experiment for validating the proposed cross-language mapping semantics, and discuss its role in creating a gold standard that can be used in assessing cross-language matching systems.

## 1 Introduction and Motivation

Sharing data on the Web meaningfully requires capturing the semantics behind the data. On the word level, meaning can be represented in digital lexical resources (lexicons) that are amenable to automatic processing and reasoning for a range of intra- and interlingual applications.

A *lexicon* is the inventory of word forms and meanings of a language. Each lexical entry specifies several linguistic properties of a word (such as its phonetics, morphology, and syntax) as well as its semantics. In a relational model of the lexicon, a words meaning is reflected in its relations to other words (Miller and Fellbaum 1991).

With the emergence of the Semantic Web, ontologies have gained great attention in research as well as in industry for enabling knowledge representation and sharing. An *ontology* in general, is a formal representation of critical knowledge that enables different systems sharing this knowledge to communicate meaningfully. Ontologies are perceived as language-independent representations of concepts and their interrelations, thereby allowing intelligent agents and applications to access and interpret the Web contents automatically.

Because some lexicons combine aspects of a lexicon with those of an ontology, they are often called linguistic ontologies (Hirst 2004, Jarrar 2010). A *linguistic ontology* can be seen both as a lexicon and as an ontology (Hirst 2004; Jarrar 2010), and is significantly different from domain ontologies. Because it is not constructed for a specific domain. Linguistic ontologies can be seen as semantic networks covering most common concepts in a natural language and provide knowledge structured on lexical items (words) of a language by relating them according to their meanings (concepts).

One such commonly used linguistic ontology is WordNet (Fellbaum 1998). WordNet was conceived as a lexicon, but the emergence of wordnets in other languages and the need to map them have raised the need to consider not just the lexical inventory of these languages (i.e., the word forms, word senses and their interrelations) but also their conceptual inventory, a set of categories of objects (concepts) that share the same properties and the relations among them.

In this paper we discuss the role of cross-language ontology matching methods in linking linguistic ontologies in different languages. In particular we investigate the semantics of cross-language mappings, and the problem of creating a gold standard to evaluate alternative ontology matching methods. We propose a classification-based semantic approach for mappings among concepts lexicalizations. We define a linguistic-based classification task that allows us to support the design of experiments to validate cross-language mappings and to enable us to build a gold standard that can be used to assess the performance of automatic cross-language matchers. Then, such mapping methods can be used to discover mappings at large-scale and solve the problem of creating large-scale linguistic ontologies in a (semi)-automatic way.

**The construction of linguistic ontologies** followed the success of WordNet and was motivated by the need for similarly structured lexicons for individual and multiple languages (multi-language lexicons). Both the "merge" (where a wordnet is first built manually from scratch) and the "expand" model (which proceeds largely by translation, Vossen 1998) are used to build wordnets in languages other than English. EuroWordNet (Vossen 2004) and MultiWordNet (Pianta et al. 2002) cover a number of European languages. In the EuroWordNet approach both models were used. Mappings among the different wordnets are represented in the Inter-Lingual Index, which is considered to be language independent. Whenever possible, entities from the individual wordnets are linked to the Inter-Lingual Index by means of equivalence and near-equivalence relations. MultiWordNet applied the expand model, and all wordnets are aligned as strictly as possible to the English WordNet under the assumption that most of the concepts are universally shared. However, Vossen (1996) argued that wordnets developed using the expand technique are overly influenced by English WordNet and thus retain its mistakes and structural drawbacks. However, the merge model strategy is more labor and cost-intensive. Wordnets for many languages have been constructed under the guidelines of Global WordNet Association[1], which aims to coordinate the production and linking of wordnets.

**Automatic construction of wordnets** is another method for building and linking wordnets, using machine translation techniques. The BabelNet project (Navigli and Ponzetto, 2012) used machine translation to provide equivalents in various languages for English WordNet synsets. While this approach might be suitable for certain NLP applications (de Melo and Weikum, 2012), it usually fails to account for the fact that different languages encode subtle socio-cultural aspects that do not always have straightforward translation equivalents. Cimiano et al. (2010) argued that translation tools (to some extent) might remove the language barrier but not necessarily the socio-cultural one; there is a need to find the appropriate word sense of the translated word that is not reflected in the literal translation equivalent. Moreover, Hirst (2004) argued that languages do not cover exactly the same part of the lexicon and, even where they seem to be common, several concepts are lexicalized differently.

**Ontology-based cross-language matching** is the process of establishing correspondences (find relations) among the ontological resources from two independent ontologies where each ontology is lexicalized in a different natural language (Spohr et al. 2011).

A common approach for cross-language ontology matching is based on transforming a cross-language matching problem into a mono-language one by translating the ontology elements of one ontology in the language adopted by the other ontology using automatic machine translation tools (e.g., Fu et al. 2012). Spohr et al. (2011) argued that the quality of machine translation systems is limited and depends greatly on the pair of languages considered. As a consequence, a pure *translation-based* approach is not sufficient to find a significant amount of mappings.

Although some techniques such as explicit semantic analysis (Gabrilovich and Markovitch 2007) proved to perform well in cross-language ontology matching (Narducci et al. 2013), it is important to understand how reliable automatic matching methods are in this domain. Before selecting and/or extending the more appropriate existing cross-language ontology matching techniques, we need to be able to compare alternative methods and to assess the quality of their output. Moreover we recognized that although a variety of cross-language ontology matching methods have been proposed, the semantic nature of cross-language mappings that cross-language ontology matching methods are expected to find has not been sufficiently investigated.

This motivated us to understand the *formal semantics of mappings among linguistic ontologies – lexicalization patterns across different languages*, and to investigate the specification of their intended meaning. In other words, providing a formal interpretation of the mapping semantics allows us to define a set of inference rules and to derive mappings (relations) from a set of existing mappings.

The research presented here aims to contribute to the Arabic Ontology project (Jarrar 2011). Our idea is to semi-automate this process by (1) matching Arabic concepts to English WordNet concepts, and (2) deriving the semantic relations among the Arabic concepts using relations among concepts in the English WordNet.

[1] http://globalwordnet.org/

The rest of the paper is structured as follows. In section 2, we introduce the Arabic Ontology project and describe the semi-automatic method by which it was created. Section 3 describes the cross-lingual ontology matching problem. In section 4, we illustrate the proposed approach. In section 5, we define an experimental setting for validating the proposed approach and its role in creating a gold standard for assessing cross-language mapping methods. In section 6, we conclude and outline possible future steps.

## 2 The Arabic Ontology

The Arabic Ontology (Jarrar 2010) aims to build a linguistic ontology for Arabic. The Arabic Ontology is a formal representation (using FOL) of the concepts that the Arabic terms convey. The Arabic Ontology can be seen and used as an *Arabic wordnet*; however, unlike WordNet, the Arabic Ontology is logically and philosophically well-founded, and follows strict ontological principles (Jarrar 2011).

The "*top levels*" of the Arabic Ontology are derived from philosophical notions (Jarrar et al. 2013), which are used to ensure the ontological correctness of the lower levels. The top levels of the Arabic Ontology constitute a classification of the most abstract concepts (i.e., meanings) of the Arabic terms. All concepts in the Arabic Ontology are classified under these top levels. These concepts are designed based on a deep investigation of the philosophy literature and well-recognized upper level ontologies like BFO (Smith. 1998), DOLCE (Gangemi et al. 2003a), SUMO (Niles and Pease 2001), and KYOTO (Casillas at al. 2009).

### 2.1 Semi-automatic Construction of the Arabic Ontology via Cross-Language Matching

In addition to that the Arabic Ontology that is being built manually at Sina Institute in Birzeit University [2], there are also hundreds of dictionaries that have been digitized and integrated into one lexical database. This database provides a good source for Arabic synsets (concepts), but lack semantic relations among the concepts. We argue that, by mapping such Arabic concepts into their conceptually equivalences in WordNet, one can (automatically) infer the relations among the Arabic concepts from the relations among the English concepts. The resultant relations can provide an initial set of relations that can be manually validated and corrected.

However, mapping synsets lexicalized in different languages is a challenging task. Cross-language ontology matching techniques (Spohr 2011; Fu 2012) can play a crucial role in bootstrapping the creation of large linguistic ontologies and, for analogous reasons, in enriching existent ontologies. We also remark that the above considerations do not apply to the Arabic ontology only, but our definitions and approach are general and can be reused for other languages.

## 3 Cross-Lingual Ontology Matching

Euzenat and Shvaiko (2007) defined *ontology matching* as a process that tries to establish *correspondences* among semantically related ontological entities, without explicitly specifying the natural languages used to label the ontological entities (e.g., concepts, relations, descriptions, and comments). We recall the definition of correspondence (mapping) presented in [Jung et al., 2009].

***Definition 1***: *Correspondence*, Given a source ontology $O_S$, a target ontology $O_T$, and a set of alignment relations $\mathcal{R}$, a correspondence is a quadruple: *correspondence:*$= < c_S; c_T; r; n>$, $c_S \in O_S$, $c_T \in O_T$. Where $r \in \mathcal{R}$, a set of alignment relations (e.g., $\equiv$, $\sqsubseteq$, or $\perp$), and $n \in [0, 1]$ is a confidence level (i.e., measure of confidence in the fact that the correspondence holds).

The largest part of the ontology matching strategies (see, Shvaiko and Euzenat 2013) involve syntactic and lexical comparisons, making ontologies for different languages very difficult to match. Ontology entities are expressed in natural language by associating them with terms (i.e., a lexicon) that belong to one (or more) natural languages. We denote the term *lexicalization* as the process of associating ontology entities with a set of terms that belongs to a set of natural languages, and the term *lingualization* as the process of retrieving the set of languages that the associated terms belong to.

According to Spohr et al. (2011), an ontology $O$ is lexicalized in a given language $l$, if the ontology terms are lingualized in language $l$, such that $l$ belong to the set of natural languages $L$ ($l \in L$). Ontologies can be lexicalized in one language (monolingual ontology), two languages

[2] http://sites.birzeit.edu/comp/ArabicOntology/

(bilingual ontology) or more languages (multilingual ontology). Spohr and his colleagues also distinguished between the matching tasks based on the number of languages used to lexicalize the ontology terms.

Given two ontologies $O_S$ and $O_T$ , which are lexicalized in two sets of natural languages $L_S$ and $L_T$ respectively, we can define the cross-language ontology matching as the process of establishing relations or correspondences among ontological resources from two independent ontologies, where each ontology is lexicalized in (a) different natural language(s), but they do not share any language.

In the recent past, a *translation-based* approach has been used to transform the cross-language problem into a mono-language ontology matching one (e.g., Fu 2012). However, the cultural-linguistic barriers (Gracia et al. 2012) still need to be overcome in terms of the mapping process and techniques, as well as to formally define the semantic mappings that align concepts lexicalized across different natural languages. That is, the semantics of mapping among concepts lexicalized in different natural languages is still unsolved.

In general, a community of users (speakers) would consider two concepts that are lexicalized in two languages to be equivalent if both terms are used to indicate the same meaning in a given context. The context (or discourse) that a community of speakers shares in order to decide if these two terms (lexemes) refer to the same concept is "not only to explain what people say, but also how they say it. Lexical choice, syntax, and many other properties of the formal style of this speech are controlled by the parliamentary context" (Van Dijk, 2006).

Our main objective is to define the semantics of cross-language mapping among concepts lexicalization. This includes the formal representation and interpretation (i.e., formal semantic) of these mappings. We start from definitions and approaches proposed for mono-language ontology matching and we extend them to cross-language ontology matching.

# 4 Mapping Semantics in Cross-Language Ontology Matching

This section presents the classification-based interpretation for the cross-language mapping problem. We discuss the extension of the definition of the classification-based approach from formal interpretation (Atencia et al. 2012) to an interpretation that covers the concept lexicalization.

## 4.1 Classification-based Interpretation of Mappings

Ontology mapping can be seen as an expression that establishes relations among elements of two (or more) heterogeneous ontologies. A crisp mapping tells us that a certain concept is related to other concepts in different ontologies and specifies the type of relations, which are typically a set of formal relations $\{ \equiv, \sqsubseteq, or \perp \}$. A *weighted mapping* ( see definition 2) in addition associates a number (weight) to those relations. We start from the definition of weighted mapping and its semantic presented in (Atencia et al. 2012) that we recall below.

***Definition 2*: *Weighted Mapping*,** Given two ontologies $O_1$ and $O_2$, a weighted mapping from $O_1$ to $O_2$ is a quadruple: *weighed mapping:=* $\langle C, D, r, [a, b] \rangle$, where $C$ and $D$ are two concepts such that $C \in O_1$ and $D \in O_2$, $r \in \{\sqsubseteq, \equiv, \sqsupseteq, \perp\}$, $a$ and $b$ are real numbers in the unit interval [0, 1].

Intuitively, the semantics of the mapping $\langle C, D, r, [a, b] \rangle$ is that the relation $r$ maps the concept $C$ to the concept $D$ with a confidence that falls into the closed interval $[a, b]$, where $a$ and $b$ represent respectively the lower and upper bounds of such an interval.

Following a standard model-theoretic *formal semantics* based concepts are intuitively interpreted as set of instances. An interpretation $\mathfrak{I}$ is a pair $\mathfrak{I} = \langle \Delta^{\mathfrak{I}}, \cdot^{\mathfrak{I}} \rangle$ where $\Delta^{\mathfrak{I}}$ is a non-empty set, called domain of interpretation $\mathfrak{I}$, and $\cdot^{\mathfrak{I}}$ is a function that interprets each concept (class) $C$ in the set of concepts $\mathcal{C}$ as a non empty subset of $\Delta^{\mathfrak{I}}$, and each instance identifier ($x \in X$) as an element of $\Delta^{\mathfrak{I}}$. Intuitively, for a given ontology $O$, if $\mathcal{C}$ is a set of concepts, $\mathcal{R}$ is a set of relations, and $X$ is a set of shared individuals. Then $C^{\mathfrak{I}} \subseteq \Delta^{\mathfrak{I}}$ for $C \in \mathcal{C}$, $r^{\mathfrak{I}} \subseteq \Delta^{\mathfrak{I}} \times \Delta^{\mathfrak{I}}$ for r$\in \mathcal{R}$, and $x \in \Delta^{\mathfrak{I}}$ for $x \in X$.

**Weighted mappings semantics**, Atencia et al. (2012) provide a formal semantics of weighted mapping among independent ontologies, that assumes a classification-based interpretation of mappings. Let $C$ be a concept of $O_1$ and $x_k$ an individual of $X$; we define $X$ as a *shared context* (domain) of the mapping. We say that $x_k$ is classified under $C$ according to $\mathfrak{I}_1$ if $x_k^{\mathfrak{I}_1} \in C^{\mathfrak{I}_1}$. Then, the set $C_X^{\mathfrak{I}_1} = \{x \in X \mid x^{\mathfrak{I}_1} \in C^{\mathfrak{I}_1}\}$

represents the subset of individuals of $X$ classified under $C$ according to $\mathfrak{T}_1$. Note that $C_X^{\mathfrak{T}_1}$ is a subset of $X$ ($C_X^{\mathfrak{T}_1} \subseteq X$), whereas $C^{\mathfrak{T}_1}$ is a subset of the domain of the interpretation $\mathfrak{T}_1$ ($C^{\mathfrak{T}_1} \subseteq \Delta^{\mathfrak{T}_1}$). In addition, $C_X^{\mathfrak{T}_1}$ is always a finite set, while $C^{\mathfrak{T}_1}$ may be infinite.

Figure 1, demonstrates the extensional meaning between two concepts $C$ and $D$ in the ontology $O_1$ and ontology $O_2$ respectively, with the classification-based mapping approach. $\mathfrak{T}_1$ and $\mathfrak{T}_2$ represent respectively an interpretation of $O_1$ and $O_2$. $\Delta^{\mathfrak{T}_1}$ and $\Delta^{\mathfrak{T}_2}$ represent the domain of interpretation of $\mathfrak{T}_1$ and $\mathfrak{T}_2$, respectively. The sets $C_X^{\mathfrak{T}_1}$ and $D_X^{\mathfrak{T}_2}$ represent the subsets of individuals $x_k$ in $X$ classified under $C$ according to $\mathfrak{T}_1$, and under $D$ according to $\mathfrak{T}_2$, respectively. The Individuals $z$ and $y$ represent individuals that do not belong to $X$.

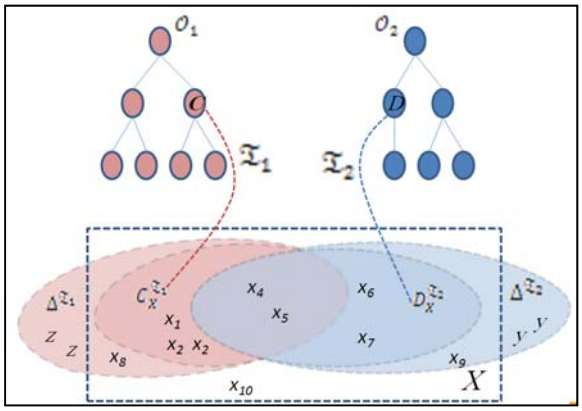

Figure 1: The extensional meaning of a concept and the common interpretation context.

The classification-based approach examines the relation among two concepts $C$ and $D$ that are in the ontology $O_1$ and $O_2$ respectively, by considering a common context (the shared domain $X$), defined as a set of common instances classified under the two ontology concepts. The different types of mappings $\langle C, D, r, [\mathrm{a}, \mathrm{b}]\rangle$ are obtained by looking at the different relation $r \in \{\sqsubseteq, \sqsupseteq, \equiv, \perp\}$. Atencia et al. use precision, recall, and F-measure, as used in the context of classification tasks, for the formalization of weighted subsumptions ($\sqsubseteq, \sqsupseteq$) and equivalence ($\equiv$)relations, respectively.

Following the classification perspective, a weighted subsumptions mapping $\langle \mathrm{C}, \mathrm{D}, \sqsubseteq, [\mathrm{a}, \mathrm{b}]\rangle$ interpreted as follows : the number of individuals of $X$ classified under C according to $\mathfrak{T}_1$ which are (re-)classified under D according to $\mathfrak{T}_2$. The weighted mapping can be seen as the recall of $\mathrm{C}_X^{\mathfrak{T}_1}$ w.r.t $\mathrm{D}_X^{\mathfrak{T}_2}$.

$$R\left(C_X^{\mathfrak{T}_1}, D_X^{\mathfrak{T}_2}\right) = \frac{\left|C_X^{\mathfrak{T}_1} \cap D_X^{\mathfrak{T}_2}\right|}{\left|C_X^{\mathfrak{T}_1}\right|} \in [a, b]$$

In the same way, the weighted mapping $\langle C, D, \sqsupseteq, [\mathrm{a}, \mathrm{b}]\rangle$ which falls in the confidence level interval [a, b], is used to express the number of individuals of $X$ classified by $D$ according to $\mathfrak{T}_2$which are (re-) classified under $C$ according to $\mathfrak{T}_1$. Then the weighted mapping can be seen as the precision of $D_X^{\mathfrak{T}_2}$ *w.r.t* $C_X^{\mathfrak{T}_1}$.

$$P\left(C_X^{\mathfrak{T}_1}, D_X^{\mathfrak{T}_2}\right) = \frac{\left|C_X^{\mathfrak{T}_1} \cap D_X^{\mathfrak{T}_2}\right|}{\left|D_X^{\mathfrak{T}_2}\right|} \in [a, b]$$

Intuitively, the F-measure can be used to express the equivalence relation that aligns two concepts $C$ and $D$ where $\langle C, D, \equiv, [\mathrm{a}, \mathrm{b}]\rangle$ represent that F-measure falls into the confidence interval [a, b]. The F-measure is the harmonic mean of precision and recall. Typically the F-measure is used to evaluate the global quality of a classifier, the *F-measure* of $C_X^{\mathfrak{T}_1}$ and $D_X^{\mathfrak{T}_2}$ is defined as:

$$F\left(C_X^{\mathfrak{T}_1}, D_X^{\mathfrak{T}_2}\right) = 2.\frac{\left|C_X^{\mathfrak{T}_1} \cap D_X^{\mathfrak{T}_2}\right|}{\left|C_X^{\mathfrak{T}_1}\right| + \left|D_X^{\mathfrak{T}_2}\right|} \in [a, b]$$

An interesting point in the above weighted mapping definition is the use of an interval [a,b] to define an uncertain (yet bounded) weight associated with a mapping. Using such intervals - as a more general notation for mapping weights - we can define the equivalence relation as a conjunction of the two subsumption relations. This in particular gives the notion of logical consequences of weighted mappings that allows to define a set of inference rules to derive a mapping from a set of existing mappings. For instance, if we have weighted mappings $\langle C, D, \sqsubseteq, [h, j]\rangle$ and $\langle C, D, \sqsupseteq, [e, f]\rangle$, then we can derive the equivalence weighted mapping $\langle C, D, \equiv, [v, w]\rangle$ with $v = \min(h, e)$ and $w = \max(j, f)$.

Notice that, if we consider the usual definition of equivalence in DLs in terms of subsumption: $\langle C \equiv D\rangle$ *iff* $\langle C \sqsubseteq D\rangle$ and $\langle C \sqsupseteq D\rangle$, when dealing with single weight values for precision ($\sqsupseteq$) and recall ($\sqsubseteq$) instead of intervals, it is usually impossible to combine them into a single value by simple conjunction (Atencia et al. 2012). Nevertheless, generally ontology matchers are used to return a single confidence level value, for instance, $n$. Accordingly, to represent the value $n$ by means of the weighted mapping interval [a, b], the authors (Atencia et al. 2012) suggest to use a pointwise interval; we can assume that $a$=$b$, then $n$=[a, a]. Thus, we can simply present the weighted mapping relation as $\langle C, D, r, n\rangle$.

Assume that the set of individuals $\{\mathrm{x}_1, \ldots, \mathrm{x}_{10}\}$ (see Figure 1) are classified under $O_1$and $O_2$. If

the individuals $\{x_1, \ldots, x_5\}$ are classified under concepts $C \in O_1$ and the elements $\{x_4, \ldots, x_7\}$ are classified under the concept $D \in O_2$, we can represent the subsumption relations $\langle C, D, \sqsubseteq, 0.4\rangle$ and $\langle C, D, \sqsupseteq, 0.5\rangle$ by computing the recall and precision, respectively. Then we can deduce the equivalence relation between $C$ and $D$ by computing the F-measure $\langle D, C, \equiv, 0.44\rangle$.

### 4.2 Classification-based Interpretation of Mappings in Cross-Language Ontologies

In what follow, we extend (Atencia et al. 2012) approach, which fits our problem and provides a good foundation for the cross-language mapping problem for several reasons. Many matching methods, in particular those for cross-language ontology matching, use metrics that evaluate the overlap between the entities (e.g., ontology individuals, documents, pieces of text) that are classified under two concepts. Also, the approach provides a very general definition of classification context (the set of instances considered for the interpretation of mappings), which can support the definition of a formal framework to interpret translations among ontology concepts that are lexicalized in different languages. Atencia et al. assume a formal interpretation of a concept denoted as class of instances in an interpretation domain.

Classification is interpreted as the task to establish whether an instance $i$ is member of a class $C$, i.e., if $i$ belongs to the extension of $C$. This extensional interpretation cannot be directly applied for ontologies that are not formally represented and interpreted in set theoretic semantics. For instance, when we annotate a document we can consider the concept as classifying an object, but the interpretation of classification here is different; in this case, saying that a concept classifies an object means that the concept represents the topic of the document. If we consider a sentence and we want to disambiguate the meaning of the words in it, we can consider the *disambiguation task* as a form of classification, namely, the classification of a word as occurrence of a word sense in the sentence.

We *hypothesize* that in order to share a meaning (concept) we have to share a domain of interpretation, and this domain represents the shared context of a community of languages speakers. Considering the extensional based approach, particularly the case of cross-lingual extensional meaning of a concept, we should keep in mind that according to a given shared context, it is *not* necessary that all objects classified under $C_S$ ($x \in C_{X,S}^{\mathfrak{T}_1}$) are also instances under $D_T$ ($x \in D_{X,T}^{\mathfrak{T}_2}$) according to an interpretation $\mathfrak{T}_1$ and $\mathfrak{T}_2$, respectively. It happens that an object $x \in C_{X,S}^{\mathfrak{T}_1}$ might *not* exist in the other language (or, ontology) ($x \notin D_{X,T}^{\mathfrak{T}_2}$), or even it might be classified under another concept such as ($x \in E_{X,T}^{\mathfrak{T}_2}$).

Recall that a synset is a set of words that all lexicalize and denote the same concept. Such words, called synonyms, are equivalent in that they carry the same meaning, even when not all synonyms are stylistically felicitous in all contexts. For example, the phrase "empty vessel" sounds good, while "vacant vassal" does not; "empty" is more frequently used than vacant in this context, in spite of the fact that both adjectives convey the same meaning. Note that "empty" and "vacant" are freely interchangeable when modifying nouns like "room" and "house."

Consider a corpus of sentences, where each sentence expresses a context and a word in the sentence represent the usage of a concept. If a majority of speakers (i.e., bilingual native speakers or lexicographers) can substitute two words, each belonging to a different language, in a sentence and both words indicate the same sense (meaning), then they can be used interchangeably to refer to the same concept (word sense).

We *hypothesize* that, if speakers can substitute two words in a given context, then these words are synonyms and give an equivalent meaning (concept) (Miller and Fellbaum 1991). This is valid also for intra- and interlingual substitution, as concepts are independent of specific languages. We assume the above hypothesis but, instead of considering the cross-language substitutability of words themselves, we consider the cross-language substitutability of meanings associated with these words, by referring to *co-disambiguation* (see definition 3) of words across ontologies in different languages.

**Definition 3**: *Co-disambiguation Task*, let $WSD(w_i)$ be a function called Word Sense Disambiguation, such that $w_i$ is an occurrence of the word $w$ in a sentence $S$. *WSD* associates $w_i$ with a sense in a lexicon (e.g., WordNet). Accordingly, we can define a *cross-language WSD* function $CL\text{-}WSD_{[L1>L2]}(w_i)$, such that *CL-WSD* associates a word $w_i$ in a language $L_1$

(where $L_1$ is the language used in $S$) with a sense in a lexicon lexicalized in another language $L_2$.

By extending the classification-based semantics defined in (Atencia et al. 2012) with the consideration of the *CL-WSD* classification task, we map a sense $C$ (lexicalized in $w_1$ using $L_1$) to a sense $D$ (lexicalized in $w_2$ using $L_2$) (i.e., represent conceptually-equivalence word senses) if *most* of the bilingual speakers accept that $CL\text{-}WSD_{[L1>L2]}(w_1) = C$, and $CL\text{-}WSD_{[L1>L2]}(w_1) = D$. At the same time accept that $CL\text{-}WSD_{[L2>L1]}(w_2) = C$, and $CL\text{-}WSD_{[L2>L1]}(w_2) = D$.

For example, in the sentence "the student sat around the table (طاولة) to eat their lunch", the words "table" and (طاولة, pronounced Tawlah) indicates the same meaning (a table at which meals are served). If most of the speakers would co-disambiguate "table" with the English word sense $Table_n^3$ (the third noun sense in WordNet for table - a piece of furniture with tableware for a meal laid out on it), and with the Arabic word sense{طاولة Tawlah, منضدة Mndada, مائدة Ma'ad, سفرة Soufra}, then $Table_3^n$ and {طاولة Tawlah, منضدة Mndada, مائدة Ma'ad , سفرة Soufra} denote the same concept.

In another words, if the substitution of the words does not change the meaning of the context, then they are conceptually equivalent. In view of this, *CL-WSD* can be seen as a classifier, where the number of agreements among the lexicographers (bilingual speakers) expresses the confidence (i.e., the weight) of the mapping.

The speakers perform the *CL-WSD* tasks, and the mapping between two word senses depends on a frequency-based function that measures the degree in which the two senses in two different languages co-disambiguate the same word sense in multiple contexts (sentences). Suppose we have a corpus of English sentences, we find a word $w_{en}$ that appears in these sentences. We disambiguate each occurrence of $w_{en,i}$ with an English word sense $C_i$; we disambiguate each occurrence of $w_{en,i}$ with a synset $D_i$ in Arabic. As a result of this operation we found two sets of distinct concepts $\bar{C}$ and $\bar{D}$ that have been used to disambiguate $w_{en}$ respectively in English and Arabic. For each $C_i \in \bar{C}$ we count the number of $C_i$ that has been co-disambiguated with every $D_i \in \bar{D}$. The co-disambiguation fraction of the two concepts $C$ and $D$ represent the degree at which we can consider $C$ *as a subclass of* $D$.

Although we use a classification task that differs from the one proposed in (Atencia et al. 2012), we can still use the inference rule they proposed to reason about mappings, to infer new mappings from existing mappings. Moreover, using the *CL-WSD* function as a classification task to evaluate the existence of relations among concepts, we can define a method to establish reference relationships between concepts by performing *CL-WSD* on sentence corpuses

## 5 Experiment Design for Cross-Language Mapping Validation

We present an experimental setting whereby the proposed cross-language mapping semantics can be evaluated and a gold standard to assess the quality and to compare alternative cross-language mapping methods can be generated.

In order to validate the equivalent relation we need to perform the following *CL-WSD* classification tasks: given a parallel corpus ( or two corpuses) which lexicalized in English and Arabic. We disambiguate each occurrence of $w_{en,i}$ in English sentences with a word sense $C_i$ and $D_i$ in English and Arabic respectively. In this way, we obtain two sets of distinct concepts $\bar{C}$ and $\bar{D}$ that have been used to disambiguate the English word $w_{en}$ respectively in senses form English and Arabic. For each $C_i \in \bar{C}$ we count how many times $C_i$ has been co-disambiguated with every $D_i \in \bar{D}$. The co-disambiguation count for the two concepts $C$ and $D$ represent the degree (confidence level) at which we can consider $C$ *as a subclass of* $D$.

In the same way, we disambiguate each occurrence of $w_{ar,i}$ in Arabic sentences with a word sense $C_i$ and $D_i$ in English and Arabic respectively. The distinct set of concepts $\bar{C}$ and $\bar{D}$ have been used to disambiguate the Arabic word $w_{ar}$ respectively in senses from English and Arabic. For each $D_i \in \bar{D}$ we count the number that $D_i$ has been co-disambiguated with every $C_i \in \bar{C}$. The proportion of the co-disambiguation for the two concepts $D$ and $C$ represent the confidence level at which we can consider $D$ *as a subclass of* $C$.

Then we use the F-measure to interpret the confidence level of the equivalent relation that aligns the two concepts $C$ and $D$.

However, it might be difficult and costly to make such experiment at large scale. One way is to use available sense annotated corpuses. Nevertheless, such an Arabic corpus is not available. Therefore, we propose to mine the subclass relations starting form a sense annotated English corpus, we *CL-WSD* the English words with the equivalent Arabic senses, and then we

check if these relations can be converted to equivalence relations by exploiting the structure (relations) of the WordNet.

The proposed experiment corresponds to a classification task; asking bilingual speakers to perform a *CL-WSD*$_{[En>Ar]}$ classification task. We collect sentences from "*Princeton Annotated Gloss Corpus"*, a corpus of manually annotated WordNet synset definitions (glosses). The selected sentences are annotated with at least one sense that belongs to "*Core WordNet"*. The reason for selecting Core WordNet concepts is that they represent the most frequent and salient concepts and thus can shared among many or most languages. Accordingly, we hypothesize that mapping the core WordNet concepts to the equivalent Arabic concepts will form the core for the Arabic Ontology. Then we can extend it to include more cultural and language-specific concepts.

For each English word sense, a number of bilingual speakers (lexicographers) are asked to provide the equivalent Arabic word sense. For each word sense, the lexicographers substitute the English word with one of the Arabic synsets, which have been developed at Sina Institute and classified under the top levels. Using available bilingual dictionaries the lexicographers select the best translation. In Figure 2, in the sentence "the act of starting to construct a *house*", the English word "house" was *CL-WSD* with the English sense *house*$_1^n$ and the Arabic sense (منزل, Mnzel) [3]. For the same sentence we substitute the sense *house*$_1^n$ with its direct hypernym (subclass) sense *home*$_1^n$ from the WordNet. We *CL-WSD* the sense *home*$_1^n$ with the Arabic sense (بيت, Baet). Ideally, we should be able to deduce the subclass relation between (منزل) and (بيت).

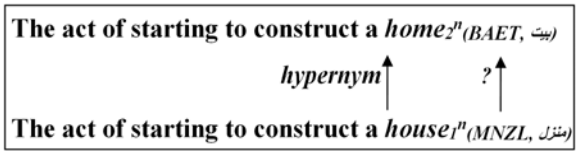


Figure 2: Example of *CL-WSD* task and a possible inference.

However, as mentioned before, not every concept is lexicalized in both (all) languages. The mappings thus obtained will form an initial semantic network. However, conflicts and overlaps might exist. The top levels concepts can control and eliminate part of this problem. For example, the associated concepts should be classified under the same top concept. This direction of work also taking into account the relations confidence level will be pursued in the future.

We plan to experiment with the proposed mapping approach on a large scale by considering all 5,000 Core WordNet concepts and to simulate the majority of speakers by incorporation larger number of bilingual speakers (lexicographers). We suggest adopting a crowdsourcing method (e.g., Amazon Mechanical Turkey (Sarasua et al. 2012) to collect feedback from larger number of lexicographers. A significance result of a full-scale version of the proposed experiment is to generate a gold standard for cross-language mappings. That can be used to assess the various automatic cross-language matching systems as well to validate the proposed semantic mapping. Thereby selecting or extending such mapping methods that can be used to discover mappings at large-scale and solve the problem of creating large-scale linguistic ontologies in a (semi)-automatic way. Moreover, we can validate the language-dependence hypothesis of the salient (core) concepts. In addition, we plan to investigate the explicit semantic analysis approach in the cross-language mapping settings (Sorg and Cimiano 2012) to enhance the word sense selection (conceptual translation) task.

## 6 Conclusion and Future Works

We introduced a classification-based mapping for cross-language matching purposes. We illustrated the proposed approach and outlined future steps. We plan to implement a large-scale experiment that covers the Core WordNet concepts and to adopt a crowdsourcing method to simulate the community agreements. In addition to bilingual dictionaries for word senses selection (conceptual translation), explicit semantic analysis techniques will be used. Moreover, we plan to investigate the extent to which the process of (semi)- automated creation is suitable for creating a linguistic ontology. We will formally define the mapping weight based on the proposed *CL-WSD* task. Finally, we aim to define and develop algorithms for semantic relations inference and to validate such methods using the cross-language mappings gold standard.

[3] Translation was obtained using Wikipedia inter-lingual links.

**Acknowledgments**
This research is funded by EU FP7 SIERA project (no. 295006).

## References

Manuel Atencia, Alexander Borgida, Jérôme Euzenat, Chiara Ghidini and Luciano Serafini. 2012. A formal semantics for weighted ontology mappings. In ISWC-2012, pp17-33.

Philipp Cimiano, Elena Montiel-Ponsoda, Paul Buitelaar, Mauricio Espinoza and Asunción Gómez-Pérez. 2010. A note on ontology localization. Applied Ontology, 5(2).

Arantza Casillas, Arantza Diaz de Illarraza, Kike Fernandez, Koldo Gojenola, Egoitz Laparra, German Rigau, Aitor Soroa. 2009. The Kyoto Project. In Proc. SEPLN´09, Spain, September.

Gerard de Melo and Gerhard Weikum. 2012. Constructing and utilizing wordnets using statistical methods. Language Resources and Evaluation, 46(2):287-311.

Jérôme Euzenat and Pavel Shvaiko. 2007. Ontology matching. Springer.

Christiane Fellbaum., editor. 1998. WordNet: An Electronic Lexical Database (Language, Speech, and Communication). The MIT Press.

Bo Fu, Rob Brennan and Declan O'Sullivan. 2012. A configurable translation-based cross-lingual ontology mapping system to adjust mapping outcomes. Journal of Web Semantics, (V15)15-36.

Evgeniy Gabrilovich and Shaul Markovitch. 2007. Computing semantic relatedness using wikipediabased explicit semantic analysis. In Proceedings of the 20th IJCAI'07, pp1606–1611, San Francisco, CA, USA.

Jorge Garcia, Elena Montiel-Ponsoda, Philipp Cimiano, Asunción Gómez-Pérez, Paul Buitelaar, John McCrae. 2012. Challenges for the multilingual web of data. JWS. (V11):63-71.

Aldo Gangemi, Nicola Guarino, Claudio Masolo and Alessandro Oltramari. 2003a. Sweetening WordNet With DOLCE, AI Magazine, 24(2003), pp. 13–24.

Graeme Hirst. 2004. Ontology and the Lexicon, in Handbook on Ontologies and Information Systems. eds. S. Staab and R. Studer. Heidelberg: Springer.

Mustafa Jarrar., 2010. The Arabic Ontology. Lecture Notes, Knowledge Engineering Course (SCOM7348), Birzeit University, Palestine.

Mustafa Jarrar. 2011. Building a Formal Arabic Ontology (Invited Paper). In proceedings of the Experts Meeting on Arabic Ontologies and Semantic Networks. Alecso, Arab League. Tunis.

Mustafa Jarrar, Hiba Olwan, Rana Rishmawi. 2013. Classification of the most Abstract Concepts in Arabic - The Top Levels of the Arabic Ontology. Technical Report, Version 1. Sina Institute, Birzeit University, Palestine.

Jung Jason J. Jung, Anne Håkansson and Ronald Hartung, 2009. Indirect Alignment between Multilingual Ontologies: A Case Study of Korean and Swedish Ontologies. In Proc. of the 3rd Inter. KES, LNAI 5559, pp.233-241.

George A. Miller and Christiane Fellbaum. 1991. Semantic networks of English. Cognition, 41, 197-229.

Fedelucio Narducci, Matteo Palmonari and Giovanni Semeraro.2013. Cross-language Semantic Retrieval and Linking of E-gov Services. 12th ISWC, October, Australia

Ian Niles and Adam Pease. 2001. Towards a Standard Upper Ontology, in The 2nd International Conference on (FOIS-2001), Ogunquit, Maine.

Emanuele Pianta, Luisa Bentivogli, Christian Girardi. 2002. MultiWordNet: developing an aligned multilingual database. 1st GWC, India, January.

Barry Smith. 1998. The Basic Tools of Formal Ontology, in Nicola Guarino (ed.), Formal Ontology in Information Systems. Amsterdam, Oxford, Tokyo, Washington, DC: IOS Press (FAIA-98), 19-28

Philipp Sorg and Philipp Cimiano. 2012. Exploiting Wikipedia for cross-lingual and multilingual information retrieval. Data&Know. Eng.,74:26–45.

Dennis Spohr, Laura Hollink and Philipp Cimiano. 2011. A machine learning approach to multilingual and cross-lingual ontology matching. In Proc. of ISWC-11, Springer.

Pavel Shvaiko and Jérôme Euzenat. 2013. Ontology matching: State of the art and future challenges. IEEE Trans. Know. Data Eng., 25(1):158-176.

Cristina Sarasua, Elena Simperl and Natalya F. Noy. 2012. CROWDMAP: Crowdsourcing Ontology Alignment with Microtasks. In ISWC-2012,. Springer.

Teun A. Van Dijk. 2006. Discourse context and cognition . Discourse Studies, 8:159-177.

Piek Vossen. 1996. Right or wrong. combining lexical resources in the EuroWordNet project. In Pro. of Euralex-96, page 715728, Goetheborg.

Piek Vossen. 1998. Introduction to Eurowordnet. Computers and the Humanities, 32(2):7389.

Piek Vossen. 2004. EuroWordNet: a multilingual database of autonomous and language-specific wordnets connected via an Inter-Lingual-Index. International Journal of Lexicography, Vol.17.