LDK 2023

Refinement of the Classification of Translations – Extension of the *vartrans* Module in OntoLex-Lemon

Frances Gillis-Webber
Computer Science Department, University of Cape Town, South Africa
fgilliswebber@cs.uct.ac.za

Abstract

In the *vartrans* module for OntoLex-Lemon, there are three categories from Translation Category Reference RDF Schema (TRCAT) used to classify translations. Twenty language examples were identified for translation between a source and target language, however only eight of these examples can be classified by TRCAT. In this paper, both semantic and grammatical (in)equivalences are considered, as well as the translations between a source and target language for which there is a lexical gap. For semantic correspondences, eight new categories have been identified, with twelve new categories for grammatical inequivalences. The *vartrans* module was then extended to include these new categories, soft-reusing two of the categories from TRCAT, with classes and object properties added for grammar rules and language features. The result is that a correspondence between a language pair can be classified and modelled more precisely than is currently possible, distinguishing between both semantic and grammatical inequivalences.

1 Introduction

In the *vartrans* module for OntoLex-Lemon, a translation between a source and a target lexical sense is classified by its category, using categories from Translation Category Reference RDF Schema (TRCAT) (Cimiano et al., 2016). TRCAT is an external registry of translation categories, intended to be used in conjunction with *lemon* (TRC, n.d.; Gracia et al., 2014). Three categories are provided for: *directEquivalent*, *lexicalEquivalent*, and *culturalEquivalent*. The *directEquivalent* category classifies the translation between two senses as semantically equivalent, and the *lexicalEquivalent* category is used when the target lexical sense is a direct translation of the source sense. The *culturalEquivalent* category is used to indicate the target translation as culturally similar to that of the source. Although each of these categories pertain to equivalences, *lexicalEquivalent* can also classify the translation between two senses as inequivalent, where a metaphrase of a source term can be indicative of a lexical gap.

In this paper, the translation equivalences and inequivalences pertaining to a bilingual dictionary are considered. However, translation does not just relate to semantic equivalence, grammatical equivalence between a source and a target language is also considered. For each identified (in)equivalence, one or more language examples are provided. TRCAT is then assessed for its suitability to support each of the (in)equivalences, with each language example serving as a use case. An extension to the *vartrans* module is then proposed, with a series of questions given to guide the user in selecting the ideal category. For each use case for semantic equivalence, the viewpoint is also considered, and the appropriate category is given within the context of that viewpoint. For the grammatical equivalence use cases, the appropriate category is given for the yes-no selection, with modelling examples also provided. The result is that the equivalence relations between a source and target language for a lexical entry/sense can be modelled more precisely than is currently possible with the *vartrans* module.

The remainder of the paper is structured as follows. In Sections 2 and 3, semantic and grammatical alignments are discussed respectively. The *vartrans* module extension is presented in Section 4, using each of the language examples from the preceding sections. Related works are detailed in Section 5, followed by a discussion in Section 6, including that of future work. The paper concludes with Section 7.

2 Semantic Alignments

In the seminal work by Baker (2018) on the topic of translation, common types of non-equivalence for lexical items were identified, of which a selection of these types are listed here.

1. Concepts that are specific to a culture.
2. A concept in a source language is not lexicalised in a target language.
3. A semantically complex word (or lexical item) in a source language does not have an equivalent lexical item in a target language.
4. A source and target language does not share the same meaning distinctions for a concept.

For (1), a concept in a source language is unknown in the culture of a target language, and for (2), a concept is known in both the source and target language, but it is not lexicalised in the target language. Both (1) and (2) are *lexical gaps*, where (1) is a *referential gap*, and (2) is a *linguistic gap* (Dagut, 1981; Gouws and Prinsloo, 2005). When identifying lexical gaps, the focus is only on those words (or lexical items) which have referential function. The reference can be concrete (for example, 'house', 'sun'), abstract ('love', 'excitement'), or purported ('unicorn', 'hell') (Dagut, 1981). Examples for (1) and (2) respectively are the isiXhosa concepts of 'hlonipha' and 'lobola'. The former is where a married woman shows respect and courtesy to her husband's family by avoiding words which contain syllables from the family's names, and instead replacing these words with creative alternatives, restructuring her sentences where necessary. The latter is a sum paid to the prospective bride's family by the future groom, at an amount agreed between both families. 'Bride price' is often given as a translation equivalent but it implies the sale of a person, and fails to capture the 'lobola' practice as a union of the two families, where originally it was paid in cows that had been accumulated by the groom's father over a period of time. Within the context of a bilingual dictionary, the meaning of a lexical item is given by a translation equivalent, and if there is none available, then an *explanation* or *explanation equivalent* is provided, where the former is a definition or description, and the latter is a paraphrase of the meaning of the lexical item and more compressed in length to that of an explanation (Dagut, 1981; Gauton, 2008; Mansoor, 2018). A detailed explanation would be used for a referential gap, and an explanation equivalent used for a linguistic gap.

Point (3) is similar to (2), where a concept is known in both the source and target language, but the source language has identified a short-hand term to represent a complex concept. An example is the English term 'adoption', the legal process where the biological parent of a child is changed to the adoptive parent or parents. The Sesotho equivalent is a paraphrase, 'ho fuwa ngwana ka molao', which has the English gloss of 'giving a child legally' (Gen, 2017). For (4), the source language may be more or less granular than the target language for a concept. An example often used in the literature is the concept of 'river' and its French equivalents: 'rivière' and 'fleuve'. The isiXhosa kinship term 'umzukulwana' is an example where it is less specific than English, with the same term used for 'granddaughter', 'grandson', and 'grandchild'.

Table 1 lists the language examples specific to semantic equivalence. The alignment is indicated in the 'Alignment' column, where a language code is used to identify the source and target languages. The concept of 'hlonipha' as a referential gap in English is UC1. Distinction is made between the concepts of 'lobola' and 'bride price', each given in UC2–5. 'Lobola' is a loanword in South African English with no morphemic modification (UC2), but a linguistic gap in US/British English (UC3). UC4 is the alignment of 'lobola' to 'bride price', where the concept of 'lobola' is more granular (or specific) to that of 'bride price'. In UC5, the alignment is between English and South African English. Within the context of South Africa, the 'lobola' borrowing would be used by South African English speakers. However, for the concept of 'dowry', this would remain unchanged in South African English. In UC6, the direct translation of 'dowry' is given for isiXhosa, although there is also a meaning distinction.

In UC5, UC9, and UC12, the alignment is shown between a language and its dialect. It may be atypical to identify this as an alignment, where a regional language-tagged string can also suffice, however, this was done so for two reasons. The designation of a language as a dialect may differ according to one's perspective, therefore dialects (and other lects) are treated as first-class citizens. Secondly, there is not necessarily full mutual intelligibility between a language and its dialects (with the dialects of Chinese being one suchß example).

The concept of 'loadshedding' (same as 'rolling blackouts', where electricity is rationed) features heavily in South Africa's lexicon (UC9). Although

Table 1: Language examples for semantic (in)equivalences. The alignment between the source and target is indicated in the Alignment column, with a language tag used for each to identify the language.

Source	Alignment	Target		
hlonipha	xh → en		**UC1**	Culture-bound term. Referential gap in English, including South African English.
lobola	xh → en-za	lobola	**UC2**	Loanword in South African English, with no morphemic modification.
lobola	xh → en		**UC3**	Linguistic gap in US/British English.
lobola	xh → en	bride price	**UC4**	Not exact meaning, isiXhosa is more granular.
bride price	en → en-za	lobola	**UC5**	Borrowing is used in South African English.
dowry	en → xh	ikhazi	**UC6**	Concept of 'dowry' from an AmaXhosa perspective has a different meaning.
adoption	en → st	ho fuwa ng-wana ka molao	**UC7**	Paraphrase as no equivalent term exists.
umzukulwana	xh → en	granddaughter grandson grandchild	**UC8**	Granularity mismatch where English is more specific.
loadshedding	en-za → en	loadshedding	**UC9**	Common term in South Africa's lexicon. Not widely used elsewhere.
loadshedding	en-za → xh	loadshedding	**UC10**	Loanword from South African English with no morphemic modification.
loadshedding	xh → st	loadshedding	**UC11**	Loanword from South African English.
traffic light	en → en-za	robot	**UC12**	A different term is used for the same concept in South Africa.
electricity	en → xh	igesi	**UC13**	The term '-gesi', a loanword with morphemic modification from the English term 'gas', has since been extended to include the concept of 'electricity'.
spoon	en → af	lepel	**UC14**	The meaning is the same, except that neither share the same hypernym.

the concept has long been lexicalised in English, the term is not widely known, unless of course, a person lives in an area where rolling blackouts occur. In the case of 'loadshedding' in South African English, the term has been borrowed by the other local languages, currently with no morphemic modification (UC10–11). For UC12, a traffic light is known as a robot in South African English.

In UC13, an example is given where an existing term is extended to include a new concept from another language, shown here for the direct equivalent 'electricity' to isiXhosa's 'igesi'. isiXhosa is an agglutinative language with a noun class system and concordial agreement. The term 'ugesi' is used for 'power' and 'gas', where the stem '-gesi', originally the loanword 'gas' from English with morphemic substitution, has since extended to include 'electricity'. Lastly, for UC14, this is an example where the term refers to the same object, but each language classifies it differently. In English, 'spoon' is a 'utensil', and in Afrikaans, it is a 'tool'.

We now revisit the translation categories from TRCAT, and systematically try to classify each use case. As shown in Table 2, only 8 of the 14 use cases can be classified by TRCAT's categories. Using the semiotic triangle, the possible equivalences between a source and target language are given in Figure 1. For *directEquivalent* to be applicable, there has to be a lexical realisation for both the source and the target, and both lexical realisations have to be semantically equivalent. This is visualised in Diagram I in Figure 1. There are no categories in TRCAT to classify linguistic (Diagram II–IV) and referential gaps (Diagram VI), as well as partial equivalence (Diagram V).

3 Grammatical Alignments

As mentioned previously, isiXhosa is an agglutinative language with concordial agreement, so the prefix of a noun changes if it is singular or plural, as well as the prefixes or pre-prefixes

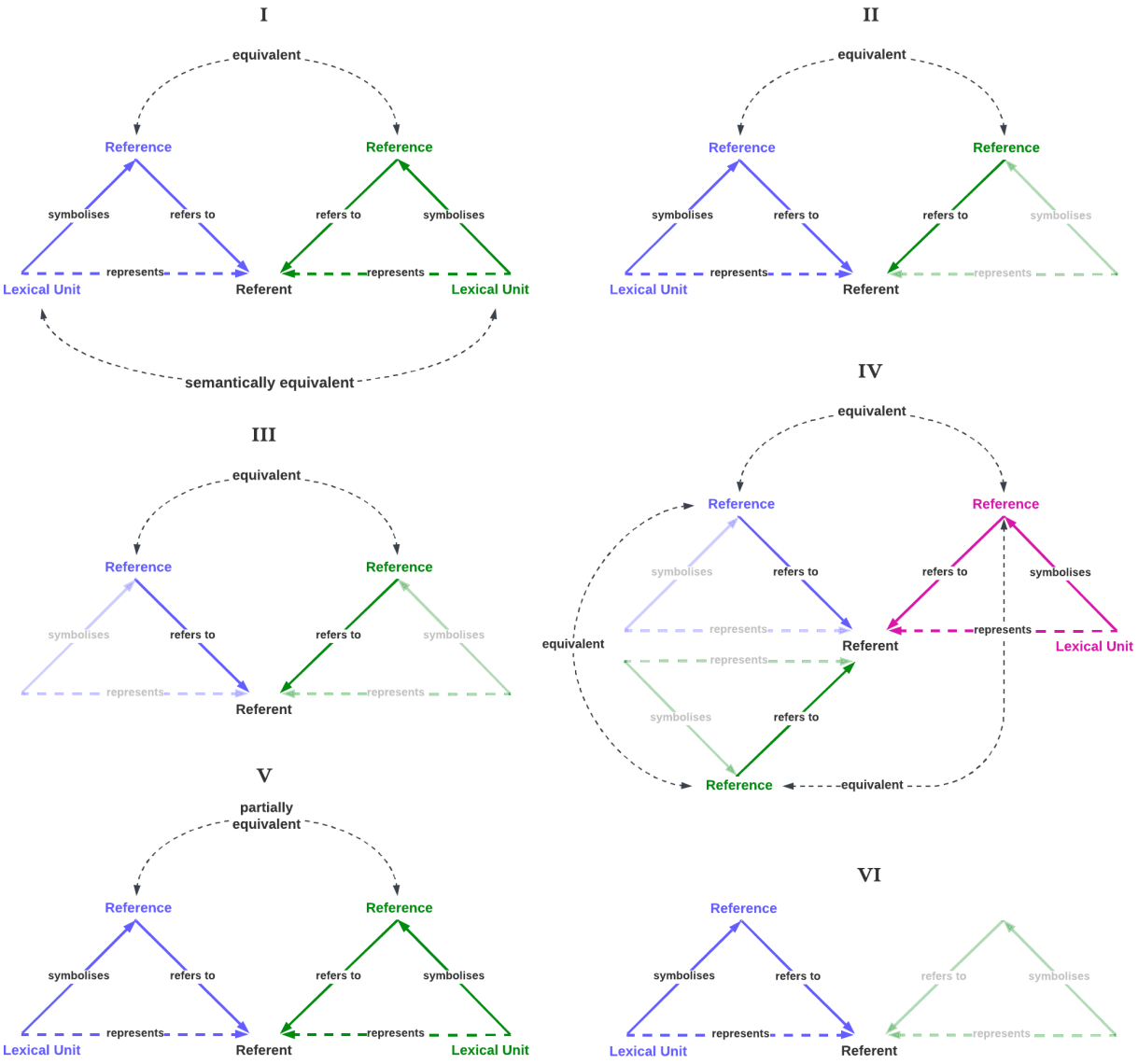

Figure 1: The semiotic triangle is used to show equivalence between two languages for a term. Language A is in purple and Language B is in green. Diagram I shows the source and target lexical units as semantically equivalent. Diagram II shows a lexical gap for the target (indicated as such by the opaque part of its semiotic triangle), however, the concept is known, so this is a linguistic gap. Diagram III shows a linguistic gap for both the source and the target. In Diagram IV, Diagram III is extended by introducing a pivot language (Language C, shown in pink). Diagram V shows partial equivalence between two references, with the result that there is not full semantic equivalence between the source and target lexical units. A referential gap for the target language is shown in Diagram VI.

Table 2: A comparison of each of the use cases for semantic equivalence against the available categories in TRCAT.

Use Case	Direct Equiv.	Lexical Equiv.	Cultural Equiv.
UC1			
UC2	✓		
UC3			
UC4			✓
UC5			✓
UC6			
UC7			
UC8			
UC9	✓		
UC10	✓		
UC11	✓		
UC12	✓		
UC13	✓		
UC14			

changing to show agreement with other parts of the sentence. As an example, the stem '-zimba' means 'body'. If the prefix 'um' is added, then 'umzimba' is singular, and if the prefix is 'imi', then it is plural. To denote modifications to the noun, such as the diminutive or feminine, then a suffix is also added. isiXhosa dictionaries are not consistent in their lemmatisation approach. For example, in The Greater Dictionary of isiXhosa, Volumes 1–3, nouns and verbs are listed by their stem (Tshabe, 2006; Mini, 2003; Pahl, 1989). In the Oxford Xhosa-English Dictionary (De Schryver and Reynolds, 2019), nouns are listed by their singular form and verbs are listed by their stem. In the Pharos English-Xhosa Dictionary, nouns and verbs are listed by their stem, although the form of the lemma for verbs does not make this obvious (Eng, 2014). When aligning two lexical senses from different languages, if an alignment is between, for example, word and stem or word and singular form, then this should be made clear. Use cases 15–16 pertain to this, given in Table 3.

Still staying with isiXhosa, using the 'subtraction' mathematical operator as an example, the stem is '-thabatha'. It is a verb by default, and to say 'to subtract' in a sentence, the prefix 'u' is used. To refer to subtraction as a noun, the prefix 'uku' is added to the stem. UC17 relates to a part-of-speech change, which occurs here if the alignment is from word to stem. UC18–19 pertains to grammatical gender. In isiXhosa, 'umfundisi' is the word for 'priest' in English. However, this is a male priest, and to refer to a female priest, the suffix 'kazi' is added. Similarly in Spanish, the label for an object property 'changed by' can be 'es modificada por' or 'es modificado por'. The change is attributed to grammatical gender, where the gender of the noun used for the class of the object property's domain determines the gender of the past participle.

Lastly, we consider alignment between a mass and count noun. In English, the word 'seed' is both a mass noun and a count noun, however we focus just on the count noun. An example sentence is "Mark planted bean seeds." In isiXhosa, the singular is 'imbewu', and this is used, even when the plural is referred to in English (UC20) (De Schryver and Reynolds, 2019).

4 The *vartrans* Module Extension

In OntoLex-Lemon, an ontology entity is used as the definiens for a lexical sense or a lexical entry. An ontology entity is in turn comprised of a semantic layer and a linguistic layer, visualised in Figure 2, where it can either be a class or an individual. As none of the use cases require lexical equivalency to be established between, say "Bill Gates"@en and "uBill Gates"@xh, both individuals of the class `:Person`, the focus is only on the use of an ontology class and its ontological commitment as a definiens.

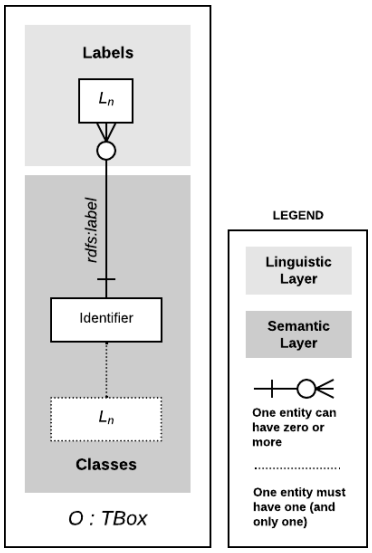

Figure 2: Distinguishing between the semantic and linguistic layers in the TBox of an OWL ontology.

Table 3: Language examples for grammatical inequivalences.

Source	Alignment	Target		
body	en → xh	umzimba	**UC15**	Singular noun in English aligned to singular form of noun stem in isiXhosa.
body	en → xh	-zimba	**UC16**	Singular noun in English aligned to noun stem in isiXhosa.
minus	en → xh	-thabatha	**UC17**	Noun in English aligned to verb stem in isiXhosa.
priest	en → xh	umfundisi / umfundisikazi	**UC18**	The isiXhosa singular form refers only to male priests. With the addition of the suffix '-kazi', the singular form now refers to a female priest.
changed by	en → es	es modificado por / es modificada por	**UC19**	The gender changes for the Spanish past participle according to the gender of the subject.
seeds	en → xh	imbewu	**UC20**	The plural is used in English, however the singular is used in isiXhosa.

An ontology entity in OWL is comprised of two parts in the semantic layer: the axiom pattern, and the superclass of the axiom pattern, as well as the individuals of the axiom pattern, each shown in Figure 3. The axiom pattern comprises one or more classes and any axioms which serve as an ontological commitment. If we let O, O' be two ontologies with vocabularies V, V', two *homogenous* ontology entities, with one entity in V and the other in V', can be aligned using an alignment axiom (Euzenat and Schvaiko, 2013). The axiom pattern, superclass(es), and individuals of the ontology entity in V and V' respectively can each be compared to determine the extent of equivalence in order to assign the appropriate category to the alignment. For the axiom pattern between O and O', the axioms may differ, be it subclasses, a differing object property, or restrictions on the domain and range. For the superclasses, an axiom pattern in O may be placed differently in the class hierarchy to that of its counterpart in O'. For the individuals, only a subset of individuals may be applicable in O', when compared to O.

Using the concept of 'River', example axiom patterns in Description Logic are given for the definiens of English's River (1), Afrikaans' Rivier (2), and French's Fleuve (3) and Riviere (4–5):

$$\exists \mathsf{flowsInto}.\mathsf{NaturalWatercourse} \sqcap \neg\exists \mathsf{flowsInto}.\mathit{Self} \quad (1)$$
$$\exists \mathsf{inVloei}.\mathsf{NatuurlikeWaterloop} \sqcap \neg\exists \mathsf{inVloei}.\mathit{Self} \quad (2)$$
$$\exists \mathsf{couleDans}.\mathsf{CoursDeauNaturel} \sqcap \exists \mathsf{couleDans}.\mathsf{Mer} \quad (3)$$
$$\exists \mathsf{couleDans}.\mathsf{CoursDeauNaturel} \sqcap \exists \mathsf{couleDans}.\mathit{Self}, \quad (4)$$
$$\mathsf{Riviere} \sqsubseteq \neg \mathsf{Fleuve} \quad (5)$$

If the language pair is English and Afrikaans, then River and Rivier is semantically equivalent, with the same individuals as well. If the language pair is English's River to French's Fleuve, the axiom pattern is not equivalent, and only a subset of the individuals apply to Fleuve.

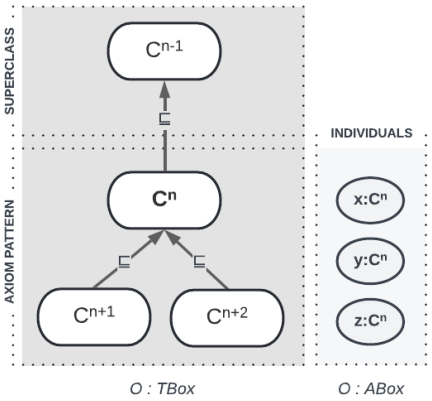

Figure 3: The 'parts' of an ontology entity in an OWL ontology. The axiom pattern and its superclasses are in the TBox. C^n is the starting point of the axiom pattern, and C^{n-1} is its immediate parent. The individuals are an assertion of class C^n.

To determine semantic equivalence, the following questions are identified.

Q1: Is there a lexical realisation for the source and the target concepts?

Q2: Are the individuals the same for both the source and the target?

Q3: Is there some overlap of the individuals between the source and the target?

Q4: Are the individuals of the target a subset of the source (or vice versa)?

Q5: Is the axiom pattern the same for both the source and the target?

Q6: Is the superclass(es) the same for both the source and the target?

Q7: Is there a lexical realisation for either the

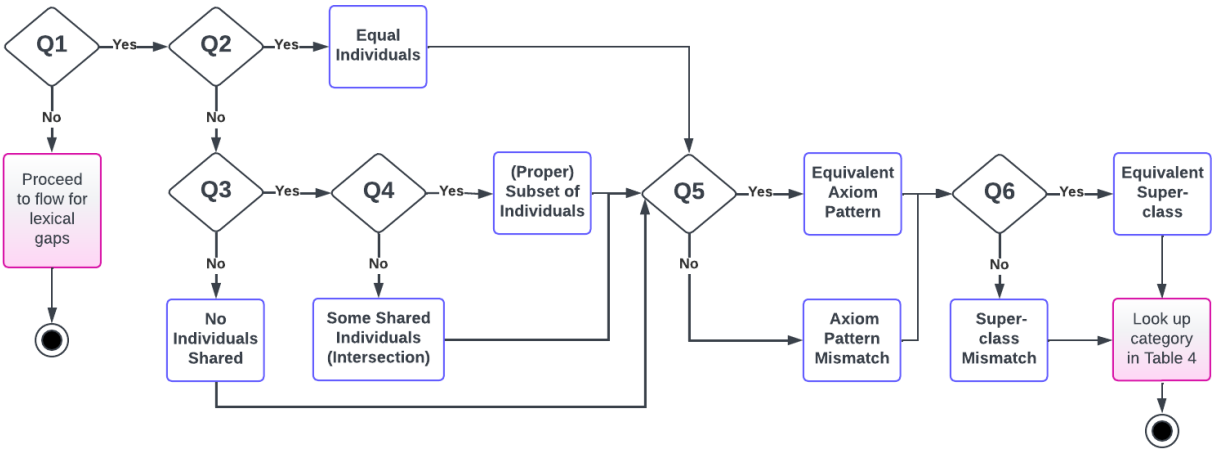

Figure 4: The decision tree diagram for Q1–6, for those alignments where there is a lexical realisation for both the source and the target. The diamond symbol denotes a decision that has to be made, where there is a 'yes' or 'no' answer. Each of the questions from Q1–6 are posed as decisions, and the starting point is Q1. The purple block indicates the feature that applies, based on the previous yes-no answers, and the small circles show the end of the flow for that question-answer selection.

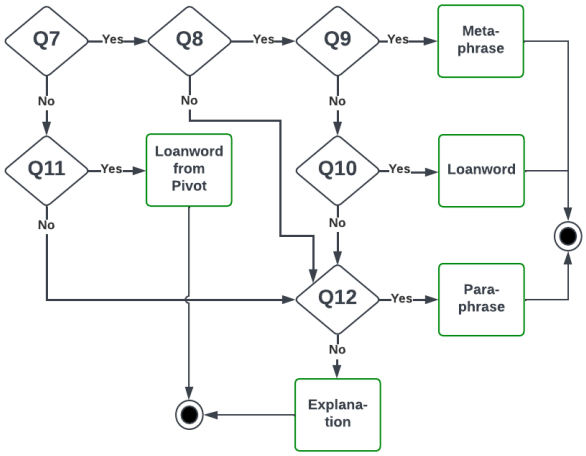

Figure 5: The decision tree diagram for Q7–12, for those alignments where is no lexical realisation for the source and/or the target. Each green block is the proposed category to use for that question-answer selection.

source or the target?

Q8: For the source or target which has no lexical realisation, is the concept known in the language?

Q9: For the target which has no lexical realisation, can the source be directly translated as a metaphrase?

Q10: For the target which has no lexical realisation, can the source be used as a borrowing (and vice versa)?

Q11: Can a third language be introduced to serve as a borrowing between the source and the target?

Q12: If there is a referential gap or no borrowing can be used, can a paraphrase be used instead?

If both source and target is lexicalised, then Q1–6 applies, with the question flow shown in Figure 4. If neither source nor target is lexicalised, then Q7–12 applies. The question flow is given in Figure 5. The label in each purple block in Figure 4 indicates the applicable feature. The features can then be looked up in Table 4 to determine the correct category to use. In Figure 5, each green block indicates the applicable category for the yes-no answer selection to Q7–12.

In Table 4, reference is made to an 'interpretation' where a correspondence between a source and target language can be equivalent in some interpretation. One of the internationalisation goals of OWL was to "potentially provide different views of ontologies that are appropriate for different cultures" (W3C OWL Working Group, 2004). If we consider ontology A which has a 'universal' viewpoint, then this ontology has, theoretically-speaking, all possible individuals for the interpretation $\mathcal{I}$. However, we can modify $\mathcal{I}$ to obtain another interpretation $\mathcal{I}_{xh}$, which is specific to the speakers of one natural language, say isiXhosa, where individuals not applicable to isiXhosa speakers are removed, and the interpretation of class names and names of object properties are also changed so that they are specific to the isiXhosa viewpoint or perspective. The result is that the individuals of $\mathcal{I}_{xh}$ is a subset of the individuals of $\mathcal{I}$ (i.e., a proper subset in set theory).

The extended *vartrans* module (*extvartrans*) is located at: `https://w3id.org/EXTVARTRANS`. A new object property, `#semanticCategory` was

created as a subproperty of `#category` in *extvartrans*. The domain is a 'lexico-semantic relation' from *vartrans*, and its range has been set to one class: `#SemanticCorrespondence`. The subclasses of `#SemanticCorrespondence` are shown in Figure 6.

Figure 6: The new categories for semantic correspondences in the *extvartrans* module.

If the individuals are equal and the axiom pattern and superclass is equivalent between a source and a target, then this is a 'Direct Equivalent', and the category from the *vartrans* module is used. If the individuals are equal but either the axiom pattern or superclass (or both) are not equivalent between a source and a target, then this is an 'Indirect Equivalent'. If the axiom pattern and superclass is equivalent, but the individuals are not equal but are instead a proper subset[1], then this is a 'Direct Equivalent in Some Interpretation' (but not all). For 'Overlapping Meaning', only some individuals are shared (instead of being a subset), and the axiom pattern and superclass can be a mismatch or equivalent between a source and a target. Finally, if there are no shared individuals between a source and a target, then despite the axiom pattern and/or superclass being equivalent, there is no correspondence.

4.1 Solving for the semantic use cases

Before each of the use cases are reviewed, we first identify the viewpoints by which a use case is considered (using the source and target language codes in the 'Alignment' column in Table 1 as a guide).

VP1: first language speakers of isiXhosa
VP2: language speakers of all English variations
VP3: speakers of South African English
VP4: speakers of English spoken in USA/UK

[1]For 'subset' to apply, a subset of A can also be equivalent to A. For 'proper subset' to apply, a subset of A is not equivalent to A.

Table 4: A lookup table to determine the appropriate category to use, according to each of the 'parts' of an ontology entity: axiom pattern, superclass, and set of individuals, where the selection for each is an outcome of the yes-no answers selected in the decision tree diagram of Figure 4. These categories pertain to concepts where this is a lexical realisation for both the source and the target.

Axiom Pattern	Superclass	Individuals	Category
Equivalent	Equivalent	Equal	Direct Equivalent
Equivalent	Equivalent	Proper Subset	Direct Equivalent in Some Interpretation
Equivalent	Equivalent	Intersection	Overlapping Meaning
Equivalent	Equivalent	None	*No correspondence* in Some Interpretation
Equivalent	Mismatch	Equal	Indirect Equivalent
Equivalent	Mismatch	Proper Subset	Granularity Mismatch
Equivalent	Mismatch	Intersection	Overlapping Meaning
Equivalent	Mismatch	None	*No correspondence*
Mismatch	Equivalent	Equal	Indirect Equivalent
Mismatch	Equivalent	Proper Subset	Granularity Mismatch
Mismatch	Equivalent	Intersection	Overlapping Meaning
Mismatch	Equivalent	None	*No correspondence*
Mismatch	Mismatch	Equal	Indirect Equivalent
Mismatch	Mismatch	Proper Subset	Granularity Mismatch
Mismatch	Mismatch	Intersection	Overlapping Meaning
Mismatch	Mismatch	None	*No correspondence*

VP5: first language speakers of Sesotho
VP6: first language speakers of Afrikaans
VP7: language-independent

UC1 can be considered from three viewpoints: VP1, VP2, and VP7. For VP1, as there is a referential gap in English, a translation is required. If the flow diagram in Figure 5 is followed, then the proposed category is `#ExplanationAsTranslation`, where the axiom pattern and superclass(es) from the source are applied to the target as well. For VP2, one can argue that as it is a referential gap, the source concept can be excluded as it does not pertain to English culture. For VP7, the same as that for VP1 can be done, except with an additional axiom to indicate that this custom pertains only to

AmaXhosa culture.

For UC2, VP3 applies. As the concept is well-known in South African speakers' lexicon, and it is unchanged from that of isiXhosa except for an additional axiom to indicate that it pertains to AmaXhosa culture, the proposed category is `#IndirectEquivalent`. For UC3, VP4 applies. There are two possibilities for this use case: ignore the concept on the basis that it has no relevance within US/UK English culture; alternatively, model the alignment as a subclass of 'bride-price' (as 'lobola' is a more granular notion), with an axiom to indicate that it pertains to AmaXhosa culture. For the latter, the `#ParaphraseAsTranslation` is suitable. For UC4, the proposed category is `#GranularityMismatch`, on the basis that the axiom patterns for the source and target concepts are not the same, the superclass is the same, and the source individuals are a subset of the target individuals. For UC5, VP3 applies. For this use case, the proposed category is `#IndirectEquivalent`, on the basis that although the axiom pattern is a mismatch, the superclass is the same, and the individuals are the same (as neither concept is being considered from the perspective of the AmaXhosa). For UC6, two viewpoints can be considered: VP1 and VP2. If the alignment is considered from VP1, then this is a `#GranularityMismatch` as the target concept is more precise than the source, and it only applies to a subset of individuals. If VP2 is considered, then the `#IndirectEquivalent` category applies, and the term 'ikhazi' can be used interchangeably.

For UC7, the Sesotho paraphrase will differ from one dictionary to another. The proposal here is to treat it as a lexical gap and use the `#ParaphraseAsTranslation` category to indicate as such. For UC8, the category is `#GranularityMismatch`. If each target term is considered individually, then there is an axiom pattern mismatch with the source, as well as the individuals being a subset (where 'granddaughter' refers to female grandchildren, but 'umzukulwana' refers to both female and male grandchildren).

For UC9–11, the category is `#directEquivalent`. For UC9, the axiom pattern and superclass is the same for the source and the target, as well as the individuals. An additional synonym can be provided for the target of UC9: 'rolling blackout'. For UC10 and UC11, VP1 and VP5 applies respectively. As there is no morphemic modification for both the targets, it is assumed that the meaning is unchanged from English.

UC12 is a `#directEquivalent`. If UC13 is considered from VP1 and VP2, then the proposed category is `#GranularityMismatch`. Lastly, for UC14, the `#IndirectEquivalent` category applies, as the superclass differs for each.

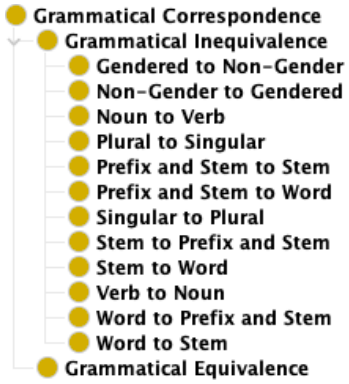

Figure 7: The new categories for grammatical correspondences in the *extvartrans* module.

4.2 Solving for the grammatical use cases

A new object property, `#grammarCategory` was created as another subproperty of `#category` in *vartrans*. Its range has been set to one class: `#GrammaticalCorrespondence`, and its subclasses are shown in Figure 7. The category `#GrammaticallyInequivalent` has subclasses, of which `#NounToPrefixAndStem` is the class selected for UC15, shown in Lines 6–7, in Listing 1. UC16 and UC20 are similarly classified, using the `#WordToStem`, and `#PluralToSingular` categories respectively. In each Turtle fragment that follows, the namespaces[2] are assumed defined.

```
1 :UC15 a vt:Translation ;
2   vt:source :sense_en_body ;
3   vt:target :sense_xh_umzimba ;
4   vt2:semanticCategory
5       trcat:directEquivalent ;
6   vt2:grammarCategory
7       vt2:WordToPrefixAndStem .
```

Listing 1: Turtle fragment for the translation of UC15.

[2] @prefix : <http://example.com#> .
@prefix vt: <http://www.w3.org/ns/lemon/vartrans#> .
@prefix vt2: <https://w3id.org/EXTVARTRANS#> .
@prefix trcat: <http://purl.org/net/translation-categories#> .
@prefix ontolex: <http://www.w3.org/ns/lemon/ontolex#> .
@prefix lexinfo:
<http://www.lexinfo.net/ontology/3.0/lexinfo#> .

For UC17, two categories are used, shown in Line 4 of Listing 2.

```
1 :UC17 a vt:Translation ;
2   ...
3   vt2:grammarCategory
4       vt2:WordToStem , vt2:NounToVerb .
```

Listing 2: Turtle fragment for the categories of UC17.

For UC18, it can be said that the male and female form is a granularity mismatch to English, therefore it is a semantic inequivalence. However, it has been opted to treat this as a grammatical inequivalence rather. As a gendered suffix is not applied consistently to the part of speech of type 'noun' in isiXhosa, a grammar rule has been created specific to a lexical item, and this is used, along with a grammar inequivalence category. To do this, a new class was created: `#GrammarRule`, for which there are two subclasses: `#PartOfSpeechSpecificRule` and `#LexicalItemSpecificRule`. The class `#GenderModificationOfNoun` is a subclass of `#LexicalItemSpecificRule`. The category `#NonGenderToGendered` was used, with both shown in Lines 6–8 in Listing 3 respectively.

```
1  :UC18 a vt:Translation ;
2    vt:source :sense_en_priest ;
3    vt:target :sense_xh_umfundisa ;
4    vt2:semanticCategory
5       trcat:directEquivalent ;
6    vt2:grammarCategory
7       vt2:WordToPrefixAndStem ,
8       vt2:NonGenderToGendered ;
9    vt2:targetRule
10      :rule_xh_fem_kazi .
11
12 :rule_xh_fem_kazi a
13   vt2:GenderModificationOfNoun ;
14   vt2:addSuffix :xh_kazi .
15
16 :xh_kazi a lexinfo:Suffix ;
17   ontolex:canonicalForm :xh_kazi_lemma ;
18   lexinfo:gender lexinfo:feminine .
19
20 :sense_xh_umfundisa a
21   ontolex:LexicalSense;
22   ontolex:reference dbp:Priest ;
23   lexinfo:gender lexinfo:masculine .
```

Listing 3: Turtle fragment for UC18.

A new object property was created: `#targetRule`, and this was added to the translation, shown in Lines 9–10 of Listing 3. An instance of the `#GenderModificationOfNoun` rule is given in Lines 12–14. A new object property was created for this rule `#addSuffix`, where the range is a lexical entry of type 'Suffix'. The creation of the suffix is shown in Lines 16–18, where LexInfo is used.

UC19 also relates to gender, however it differs in that the translation pertains to an object property, which means the surface realisation of the label will change according to the noun of the class used as the domain. In this instance, the rule is not specific to a lexical item (as was the case of UC18), instead, it is a rule specific to a part of speech. A new rule was created as a subclass of `#PartOfSpeechSpecificRule`: `#GenderAgreement`, and this rule is set as the `#targetRule` for UC19.

```
1  :UC19 a vt:Translation ;
2    vt:source :lex_en_changed_by ;
3    vt2:targetMasculine
4       :lex_es_es_modificado_por ;
5    vt2:targetFeminine
6       :lex_es_es_modificada_por ;
7    vt2:semanticCategory
8       trcat:directEquivalent ;
9    vt2:grammarCategory
10      vt2:NonGenderToGendered ;
11   vt2:targetRule
12      :rule_es_rule_gender .
13
14 :rule_es_rule_gender a
15   vt2:GenderAgreement .
```

Listing 4: Turtle fragment for UC19.

5 Related Works

Ontologies pertaining to linguistics were reviewed in the Linked Open Vocabularies (LOV) repository[3], of which a selection are listed here. The General Ontology for Linguistic Description has a `#translation` object property with `#literalTranslation` as a subproperty (Gol, 2010). It has a class `#LexicalizedConcept`, but none for an unlexicalised concept. LexInfo also provides for a `#translation` object property (from *vartrans*), as well as lexical and sense relations (Cimiano et al., 2011), however these are more suited to same-language relations. The property `#geographicalVariant` can be used for dialects, and the properties `#exact`, `#approximate`, and `#quasiEquivalent` can be used for lexicalised translations, although when to use the latter two is not made clear. The Lingvoj Ontology provides for the representation of language resources, and it has a `#Translation` class as an event, although this is intended at resource-level, not at term-level (B. Vatant, n.d.). The Lexvo.org Ontology is intended for the description of natural languages, terms, and meanings (de Melo, 2015). It provides

[3] https://lov.linkeddata.es/dataset/lov

for the thesaurus hierarchy of `#broader` and `#narrower`, as well as `#somewhatSameAs` and `#nearlySameAs`, where the latter two are intended as an alternative to `owl:sameAs`, all as object properties. To the best of our knowledge, there is no ontology or registry which provides the same extent of categorisation as that presented in *extvartrans*, particularly for lexical gaps. Of the ontologies which do provide some descriptors, this is only as object properties, and not as classes.

6 Discussion & Future Work

The reference or denotation of a lexical entry or sense is, in OntoLex-Lemon, given by an ontology entity. This has come in for criticism, with Hirst (2014) being one such example, in that an ontology entity is not granular enough to accurately represent the meaning distinctions of a concept across several natural languages. Direct equivalence between terms of different languages is not always possible, and even more so for concepts which are culture-bound (Culler, 1976; Kramsch, 1998; Zgusta, 1971; Hirst, 2014). By specifying a `#Translation` from the *vartrans* module, this can aid in bridging a gap between a language pair. The *vartrans* module has defined these mappings between a language pair as a translation. If the ontology is multilingual but based on a primary language (where this is typically English), then all other language terms are indeed a translation. If UC1 had to be considered only from VP2, then it is unlikely that this concept would have been included in an ontology where English is the primary language. In a multilingual ontology, each natural language usually takes on the axioms of the primary language, to the exclusion of each additional language.

Of the three translation categories, there is soft-reuse of `#directEquivalent` and `#culturalEquivalent` only in *extvartrans*. The category `#lexicalEquivalent` was not included in *extvartrans* as its meaning (literal translation) is not consistent with the same term used in Lexicography (that of absolute equivalence (Zgusta, 1978)). The category `#MetaphraseAsTranslation` was created as an alternative.

The *extvartrans* module aims to get closer to realising one of the internationalisation goals of the OWL specification, and that is to develop different views of the same ontology, where each view is specific to a culture. Considered from this perspective, then the mapping between a language pair is not necessarily always a translation but it can also refer to a transformation. It is for this reason that the word 'Correspondence' was used in the *extvartrans* module, instead of the word 'Translation'. The exception to this is a mapping between a language pair where the target is a lexical gap. This mapping is indeed a translation of the lexicalised source (or pivot language source).

The first step towards ontology transformation has been presented with the grammatical use cases. Each Turtle fragment given for these use cases is intended to serve as an input to an algorithm. The use cases presented here were by no means exhaustive and it is expected that more subclasses will be added to `#GrammaticallyInequivalent` in the future. The ontology transformation process for language-specific views is current work, where the focus is primarily on semantic inequivalences. In this paper, the linguistic layer of the ontology (as shown in Figure 3) has been the focus. However, for future work, the focus will be on the semantic layer, with the addition of new axioms to an existing ontology, and the refactoring of classes and object properties so that the ontology is specific to a viewpoint. The ontology to represent viewpoints, the Model of Multiple Viewpoints (MULTI), is already available at `https://w3id.org/MULTI` (Gillis-Webber, 2023). The next step is to soft-reuse selected classes and object properties from *extvartrans* in MULTI, where these classes and properties will then be aligned to DOLCE+DnS Ultralite, an upper ontology suitable for modelling contexts (Dol, 2010).

7 Conclusion

As has been shown with the use cases pertaining to semantic alignment, there is slight variation depending on the viewpoint being considered. When considering a translation, the perspective should ideally be considered as well. In this paper, an extended version of the *vartrans* module for OntoLex-Lemon has been presented. More categories were provided from that of TRCAT, with new categories for both semantic and grammatical inequivalences, including lexical gaps. Additional classes and object properties were included in *extvartans* for grammar rules and language features. For grammatical inequivalences, the code fragments provided were the first step to ontology trans-

formation, where an ontology is transformed to a language-specific view, in line with the internationalisation goal of the OWL specification.

Acknowledgements

This work was financially supported by Hasso Plattner Institute for Digital Engineering through the HPI Research School at UCT.

References

2010. General Ontology for Linguistic Description (GOLD). http://linguistics-ontology.org/. Online; accessed: 2023, March 19.

2010. Ontology:DOLCE+DnS Ultralite. http://ontologydesignpatterns.org/wiki/Ontology:DOLCE+DnS_Ultralite. Online; accessed: 2023, January 24.

2014. *English-Xhosa / Xhosa-English Dictionary*, 14 edition. Pharos Dictionaries.

2017. *Gender Terminology: Sesotho 2017/18*. Commission for Gender Equality.

n.d. Translation Category Reference RDF Schema. http://purl.org/net/translation-categories. Online; accessed: 2023, May 27.

B. Vatant. n.d. The Lingvoj Ontology (lingvo). https://lov.linkeddata.es/dataset/lov/vocabs/lingvo. Online; accessed: 2023, March 19.

M. Baker. 2018. *In Other Words: A Coursebook on Translation*. Routledge, Oxon, UK.

P. Cimiano, P. Buitelaar, J. McCrae, and M. Sintek. 2011. LexInfo: A declarative model for the lexicon-ontology interface. *Journal of Web Semantics*, 9(1):29–51.

P. Cimiano, J.P. McCrae, and P. Buitelaar. 2016. Lexicon Model for Ontologies: Community Report. Final Community Group Report, 10 May 2016, Ontology-Lexicon Community Group under the W3C Community Final Specification Agreement (FSA).

J.D. Culler. 1976. *Saussure*. Fontana/Collins.

M. Dagut. 1981. Semantic “Voids" as a Problem in the Translation Process. *Poetics Today*, 2(4):61–71.

G. de Melo. 2015. Lexvo.org: Language-related Information for the Linguistic Linked Data Cloud. *Semantic Web*, 6(4):393–400.

G. De Schryver and M. Reynolds. 2019. *Oxford English-isiXhosa School Dictionary*, 7 edition. Oxford University Press Southern Africa.

J. Euzenat and P. Schvaiko. 2013. *Ontology Matching: Second Edition*. Springer-Verlag Berlin Heidelberg.

R. Gauton. 2008. Bilingual Dictionaries, the Lexicographer and the Translator. *Lexikos*, 18:106–118.

F. Gillis-Webber. 2023. Towards an Ontology of Viewpoints. In *Proceedings of the 13th International Conference on Formal Ontology in Information Systems (FOIS 2023), 17–20 July, Sherbrooke, Québec, Canada*.

R.H. Gouws and D.J. Prinsloo. 2005. *Principles and Practice of South African Lexicography*. AFRICAN SUN MeDIA, Stellenbosch, South Africa.

J. Gracia, E. Montiel-Ponsoda, D. Vila-Suero, and G. Aguado de Cea. 2014. Enabling Language Resources to Expose Translations as Linked Data on the Web. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 409–413, Reykjavik, Iceland. European Language Resources Association (ELRA).

G. Hirst. 2014. Overcoming linguistic barriers to the multilingual semantic web. In P. Buitelaar and P. Cimiano, editors, *Towards the Multilingual Semantic Web: Principles, Methods and Applications*, pages 3–14. Springer Berlin Heidelberg.

C. Kramsch. 1998. *Language and Culture*. Oxford Introductions to Language Study. Oxford University Press.

K. Mansoor. 2018. Translation Across the Difficulties of Equivalence Concept. *Scientific Bulletin of the Politehnica University of Timisoara. Transactions on Modern Languages*, 17(1):55–66.

B.M. Mini, editor. 2003. *The Greater Dictionary of isiXhosa: K to P*, volume 2. IsiXhosa National Lexicography Unit, University of Fort Hare.

H.W. Pahl, editor. 1989. *The Greater Dictionary of isiXhosa: Q to Z*, volume 3. University of Fort Hare.

S.L. Tshabe, editor. 2006. *The Greater Dictionary of isiXhosa: A to J*, volume 1. IsiXhosa National Lexicography Unit, University of Fort Hare.

W3C OWL Working Group. 2004. OWL Web Ontology Language Use Cases and Requirements: W3C Recommendation 10 February 2004. W3C Recommendation, World Wide Web Consortium. Online; accessed: 2023, April 28.

L. Zgusta. 1971. *Manual of Lexicography*. Academia.

L. Zgusta. 1978. Equivalents and Explanations in Bilingual Dictionaries. In Mohammad Ali Jazayery, Edgar C. Polomé, and Werner Winter, editors, *Linguistic and Literary Studies: Vol 4, Linguistics and Literature / Sociolinguistics and Applied Linguistics*, pages 385–392. De Gruyter Mouton, Berlin, New York.