Managing Language Diversity Across Cultures: the English-Mongolian Case Study

Amarsanaa Ganbold
School of Information Technology
National University of Mongolia
Ulaanbaatar, Mongolia
amarsanaag@disi.unitn.it

Feroz Farazi
DISI
University of Trento
Trento, Italy
farazi@disi.unitn.it

Moaz Reyad
DISI
University of Trento
Trento, Italy
reyad@disi.unitn.it

Oyundari Nyamdavaa
School of Information Technology
National University of Mongolia
Ulaanbaatar, Mongolia
oyundari.n@gmail.com

Fausto Giunchiglia
DISI
University of Trento
Trento, Italy
fausto@disi.unitn.it

Abstract—**Developing ontologies from scratch appears to be very expensive in terms of cost and time required and often such efforts remain unfinished for decades. Ontology localization through translation seems to be a promising approach towards addressing this issue as it enables the greater reuse of the ontological (backbone) structure. However, during ontology localization, managing language diversity across cultures remains as a challenge that has to be taken into account and dealt with the right level of attention and expertise. Furthermore, reliability of the provided knowledge in the localized ontology is appearing as a non-trivial issue to be addressed. In this paper, we report the result of our experiment, performed on approximately 1000 concepts taken from the space ontology originally developed in English, consisted in providing their translation into Mongolian.**

Keywords: Ontology localization, space ontology, space domain, ontology, Semantic Web, knowledge, provenance

I. INTRODUCTION

This paper is a long version of [1], in which it is described that building a true, flourishing and successful Semantic Web [2] should involve the participation from all cultures and languages across the world. In the development of the traditional Web, this participation was spontaneous and has been made possible as the necessary tools and resources were available. In the Semantic Web one crucial feature is the capacity to assign precise meaning to words, for instance in order to diminish the impact of polysemy. Still for many languages, one example being Mongolian, such resources are not developed at all and for some others what is out there cannot be used effectively as they could not achieve critical mass. However, for English much progress has been made and the WordNet (http://www.princeton.edu) developed at Princeton is one of the well-known and most widely used resources in the field. Yet its coverage is often unsatisfactory when dealing with domain specific tasks [3].

Towards solving the issue of the lack of coverage and to gain a critical mass of concepts, some domain ontologies have already been developed. A prominent example is the *space ontology* [4] developed in English with comparatively very large coverage of geo-spatial features and entities around the globe. Domain ontologies can also deal with the specificity of an area of knowledge, for example, by providing relations and attributes specific to the domain. By reducing polysemy (the amount of words with same meaning), they can enable better semantic interoperability.

Ontologies that are developed to perform NLP tasks in one language can hardly be used with their full potential for another. Representing an existing ontology in a new language, taking into account cultural and linguistic diversities, is defined as ontology localization.

In this paper, we describe the development of the space ontology in Mongolian starting from its English counterpart as it is contained in the Universal Knowledge Core (UKC), as described below. Building an ontology without human-level accuracy is a potential obstacle in developing applications (e.g., word sense disambiguation and document classification). Synset base resources (linguistic representation of ontologies) such as WordNet and FinnWordNet [5] are built manually to obtain better quality.

Knowledge is created to be consumed by others in a multitude of activities including daily life, education, research and development for the advancement of our society. Through ontology development and ontology localization we create new knowledge. Trustworthiness and reliability of the produced knowledge are crucial measures that if comprehended, modelled and communicated properly would make consumers lives comparatively easier.

Being concerned about the quality and giving utmost importance to it, we followed a manual approach. The contributions of our paper include:

i) The development of an ontology localization methodology that is domain and language independent and seems to achieve very high quality

ii) The development of a methodology for dealing with diversity (e.g., lexical gaps) across cultures and languages

iii) The lessons learned from the execution of the whole process in the generation of the space ontology in Mongolian

iv) The development of the provenance model that manages information about various contributors to the ontology localization process for ensuring quality and credibility of the knowledge produced.

The paper is organized as follows. In Section 2, we provide detailed description of the UKC. Section 3 gives an overview of the space ontology. In Section 4, we describe the macro-steps of the translation process. In Section 5, we present the provenance model. In Section 6, we describe the diversity across English and Mongolian cultures in terms of space related features. Section 7 reports the experimental results. Section 8 discusses the lessons learned while Section 9 describes the related work. Finally, in Section 10, we provide the concluding remarks.

II. The Universal Knowledge Core

The UKC [4] is a large-scale ontology, under development at the University of Trento, which includes hundreds of thousands of concepts (e.g., lake, mountain chain) of the real world entities (e.g., Lake Garda, Alps). It consists of three main components: *domain core*, *concept core* and *natural language core* (See Fig. 1).

Domain core: As described in [4], the domain core consists of various **domains,** where each of them represents an area of knowledge or field of study that we are interested in or that we are communicating about [6]. In other words, a domain can be a conventional subject of study (e.g., mathematics, physics), an application of pure disciplines (e.g., engineering, mining), the aggregation of such fields (e.g., physical science, social science) or a daily life topic (also called Internet domains, e.g., sport, music). Each domain is organized in **facets,** where a facet can be defined as a hierarchy of homogeneous concepts describing the different aspects of meaning [7]. According to our methodology [8], called DERA, where D stands for Domain, facets are classified into three categories: Entity class (E), Relation (R) and Attribute (A). For example, in the space ontology, country and continent are **entity classes**. **Relations** describe relations between entities; examples of spatial relations are near, above, far etc. An **attribute** is a property of an entity, e.g., depth of a lake.

Concept core: The concept core consists of concepts and semantic relations between them. The concepts in the concept core form a directed acyclic graph, which provides the terms and the structure from which facets are defined. Entity class, relations and attributes are all codified as concepts. A **concept** is a language independent representation of a set of words (synset), which are synonym of a given word in natural language. For example, country, city, etc. The concept *city* can be represented as *city* in English, *città* (chit'a) in Italian, *хот* (khot) in Mongolian.

A semantic relation is a relation that holds between two concepts. Some examples of semantic relation are is-a (or *hyponym-of*), part-of (part-meronym-of) and value-of. An instantiation of the is-a relation can be given as *city* is-a *populated place*.

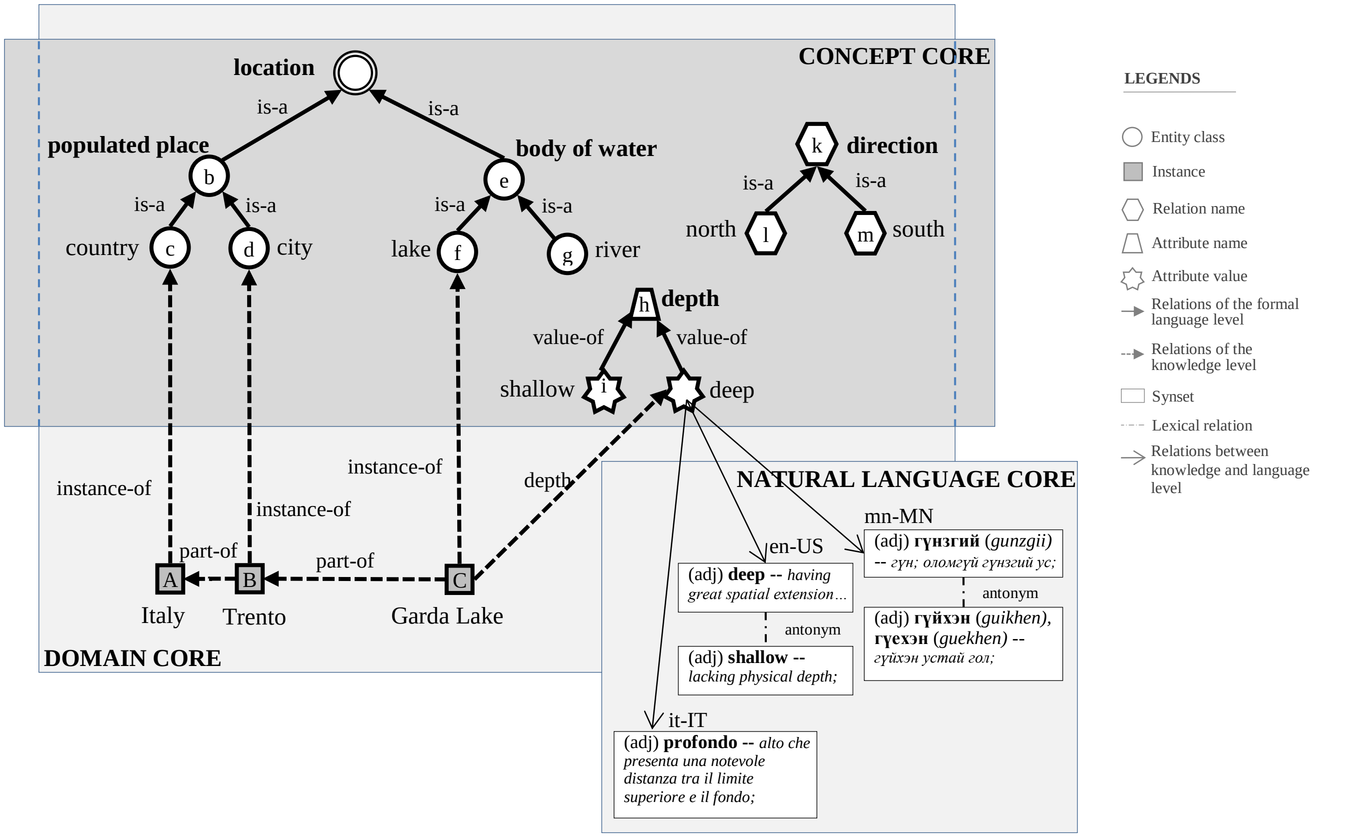

CONCEPT CORE
location
is-a
is-a
populated place
body of water
b
e
is-a
is-a
is-a
is-a
country
c
d
city
lake
f
g
river
k
direction
is-a
is-a
north
l
m
south
h
depth
value-of
value-of
shallow
i
deep
instance-of
instance-of
instance-of
depth
part-of
part-of
A
B
C
Italy
Trento
Garda Lake
DOMAIN CORE
NATURAL LANGUAGE CORE
mn-MN
en-US
(adj) **гүнзгий** (*gunzgii*) -- *гүн; оломгүй гүнзгий ус;*
(adj) **deep** -- *having great spatial extension…*
antonym
antonym
(adj) **shallow** -- *lacking physical depth;*
(adj) **гүйхэн** (*guikhen*), **гүехэн** (*guekhen*) -- *гүйхэн устай гол;*
it-IT
(adj) **profondo** -- *alto che presenta una notevole distanza tra il limite superiore e il fondo;*
LEGENDS
Entity class
Instance
Relation name
Attribute name
Attribute value
Relations of the formal language level
Relations of the knowledge level
Synset
Lexical relation
Relations between knowledge and language level

Figure 1. Knowledge Organization in the UKC

Natural language core: The natural language core consists of a set of languages, each representing a set of linguistic objects and relations between them. The objects of this core are words, senses, synsets and exceptional forms. A **word** is the basic lexical unit of the natural language core represented as a lemma. It can be multiword, phrase, collocation, etc. The words in the natural language core provide, for any given language, the translation of the concepts stored in the concept core.

Word senses are organized into four part-of-speeches -- noun, verb, adjective and adverb. One word may have more than one part-of-speech, and synonym word senses with the same part-of-speech are grouped into a synset. A **sense** is a possible meaning for a word. A word can have one or more senses each having a part-of-speech tag. Each sense corresponds to and belongs to only one synset. All senses of a given word are ranked according to most preferred usage. A **synset** is a set of words, which share the same meaning. In fact, words in a synset have semantically equivalent relations. Each synset might be accompanied by a gloss consisting of a definition and optionally example sentences.

Relations of the language core are of type lexical and semantic lexical. This kind of relations holds between the objects of the same language.

A lexical relation is a relation that holds between the words of different synsets. Antonym, derivationally-related-form and also-see are examples of such relation. An example of the antonym relation can be provided as *lowland* is an antonym of *highland*. Note that hereinafter we represent synsets with double hyphens distinguishing words (comma separated) from glosses, which are formatted in *Italics*.

(a) lowland -- *low level country*

(b) highland, upland -- *elevated (e.g., mountainous) land*

The word *highland* of the synset reported in (b) is in antonymy relation with the word *lowland* of the synset (a). Notice that the same relation does not hold between the other word *upland* of the synset (b) and *lowland*.

A semantic-lexical relation is a relation that holds between two synsets. Some examples of this kind of relation are similar-to, troponymy and verb-group. An example of the semantic-lexical relation can be *adjacent* is similar-to *near*.

(c) adjacent -- *near or close to but not necessarily touching*

(d) near, close, nigh -- *not far distant in time or space or degree or circumstance*

In this case the synset (c) that consists of only one word is in similar-to relation with the synset (d) that consists of three words. This means that the very same relation can be applied between any word of the synset (c) and any word of the synset (d).

The natural language core is built with the complete integration of hierarchically organized synset bases, as it is the case, for instance, for WordNet and the Italian part of MultiWordNet (http://multiwordnet.fbk.eu).

III. THE SPACE DOMAIN

The space domain [4], [6] is a large-scale geospatial ontology built using the faceted approach. It was developed as the result of the complete integration of GeoNames (http://www.geonames.org) and WordNet. It is also known as space ontology and in this paper, we refer to it with any of these names. It currently consists of nearly 17 facets, around 980 concepts and 8.5 million entities. The ontology (excluding entities) is integrated into the UKC. Some examples of facet are *land formation* (e.g., mountain, hill), *body of water* (e.g., sea, lake), *administration division* (e.g., state, province) and *facility* (e.g., university, industry).

In Fig. 2, we provide a partial bird's eye view of the whole set of facets. Note that facets are not connected to each other and they do not have concept overlap across or within them.

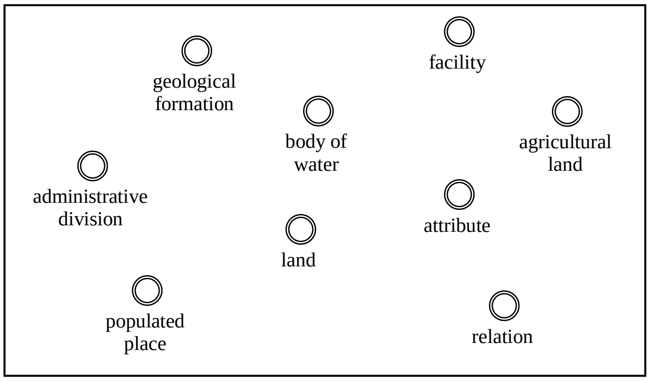


Figure 2. A subset of the facets of the *Space* domain

Fig. 3 shows a small portion of the facet *geological formation* in which the second level represents *natural elevation*, *natural depression* and the level below the *natural elevation* is organized into *oceanic* and *continental elevation*, and so forth.

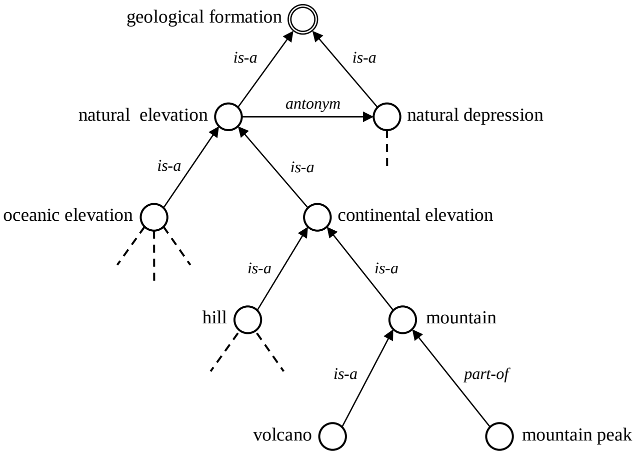


Figure 3. An entity class (E) category facet (partial view)

Note that, within a facet with double circled node we distinguish the root concept from the rest of the concepts that are represented with single circle.

In the Space domain, the *relation* category contains around 10 facets such as *spatial relation* and *primary outflow*. A partial representation of the *spatial relation* facet is shown in Fig. 4.

The *spatial relation* is the spatial property between geological physical objects or the way in which something is located. Leaf nodes of this facet represent relations between entities. For instance, Mongolia is *south* of Russia and *north* of China. The relation *primary outflow* connects two bodies of water.

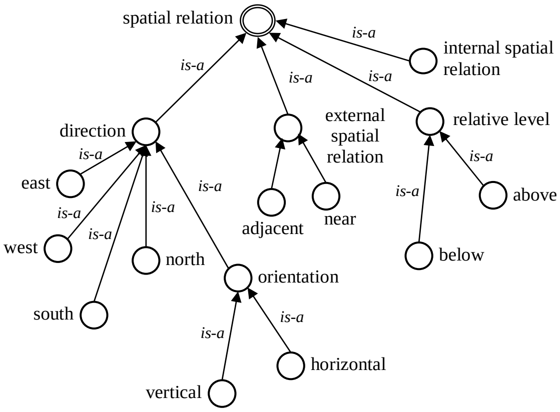


Figure 4. A relation (R) facet (partial view)

Within the domain the *attribute* category consists of around 20 facets such as *rain* and *temperature*.

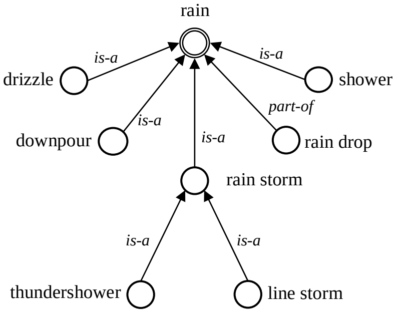


Figure 5. An attribute (A) facet (partial view)

As shown in Fig. 5 the facet *rain* includes among others *rainstorm*, *downpour*, *drizzle* and *shower*. With rain we mean *falling of water in drops from vapor condensed in the atmosphere*. The temperature indicates *the degree of hotness or coldness of an object or environment*.

IV. Translation Approach

In the following subsections we describe the general process for translation and its instantiations both for translating and creating a concept in the target language.

A. General process

The main idea of the translation process is to take the objects of the domain of interest from a source language, in this case English, and to produce the corresponding representation in a target language, e.g., Mongolian in order to extend the UKC with translations. The process includes the translation of the synset words and glosses. A direct translation of them is provided whenever possible. However, the world is full of diversity and people of a particular culture might not be aware of some concepts. For instance, Mongolia is a landlocked country, thus some terms (e.g., *dry dock*, *quay*, *pier*, etc.) related to seaport are not known to the community or are rarely used and are often a source of lexical gaps for Mongolian. Lexical gaps are those concepts that do not have a succinct representation in a given language. However, they can be expressed as a free combination of words [10].

We select English as a master source language for all localization activity since the language is the second most widespread language and will be a common language to use in scientific and research oriented discussions. For executing the translation process, English language representation of the concept core is copied to an LKC (Local Knowledge Core) repository, which contains translations in the target language.

In order to provide the most suitable translation for a synset, we follow the macro-steps described below and represented in Fig. 6.

1(a) A **language translator** takes a synset provided in the source language and gets a clear understanding of its meaning. In case of difficulty, he/she finds the corresponding images or videos of the synset word(s) on the Web to perceive the concept through visualization.

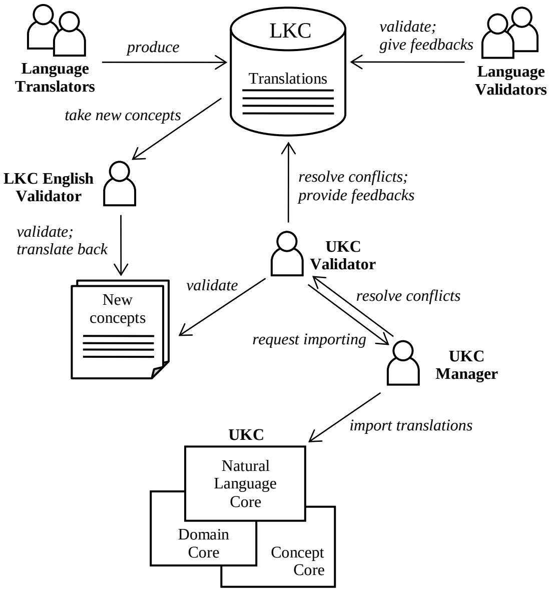


Figure 6. Translation phases of UKC

1(b) The **language translator** provides a suitable translation of the word(s) in the target language.

With suitable we mean word, multiword, co-occurrence and phrasal representation as we do not allow a free combination of words as translation of a word. In case of unavailability of the word(s) for the given meaning, the translator can mark it as a lexical gap. However, the translator always provides the translation of the gloss.

1(c) A **language validator** evaluates the translation of the word(s) and the gloss of the synset. In case the concept is marked as a gap, the validator either confirms the gap or suggests a translation for the word(s).

1(d) Upon receiving feedback on the synset, the **language translator** goes through the comments and updates the translation when necessary. In case of disagreement, the language translator provides comments including mostly the rationale about the disagreement.

1(e) The **language validator** reevaluates the updated translation. In case of disagreement, the validator generates further feedback and sends it back to the language translator (step 5). Even if after a few iterations a disagreement is not resolved, a second language validator is consulted. If agreed upon, the validation for the given synset is over.

2 In the cases where the **language translator** finds out a new concept that might be a lexical gap or a missing concept in the source language, she suggests a suitable synset for this concept in the target language and if possible also the corresponding synset in the source language. The **LKC English validator** evaluates the source language synset (if suggested) for the new concept coming from the target language. Otherwise, she translates back this target language synset into English.

3 A **UKC validator** takes the translations resulting from steps 1 and 2 to evaluate their correctness from both the language and UKC perspectives. The validator corrects the mistakes and resolves the issues (if any) communicating with the language validator and LKC English validator (if necessary), possibly in a few iterations. Finally, she asks a UKC manager for importing the translation to the UKC.

4 The **UKC manager** runs an automatic validation tool to evaluate if the provided input is compliant with the UKC. In case of errors are found, they are corrected with the help of the UKC validator (if needed) possibly iterating a few times. The manager also decides new concepts originating from the target language whether to accept or reject them. Once all the issues are resolved, the UKC manager imports the translations to the UKC.

Following these steps we translated the *space ontology* into Mongolian end-to-end, evaluated and finally imported the translations to the UKC.

To achieve optimal quality while executing the whole process depicted in Fig. 6, we set the criteria that translators and various validators must possess the competences necessary for the task. The language translator should be a native speaker from the country of origin of the target language with a good command of the source language. The language validator should be a linguist possessing the necessary language competences. The LKC English validator should be as close as possible to an English native speaker who should understand well the target language. The UKC validator is a native speaker of the target language with knowledge of the UKC. Both the UKC validator and LKC English validator are in charge of the language dependent tasks in the translation process. The UKC manager is an expert on the UKC with no specific competence on the language.

From a geographical point of view we expect that, in most cases, the language core will be developed in the countries where that language is spoken, while the UKC is and will be developed centrally. The UKC validator, whenever possible, should operate centrally where the UKC manager is. This spatial distribution of operations and operators has been designed as an attempt to preserve local diversity and, at the same time, to deal with the need for central coordination required because of existence of a unique, single UKC. The underlying model is that there is a single world, represented by the UKC, and many different views of the world, each represented by a different natural language. The diversity of the world is therefore captured, as it will be described in detail in the next section, in the mapping from the informal natural languages and the unique UKC formal concept language.

B. *Translating a concept*

Here we instantiate the general translation process described in section A by taking the concept subtree, rooted at “mountain”, shown in Fig. 3, Translation is performed from English to Mongolian. All three concepts of the subtree are provided as follows:

> mountain, mount -- *a land mass that projects well above its surroundings; higher than a hill*
>
> volcano -- *a mountain formed by volcanic material*
>
> peak, crown, crest, top, tip, summit -- *the top point of a mountain or hill*

Note that the subtree should normally get translated according to the macro steps 1, 3 and 4, whose executions are marked with prime (') and described below.

1(a)' The **language translator** perceives the meaning of the concept *mountain* by reading its gloss. The translator could understand the concept as a massive land that is highly raised than surrounding geological formations and it is more elevated than the hills. She checks whether there is at least a term to refer to this notion in Mongolian culture. In case of dilemma in understanding its availability, she also visualizes the meaning by consulting resources (e.g., images, videos) on the web.

1(b)' The **language translator** represents *mountain* as *уул (*uul*)* in Mongolian. In this case, the Mongolian representation is a lexical unit, therefore the concept is not marked as a lexical gap. The gloss is translated as *эргэн тойрон буюу хүрээлэн буй* *орчноосоо дээш өргөгдөн гарсан өндөрлөг газар; толгодоос өндөр* (a high land raised above and elevated from its surroundings and all-around; higher than hills).

1(c)' The **language validator** agrees with the term provided as the name (or label) of the concept, but she suggests an improved translation of the gloss as *эргэн тойрны орчноосоо дээш өргөгдсөн өндөрлөг газар; дов толгодоос өндөр;* (a high land raised above from its surroundings; higher than hills). Here, the **language validator** removes the words *эргэн тойрон* (all-around) and *гарсан* (elevated) from the gloss developed by the **language translator** because from her point of view without these terms in the gloss, the concept can clearly be understood by the native speakers.

1(d)' The **language translator** receives the validation feedback on the translation of the concept *mountain*. She accepts the validated result and updates the translation according to the language validator's comments.

1(e)' As the **language translator** accepts the modifications proposed by the language validator, no conflict is left to be resolved and the validator proceeds with the next steps.

3' The UKC validator checks the translation of the terms and glosses of the concept. Since there are no disagreements from both the language and UKC perspectives, she asks a UKC Manager to import the translations into the UKC.

4' Finally, the UKC Manager runs an automatic import function to integrate the translation into the UKC.

Similarly, the other two concepts *volcano* and *peak* are translated and then integrated into the UKC.

C. Adding a new concept

New concepts can be added executing macro steps 2, 3 and 4 of the general process of translations. The instantiations of these steps are marked with prime ((') and described below.

2' The **language translator** realizes that in the given subtree a concept, which is part-of the concept *mountain*, is missing. Therefore she proposes to create a concept and develops a synset for this in Mongolain as follows:

гэзэг (gezeg) -- *уул толгодын ар шил*

The **LKC English validator** recognizes that this concept is a gap in English and translates its gloss back to English as follows:

GAP -- *northern ridge of a mountain*

3' The UKC validator verifies the translation of the gloss, confirms that it is a gap in English and evaluates the correctness of the added relation of the concept in the subtree. As she conceded with the produced knowledge from linguistic and UKC viewpoints, she proceeds through asking the UKC Manager to incorporate the translations into the UKC.

4' The UKC Manager takes the necessary steps for the integration of the new knowledge to the UKC.

Note that, the UKC infrastructure was developed taking into account the fact that new knowledge can come to the system at any point in time. Therefore, it supports integration of new objects (e.g., concepts and relations). UKC manages the lifecycle and evolution of the objects by exploiting the timestamp attached to them.

V. Handling Provenance

Provenance can be defined as the source of a piece of knowledge. In the context of this work, we assign fine grained provenance in terms of knowledge objects and their contributors.

For ensuring trustworthiness and reliability, we employ a provenance model that is defined and then instantiated with the use cases -- concept translation and concept addition -- in the following subsections.

A. Representing provenance

In this section, we present the data structure that we use to represent the provenance. Our provenance model is designed to maintain information about the elements that can be created in the knowledge base during the ontology localization process. These elements are concepts, lexical gaps, lexical relations, semantic lexical relations, synsets, senses, words and sense ranks.

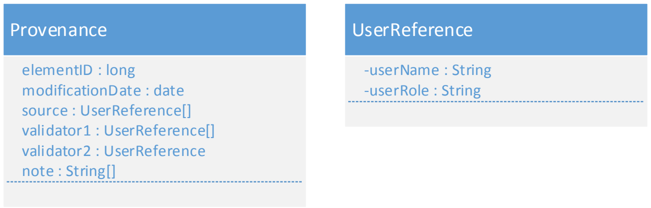


Figure 7. UML Diagram for provenance

Fig. 7 shows the UML diagram of our provenance model, which consists of two classes – *Provenance* and *UserReference*. *Provenance* represents information about the source of an element and *UserReference* represents a human user who is involved in the localization process.

Provenance is a 5-tuple <*elementID*, *modificationDate*, *source*, *validator1*, *validator2*, *note*^{op}>, where *elementID* is the unique identifier of the knowledge base element to which the provenance is applied, *modificationDate* is a timestamp specifying the latest date of modification of the provenance, *source* refers to a list of contributors who are language translators responsible mainly for translating an element and also for proposing a missing concept in the target language, *validator1* refers to a list of contributors who are language validators responsible for validating translations, *validator2* is the UKC validator and *note*^{op} refers to the additional remarks that can optionally be provided by LKC developers/ validators or UKC validator/manager.

UserReference is a pair <*userName*, *userRole*>, where *userName* represents the name and email address of a human user and *userRole* indicates one of the following editorial roles: *LKC_Developer*, *LKC_Validator* and *UKC_Validator, UKC manager*.

Note that, in this provenance model, for the translation tasks, there are references to at least three human contributors – *source*, *validator1* and *validator2*. That means each manually translated element is validated by one or more user, who can also be communicated via email to discuss the rationale behind the translations and concepts developed by them. This we believe is the main strength of our provenance modelling and that helps increase the reliability of the ontologies localized in any target language following our approach.

B. *Provenance in concept translation*

In the concept translation macro-steps, we create and update provenance in the following cases.

a. If a language translator develops a word, a synset, a lexical gap or a concept in Steps 1(a)-1(b) or Step 2, for each element created, a new provenance will be generated with the *source* referring to the translator.
b. If a language validator confirms an already developed element in Step 1(c) and 1(e), the provenance of that element is updated by linking *validator1* to the instance of the UserReference, which corresponds to the name and email and the role of the language validator that is LKC_Validator. This marks the element as validated.
c. As soon as the UKC validator confirms an element in Step 3, the provenance of that element is updated with the instantiation of the attribute *validator2* referring to the name and email and role of the validator (i.e., UKC_Validator) of the given context. This marks the element as completely validated and accepted.

Note that the modification date changes whenever a new operation is performed on the provenance. In (a) it refers to the date of translation, in (b) it is updated with the date of the language validation and in (c) it is replaced with the date of the UKC validation.

C. *Provenance in concept addition*

While adding concept, we create and update provenance in the following possible scenarios.

a. In the concept addition phase in Step 2, the language translator may create a new concept and its related lexical components such as synset, word, etc. in the target language and optionally in the source language In this case the provenance *source* for each of the objects will be instantiated with the translator for her LKC developer role.
b. Again in Step 2, (i) if the LKC English validator evaluates the source language synset provided by the language translator, the provenance *validator1* is instantiated with her LKC Validator role (ii) if it happens that the LKC English validator translates back the new concept into the source language, in this case the *source* is filled in with her role as LKC Developer and the *validator1* is left un-instantiated.
c. Similarly to the concept translation, a UKC validator checks the correctness of the concept addition and she becomes *validator2* with the corresponding role.

VI. TYPES OF DIVERSITY

The translation or localization is the adaptation of a piece of knowledge to a particular language and culture [9]. This is nontrivial and linguistic experts might help in this task. Moreover, the localization should be based on the perception of the concepts and entities in the real world within the local communities and not on the literal translation.

A. *Concepts*

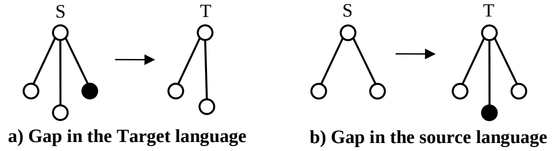
We assume concepts to be universal. However, their representation in the different natural languages changes. Within the same language a concept might be referred with multiple terms (known as synonymy) and multiple concepts might be referred with the same term (known as polysemy).

The concepts *valley*, *dale* and *hollow* are represented with the same term in Mongolian.

> ***valley*** *– (a long depression in the surface of the land that usually contains a river)*
>
> ***dale*** *– (an open river valley (in a hilly area))*
>
> ***hollow*** *– (a small valley between mountains; "he built himself a cabin in a hollow high up in the Appalachians")*

Moreover, in the UKC *dale* and *hollow* are subordinate concepts of *valley*. In this case, translating them into the target language increases polysemy. However, we translate them because within the Mongolian culture people can classify their (real world) entities under the specific concept.

Moreover, a concept might not have a name in a target language the fact that it can be a lexical gap. For example, the concept *parish - (the local subdivision of a diocese committed to one pastor)* is a lexical gap in Mongolian. The variation in the concept lexicalization from the source language (S) to the target language (T) is depicted in Fig. 8(a).



a) Gap in the Target language **b) Gap in the source language**

Figure 8. Variations of concept localization

As the lexical gap is a feature of the languages, it does happen with all of them. There can be a gap also from the target to source language. For instance, the Mongolian words **бууц** (buuts) and **буйр** (buir) are gaps in English. The word *buuts* can be represented in English as *an area of dried and accumulated manure where a nomadic family was living* and the word *buir* can be represented in English as *a round shaped spot where a nomadic yurt was built*. Note that these words lack a succinct representation in English. Therefore we consider them as gaps. This phenomenon is drawn in Fig. 8(b).

The nomadic lifestyle of Mongolians is the source of these concepts that are not used in the English speaking cultures across the globe.

Words pointing to lexical-gap concepts might appear also in the glosses. For instance, the term *piers* appearing in the gloss of *Romanesque architecture* is a lexical gap in Mongolian. In such cases, the translation is produced with a free combination of words.

> *Romanesque architecture – (...characterized by round arches and vaults and by the substitution of* ***piers*** *for columns and profuse ornament and arcades)*

B. Senses

In the space ontology, some words have multiple senses that have subtle difference in meaning. For instance, the word fissure has two senses:

> [S1]: *crack, cleft, crevice,* ***fissure****, scissure – (a long narrow opening)*
>
> [S2]: ***fissure*** *– (a crack associated with volcanism)*

The two concepts associated with the given word are hyponyms of *continental depression* and they can be represented with the same word(s) in the target language. This phenomenon is shown in Fig. 9(a).

Polysemous words in the source language might correspond to lexical gaps for a subset of senses. For instance, *gorge* has two senses within Space ontology and one of them is a gap as depicted in Fig. 9(b), where 'mn' and 'en' denote Mongolian and English, accordingly.

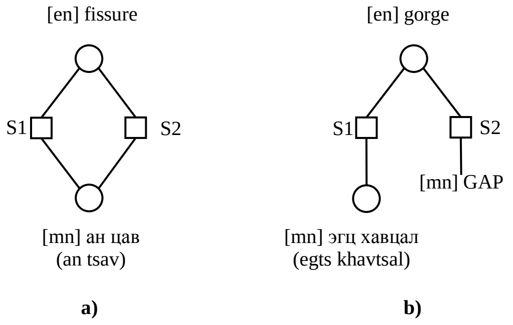


Figure 9. Word sense diversity

C. Synsets

Words in a synset can be directly translated into the target language. However, for some of them there might be a lack of translation. For example, the synset *mountain peak (the top point of a mountain or hill)* has 6 words of which 3 of them lack translation into Mongolian as shown below.

1 peak	→	оргил (ogril)
2 crown		
3 crest		
4 top	→	орой (oroi)
5 tip		
6 summit	→	дээд оргил (deed orgil)

In gloss paraphrasing, some parts of the glosses sometimes are obtained using words with a very close or similar meaning instead of exact translation. Though our first preference is to provide the exact translation, in many cases this could not be achieved. The following example shows a paraphrased translation where the phrase "near a shore" is eliminated from Mongolian version. In this situation, there is no difference between bank and shore in Mongolian language.

> [in English] ***oceanic sandbank*** *– a submerged bank of sand near a shore, can be exposed at low tide*
>
> [in Mongolian] ***далайн элсэн эрэг*** (gl. oceanic bank of sand) *– шунгаж орсон далайн элсэн эрэг, далайн давалгааны намхан хаялганд үзэгддэг (gl. a submerged sea bank of sand, visible at low tide)*

Example sentences in glosses were also paraphrased or added newly in order to provide a better explanation. For example, well-known place names are often substituted in the target language because famous names within a culture might give better understanding about a concept being translated. The highest mountain peak of the Alps ridge is Mont Blanc that is substituted with Everest as it is known to the most of the people in the East Asian region. Moreover, symbols are kept in their original forms, e.g., measurement unit symbol, pH.

Date and time format, measurement unit and currency were converted into the ones used regionally. For example, 5 inches is converted into 12.7 centimeters because of the pervasive use of MKS system in Mongolia. Note that these types of words appear only in glosses. However, using these types of word might not be suitable as fractions are less intuitive than whole numbers. For example, 3 feet is converted into 0.9144 meters. Such fractions cannot be mapped easily to the real world entities and most often become tedious to remember.

VII. Results

In this section, we report the results of our experiment. We could translate 91.88% of the concepts of the space ontology into Mongolian and the remaining 8.12% were identified as lexical gaps. In Table I, we report the detailed statistics of the translation task and the obtained results.

In Table I, the number of concepts per facet is shown separately, e.g., administrative division has 18 concepts, agricultural land has 19 concepts and so on. Note that for the sake of space, we group the statistics of all attribute facets as attribute and relational ones under relation.

Language Translators provided the Mongolian translation for 905 concepts *Language Validators* provided feedback on each of the produced synset words and glosses separately that help us achieving better quality. The validation procedure identified 188 disagreed words and 243 disagreed glosses. Cases such as disagreements and modifications for improvement were solved in iterations (as many as needed) between the translators and validators until they reached to an agreement. The highest number of iterations was recorded as 4.

TABLE I. LOCALIZATION RESULT OF THE SPACE DOMAIN

Facets	Concepts	Translated	Disagreed words	Disagreed glosses	Translator Identified Gaps	Finally accepted Gaps	Finally Localized Concepts
administrative division	18	18	2	4	0	0	18
agricultural land	19	19	2	1	0	0	19
attribute	85	73	1	23	12	10	75
barren land	7	7	1	0	0	0	7
facility	357	357	54	64	0	2	355
forest	5	5	5	4	0	0	5
geological formation	200	150	73	87	50	52	148
land	15	15	2	3	0	2	13
plain	12	12	0	0	0	3	9
rangeland	8	8	1	4	0	0	8
region	46	44	6	0	2	2	44
relation	54	54	8	32	0	0	54
wetland	8	8	3	1	0	0	8
abandoned facility	16	15	4	1	1	1	15
body of water	116	106	24	17	10	3	113
populated place	13	10	2	1	3	2	11
seat of government	6	4	0	1	2	2	4
Total number of objects	**985**	**905**	**188**	**243**	**80**	**79**	**906**

Language Validators' evaluation of the lexical gaps revealed that the translators proposed 10 false positives out of 80. We also identified that the translators produced 9 false positive translations of the concepts whereas they are gaps. In the end, we found that there are in total 79 gaps and 906 concept translations being accepted. The *UKC Language validator* and *UKC validator* reported a few (around 5) conflicts, which were then solved with little effort. It is worth mentioning that *Language Translators* proposed to add 7 new concepts to the space ontology. This is only initial work and we expect that a few more concepts will be added with the evolution of the space ontology.

VIII. LESSONS LEARNED

Assigning word sense rank appears as a difficult task to accomplish since the *Language Translators* provide their the results indipendently. In the translation work, they were aware of the fact that concepts translated by others might have the same word label. But it remained obscure until the whole translation task was finished. This ranking could be defined once all the concepts are translated. This is a non-trivial task to accomplish because deciding acceptable ranks might require local community agreement or the consultation of high quality linguistic resources that are often insufficient for domain specific tasks in many languages.

Synonymous words within the synsets were often increased after translations were evaluated by the Language Validators. This was the case since Language Translators concentrate in providing the target language correspondence representation of the knowledge objects taken from the source language within a reasonable amount of time. This often results in the postponement of the addition of synsets.

In the cases where an example sentence in a gloss contains a number that has to be converted according to some suitable measurement, we should freely change values and corresponding units since the numbers always give some extra information to provide glosses. For instance, 6000 meters can be changed to 6 km (while value remains same) and 3 kilograms to 3 pounds (while value modifies). Nevertheless, in case of sensitive information found in a gloss, we should exactly convert the number to relevant measurement unit in order to preserve the meaning of the gloss. For example, for understandable measuring unit of the target users 500 feet can be converted into 152.4 meters.

Parts of the glosses that follow the same syntactic pattern in the source language can be translated with little effort. For instance, the gloss part *a facility for [verb]+ing [object]* appeared in around one tenth of the concepts. We repeated the same translation for the part that matched completely. Moreover, we used the translation memory technique that provides a translation with recurrent structure in the same way as previous translations.

In order to introduce foreign cultures to the community, we can translate lexical gaps as free combination of words. However, this should not always be the case. A first reason is computational: the explicit marking of the lexical gaps could support the KB-based applications in reducing computation time by avoiding the management of (multi)words that will be very rarely or never used. A second, more important reason, is related to the actual existence of a free combination of words capable of capturing, in the mind of a native speaker with no knowledge of the original concept (as it exists in the foreign culture) what the concept actually means, in the real world.

IX. RELATED WORK

MultiWordNet [10] consists of several European language WordNets. It was developed under a model that reuses semantic relations from WordNet as follows: when there are two synsets and a relation holding between them, the same relation is assumed to hold between corresponding synsets in the new language. There is no literal translation in the case of developing Italian version of MultiWordNet of the synsets, words and exceptional forms but the contributors have produced the best possible Italian equivalents according to their skills and experiences in knowledge organization and linguistics. However, a limited number of glosses has been provided, e.g., around 2k in Italian over 33k.

The ontology localization activity described in [11] is an attempt to address the localization and diversity issues. They proposed guidelines and methodologies for enriching ontology with multilingual information. However, we differ from them with respect to the target language and the development approach.

Universal Multilingual Knowledge Base also known as UWN [12] was developed leveraging on the Wikipedia data and linking multilingual terms that are connected to the same page. However, automatically built KB resources often suffer from quality issues, e.g., around 10% of the terms in UWN are attached to the wrong senses, whereas we achieved human-level accuracy.

FinnWordNet [5] was produced from WordNet with the help of professional translators and the output is monitored by bulk validation. While producing the whole WordNet in Finish in 100 days, they traded off the quality for reducing the amount of translation time. Diversity in the languages such as lexical gaps is overlooked in this task.

Concerning provenance modelling and representation, the PROV-O ontology [15] was developed to be used to trace resources belonging to any domain. Despite its richness and well-coverage in terms of classes and relations, it could fulfill our need only partially.

X. Conclusion

In this paper, we proposed an experiment for generating ontologies through translation from one language into another. This experiment was developed to be applied independently of domain and language and to deal with the diversity across the languages. While translating the ontologies, we identified the various diversity features and their presence in a given target language by working together with the linguistic experts and/or native speakers living in the country where it is spoken. We evaluated the effectiveness of the methodology by performing a case study for translating the space ontology into Mongolian.

Thanks to the reuse of the ontological backbone structure, we achieved space ontology in Mongolian that is as high quality as the original one in English. Though manual approach is usually known to be time consuming, adopting this methodology in a crowdsourcing setting can help increase throughput and make this suitable for dealing with large ontologies. We also have presented a provenance model for ontology localization tasks to keep track of the translators and validators the fact that it helps increase the reliability of each single object of the knowledge base. The generated ontology can be exploited to improve the accuracy of *NLP tasks* [13] and *Concept Search* [14] in space domain. The Mongolian version of space domain is currently in use in the School of Information Technology of the National University of Mongolia as background knowledge in their NLP pipeline.

Acknowledgment

The research leading to these results has received funding (partially) from the European Community's Seventh Framework Program (FP7/2007-2013) under grant agreement n. 600854 Smart Society: hybrid and diversity-aware collective adaptive systems: where people meet machines to build smarter societies http://www.smart-society-project.eu/. The experiment described in this paper is part of a long term project, whose goal is to create the Universal Knowledge Core (UKC), a worldwide multilingual linguistic and ontological resource. The work on provenance is part of the PhD thesis of the third author. We are thankful to Vincenzo Maltese for his valuable feedback.

References

[1] A. Ganbold, F. Farazi, and F. Giunchiglia. “An Experiment in Managing Language Diversity Across Cultures,” in eKNOW 2014, pp. 51–57.

[2] T. Berners-Lee, J. Hendler, and O. Lassila, “The Semantic Web,” Scientific American, vol. 284, no. 5, 2001, pp. 34–43.

[3] F. Giunchiglia, V. Maltese, F. Farazi, and D. Biswanath, “GeoWordNet: a resource for geo-spatial applications,” in ESWC’10, Volume Part I, 2010, no. December 2009, pp. 121–136.

[4] F. Giunchiglia, V. Maltese, and D. Biswanath, “Domains and context: first steps towards managing diversity in knowledge,” Web Semantics: Science, Services and Agents on the World Wide Web, vol. 12–13, 2012, pp. 53–63.

[5] K. Lindén and L. Carlson, “FinnWordNet – Finnish WordNet by Translation,” LexicoNordica – Nordic Journal of Lexicography, vol. 17, 2010, pp. 119–140.

[6] F. Giunchiglia, B. Dutta, V. Maltese, and F. Farazi, “A facet-based methodology for the construction of a large-scale geospatial ontology,” Journal on Data Semantics, vol. 1, no. 1, 2012, pp. 57–73.

[7] F. Giunchiglia, B. Dutta, and V. Maltese, “Faceted Lightweight Ontologies,” in in Conceptual Modeling Foundations and Applications, vol. 5600, 2009, pp. 36–51.

[8] F. Giunchiglia, B. Dutta, and V. Maltese, “From Knowledge Organization to Knowledge Representation,” in ISKO UK Conference, 2013, pp. 44–56.

[9] M. C. Suárez-Figueroa and A. Gómez-Pérez, “First Attempt towards a Standard Glossary of Ontology Engineering Terminology,” in TKE08, 2008, pp. 1–15.

[10] L. Bentivogli and E. Pianta, “Looking for lexical gaps,” in EURALEX International Congress, 2000, pp. 663-669.

[11] M. Espinoza, E. Montiel-Ponsoda, and A. Gómez-Pérez, “Ontology localization,” in K-CAP ’09, 2009, pp. 33-40.

[12] G. De Melo and G. Weikum, “Towards Universal Multilingual Knowledge Bases,” in Principles, construction, and applications of multilingual wordnets : proceedings of the Fifth Global WordNet Conference, 2010, pp. 149–156.

[13] I. Zaihrayeu, L. Sun, F. Giunchiglia, W. Pan, Q. Ju, M. Chi, and X. Huang, “From Web Directories to Ontologies : Natural Language Processing Challenges,” in ISWC’07/ASWC'07, 2007, no. 60673038, pp. 623–636.

[14] F. Giunchiglia, U. Kharkevich, and I. Zaihrayeu, “Concept search,” The Semantic Web Research and Applications, vol. 5554/2009, 2009, pp. 429–444.

[15] T. Lebo, S. Sahoo, D. McGuinness, K. Belhajjame, J. Cheney, D. Corsar, D. Garijo, S. Soiland-Reyes, S. Zednik, and J. Zhao. PROV-O: The PROV Ontology, 2013, W3C Recommendation. URL: http://www.w3.org/TR/2013/REC-prov-o-20130430/.