УДК 681.3

А.А. Никоненко, аспирант
Киевский национальный университет им. Т. Шевченко,
факультет кибернетики, г. Киев, Украина
andrey.nikonenko@gmail.com

# Подходы к верификации знаний в лингвистических онтологиях

*В статье описан подход к процессу наполнения лингвистической онтологии данными, предложена модель оценки качества полученных знаний, приведены практические рекомендации по выявлению и автоматическому исправлению ошибок в знаниях.*

***Ключевые слова: онтологические базы знаний, верификация знаний, оценка качества концептов, оценка качества связей, автоматическое исправление ошибок***

## *Вступление*

В современной компьютерной лингвистике онтологии занимают одно из ключевых мест. Значимость данного типа ресурсов в решении задач инженерии знаний (knowledge engineering) связана с потребностью в хранении информации о языке в доступном для понимания компьютером виде. Одним из основных мест использования онтологий является разработка систем и методов семантического анализа естественноязыковой текстовой информации. В решении задач данного вида онтологии выполняют функции базы знаний, которая хранит информацию об объектах внешнего мира и связях между ними.

Статья посвящена вопросу создания универсальных лингвистических онтологий и проведения верификации полученных знаний с целью оценки их пригодности для использования в решении прикладных задач ассоциативно-семантического анализа текста на естественном языке. Актуальность приведенного в статье материала обусловлена как широкой популярностью использования онтологий в зарубежной компьютерной лингвистике, так и отсутствием общелингвистических онтологий для украинского языка. Приведенные в данной работе результаты являются продолжением исследований описанных в [1], часть статьи, описывающая принципы построения украиноязычной онтологии, использует идеи близкие к изложенным в [2-3] с их дальнейшей доработкой и расширением в сторону принципов описанных в [4]. Исследования модели верификации знаний онтологии идеологически близки работе [5], посвященной оценке качества онтологии корейского языка, но не являются ее повторением для украинского языка. В статье предложена новая модель автоматического выделения и исправления ошибок в лингвистической онтологии, использование которой позволяет значительно повысить практическую применимость данного вида ресурсов путем повышения их качества.

**Цель статьи** - исследовать вопрос адаптации онтологии одного языка для другого и систематизировать в виде модели экспериментально полученные результаты оценки качества созданных знаний. **Основные задачи** статьи:

1. Исследовать вопрос построения онтологии путем адаптации.
2. Сформулировать общие принципы совместной работы над онтологией.
3. Провести систематизацию основных видов ошибок снижающих качество знаний онтологии.
4. Предложить автоматические или автоматизированные методы исправления ошибок.

Формальная модель онтологии выглядит как O = <X,R,F>, где

- X — конечное множество понятий предметной области;
- R — конечное множество отношений между понятиями;
- F — конечное множество функций интерпретации.

Данная модель отвечает описанию онтологии как формальной системы, основанной на математически точных аксиомах, и соответствует формальным онтологиям, построенным на основе различных видов логики: предикатов первого порядка, дескриптивной, модальной и т.п. [1].

Другим широко распространенным видом онтологий являются лингвистические (лексические) онтологии типа WordNet [6]. Данный класс ресурсов более приспособлен для хранения информации о структуре естественного языка за счет менее строгой системы представления знаний. Простота спецификации лингвистических онтологий позволяет создавать базы знаний размером в сотни тысяч концептов. Для формальных онтологий достижение такого размера не является возможным. Объем знаний и близость структуры лингвистической онтологии к структуре естественного языка привели к

широкому распространению этого вида ресурсов в решении задач автоматической обработки языка.

Модель лингвистической онтологии также может быть представлена в виде тройки <X,R,F>, при наложении дополнительных условий, а именно: множество F должно быть пустым, а множество R должно состоять из объединения множества семантических (несущих смысловую нагрузку и соединяющих концепты) и лексических (отображающих структуру языка и соединяющих слова концептов) связей онтологии.

### *Построение онтологии*

В данной работе мы обойдем стороной вопросы архитектурного характера: выбор структуры хранения данных, выбор языка представления знаний и дизайна системы для обеспечения взаимодействия с онтологией. Эти вопросы подробно рассмотрены в статьях [7,8], поэтому основной акцент сделаем на процесс наполнения лингвистической онтологии. Под наполнением мы будем понимать задачу адаптации знаний онтологии одного языка к другому.

Данная задача часто возникает при попытке построения WordNet-подобной онтологии, в частности, для европейских языков (см. например, EuroWordNet [2,9], BalkaNet [3]). Для построения локальной версии онтологии типа WordNet обычно используются методологические наработки, полученные во время создания оригинальной англоязычной версии, что позволяет исследователям перейти сразу к процессу создания концептов на своем языке, не заостряя внимание на построении смысловой иерархии понятий языка. Полученные таким образом данные тесно связаны со смысловыми единицам оригинальной онтологии, т.е. выполняется не столько работа по созданию лексической базы с нуля, сколько адаптация смыслов англоязычных концептов к локальному языку, а сам процесс построения онтологии сводится к получению проекции смыслов оригинального WordNet на конкретный язык.

Работа по адаптации смыслов с одного языка на другой является более простой, чем полный цикл разработки онтологии, однако она также требует значительных временных затрат и поэтому не может быть выполнена исключительно экспертами. Подразумевается, что такую работу должны производить специально обученные специалисты, а за контроль качества полученных данных отвечает команда экспертов. Характер и средства выполнения работы целиком зависят от договоренностей внутри команды проекта, создание знаний может проводиться как с использованием специальных программных систем и методов, так и без них в режиме разделения исходной онтологии на участки и передачей каждого из них в ведение определенной группе. Также существуют специальные стратегии адаптации, например, путем спуска по IS-A иерархии [2].

В проекте по построению универсальной украиноязычной онтологии UWN [8] нами было опробовано несколько подходов к адаптации знаний. В результате наиболее эффективным был признан wiki-подобный подход [10], который позволяет большому количеству пользователей одновременно вносить правки в большое количество синсетов. Для повышения качества создаваемых знаний мы разделили всех пользователей на три основных категории: Читатели, Редакторы и Модераторы. Каждая из категорий имеет свои права и политики работы с данными. В общем виде процесс работы с онтологией выглядел следующим образом: Редактор вносит основные правки в определенный синсет, после чего синсет поступает на проверку Модератору, который оценивает качество, вносит правки и подтверждает корректность концепта либо отклоняет его. Отклоненные концепты возвращаются на доработку автору-Редактору. Повторно обработанный синсет поступает на проверку тому же Модератору, который проверял его ранее. В общем виде процесс обработки синсета изображен на рис. 1.

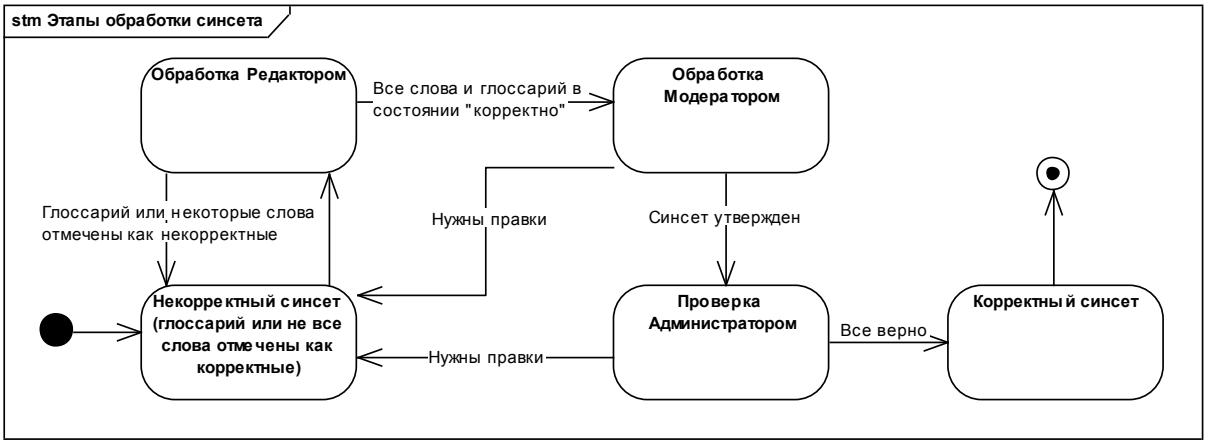


Рисунок 1 - Этапы обработки синсета

Существуют также механизмы, которые позволяют в некоторых случаях произвести смену Редактора-владельца синсета, либо передать его на проверку другому Модератору. Дополнительно, в разработанной системе существует третий уровень проверки качества – Администраторы, пользователи с данным профилем осуществляют периодические проверки наиболее слабых участков онтологии, определяемых автоматическими оценочными функциями. Детально процесс совместного наполнения онтологии описан в работе [10].

### *Оценка качества знаний онтологии*

В разных источниках процесс проверки качества знаний называют по-разному. Два наиболее распространенных варианта: верификация и валидация. С нашей точки зрения, в данной ситуации более корректным будет использовать термин верификация, поскольку мы проводим общий анализ характеристик полученной базы знаний на соответствие требованиям, которые предъявляются к качеству онтологических баз, а не выполняем проверку соответствия внесенных знаний специфике определенной ПРО (прикладной области) для разработки конкретного продукта. Сама проводимая нами оценка качества ставит своей целью не доказательство формального соблюдения всех требований процесса построения онтологии, а выявление слабых мест и некорректных смысловых единиц с целью их дальнейшей доработки и улучшения.

Проблема точности и полноты наполнения концептов является одной из наиболее важных при составлении онтологии, а ее корни находятся в универсальном характере онтологии. Универсальная онтология – это попытка всеобъемлющего описания мира, поэтому, в отличие от онтологий ПРО, она не может быть составлена узким кругом экспертов определенной тематики и требует от составителей энциклопедических знаний по большому количеству тематик. Неудивительно, что даже трехуровневая система проверки качества не позволяет избежать попадания в финальную версию онтологии синсетов с наборами слов не всегда отвечающими смыслу концепта или синсетов с «бедным» набором слов. Особняком стоит вопрос о характере изменения связей семантической сети при переносе (адаптации) онтологии одного языка на другой. Таким образом, оценка качества лингвистической онтологии делится на две основных задачи: оценка качества концептов и оценка полноты и корректности связей меж концептами.

К сожалению, в литературе чаще всего под верификацией онтологии понимают исключительно формальную проверку правильности процесса построения [11], соответственно, утверждать об успешной верификации универсальной онтологии в данном ключе можно только после проверки экспертами каждой предметной области онтологии. Под проверкой предметной области мы понимаем ручную проверку каждого понятия ПРО и его связей с другими понятиями ПРО. Выполнить такую проверку на универсальной онтологии в достаточном для верификации объеме не представляется возможным. Обзорно ознакомиться с основными вопросами, возникающими при попытке формальной верификации онтологических баз знаний, можно в [12], мы же сосредоточимся на методах выявления ошибок и возможностях расширения онтологии.

### *Оценка концептов*

В данный момент составляемая нами онтология UWN содержит более 80000 концептов-существительных. Практический опыт работы с полученной базой знаний позволяет построить следующий список наиболее распространенных проблем:

1. Некорректное/неточное соответствие между смыслами англоязычного и украиноязычного глоссария синсета (подразумевается, что при адаптации онтологии каждому оригинальному концепту должен быть подобран близкий по смыслу концепт на украинском языке).
2. Бессмысленный или сложный для понимания глоссарий (как правило, это предложения, составленные с нарушением правил синтаксиса).
3. Нарушение логической структуры глоссария (при адаптации мы стараемся сохранить исходную структуру глоссария, например, если оригинальный глоссарий содержит два примера использования терминов, то и украиноязычный вариант должен содержать столько же).
4. Ошибки грамматического характера в глоссарии, в т.ч. орфографические ошибки, несогласованность элементов предложения по родам и числам и т.д.
5. Наличие терминов, соответствующих смыслу синсета, в примерах в глоссарии, но отсутствие их в списке слов.
6. Наличие в синсете слов, смысл которых не соответствует глоссарию.
7. Наличие в синсете англоязычных слов вместо их украиноязычных соответствий.
8. «Бедность» синсета – ситуация, когда при наличии в языке большого количества

синонимов, описывающих один и тот же смысл, в синсете используется лишь малое их количество.

9. Наличие в синсете дублей – ситуация, когда в списке слов одно слово или словосочетание используется несколько раз в виде отдельных элементов. Существующие подвиды:
   a. простое дублирование;
   b. дублирование с добавлением в конце/начале слова дополнительных символов, например, цифр;
   c. дублирование словосочетаний с использованием различных символов (различного количества символов) в виде разделителя.
10. «Шум» - занятие одной или нескольких ячеек для слов бессмысленными наборами символов.

### *Методы повышения качества концептов*

Рассмотрим основные подходы, которые позволяют выявить или исправить перечисленные выше проблемы:

1. Оценка степени соответствия смыслов двух текстов является достаточно сложной задачей и полностью может быть решена только благодаря участию человека-эксперта. Мы допускаем возможность частичного решения данной задачи на основе построения автоматической системы выделения ключевых понятий (entities) текста и построения графа отношений между ними. Для построения такой системы понадобится, как минимум, построение синтаксического дерева каждого предложения текста, установление смысловых связей меж его элементами, решение неоднозначностей (переход от слов к смыслам), объединение полученных подсетей в единую сеть, представляющую собой некую проекцию смысла глоссария на узлы онтологии. Построив такие сети (узлами выступают номера синсетов, которые в обоих вариантах онтологии одинаковые; ребра принадлежат единому множеству семантических связей онтологии) для англоязычного и украиноязычного глоссария синсета мы сможем автоматически определить степень их подобия. Несмотря на простоту идеи, перед прикладным использованием описанная система обязательно должна пройти проверку качества работы.
2. Чрезмерная сложность и загроможденность глоссария может быть определена с помощью построения синтаксических деревьев и их дальнейшего анализа. Мы предполагаем, что связность глоссария обратно пропорциональна размеру синтаксического дерева после проведения свертки, однако, определение точного числового значения соотношения количества слов в предложении к величине его дерева после свертки предстоит определить экспериментально. Нарушение синтаксических правил при составлении предложения также должно характеризоваться количеством висящих (изолированных) узлов дерева синтаксического разбора и, соответственно, также будет учитываться при свертке. Сложнее ситуация с определением бессмысленности глоссария, поскольку предложение может быть построено по всем правилам грамматики, но не нести при этом смысла, например: «Релиз птицы невероятно интересно придумать спиной», мы предполагаем, что в данном случае узлы семантической сети, полученной путем проекции слов предложения на онтологию, будут разнесены на большие расстояния (имеются в виду расстояния по одной из классических метрик определения семантической связности [13]). Спорным моментом в данном методе решении задачи является то, что для оценки качества ресурса мы используем сам ресурс, причем предполагается, что он содержит некоторое количество ошибок.
3. По сравнению с двумя предыдущими пунктами сравнение структур глоссариев выглядит тривиальной задачей и может быть осуществлено с помощью регулярных выражений (т.к. шаблоны глоссариев весьма единообразны), либо с помощью разбора предложения на составные части с выделением и сопоставлением порядка основных структурных элементов: кавычки, точки с запятой и т.д.
4. Орфографические ошибки в глоссарии и словах должны выявляться с помощью утилиты по проверке правописания (spell checker). В состав UWN входит такая утилита для проверки украиноязычного текста (также она встроена в онтокорректор, что должно снизить количество ошибок данного типа). Ошибки грамматического характера могут быть определены с помощью морфосинтаксической разметки текста (part-of-speech tagging [14]) и дальнейшего определения всех структур синтаксического дерева, элементы которых не согласованы по родам, числам, лицам, падежам и т.д.
5. Проверка вхождения примеров из глоссария в список слов синсета достаточно проста. Поскольку все примеры в глоссариях заданы одним общим шаблоном, то они

могут быть выявлены с помощью регулярных выражений и, по необходимости, внесены в список слов синсета.

6. К сожалению, в данный момент нет единой методики, которая позволяла бы выявлять соответствие слова смыслу предложения. В качестве методов частичного решения данной задачи нами предложено два метода:
   a. Подсчет расстояний в онтологии от семантической сети, представляющей проекцию смысла предложения, до каждого из слов, приведенных в синсете. Слова, семантические расстояния от которых до предложения будут намного больше среднего расстояния по всему множеству слов, с большой вероятностью могут считаться отличными от глоссария по смыслу.
   b. Другим вариантом будет проверка синонимичности слов синсета. Подход основывается на предположении, что большинство слов в синсете соответствует его смыслу, тогда выделить лишние слова можно опираясь на словари синонимов.
7. В общем случае украиноязычная онтология должна содержать лишь украиноязычные слова, исключения из этого правила составляют латинские названия растений и животных, некоторые латинские же термины и небольшое количество синсетов, слова которых не имеют соответствия в украинском языке. Как правило, отсутствие соответствия связано с более развитой классификацией некоторых объектов в английском языке. Пример такого синсета приведен на рисунке 2, в данном случае слова hero, grinder и другие не могут быть просто переведены на украинский язык, поскольку несут отличный от простого перевода смысл. Для решения задачи, поставленной в данном пункте, нам понадобится найти все синсеты в онтологии, которые содержат латиницу, затем вычесть из них множество синсетов с латинским названиями родов, видов, растений, животных и т.д., а также синсеты, содержащие фразы на латинском языке. Оставшееся множество синсетов (по нашим оценкам, оно составляет одну-две сотни) должно быть проверено на корректность вручную.

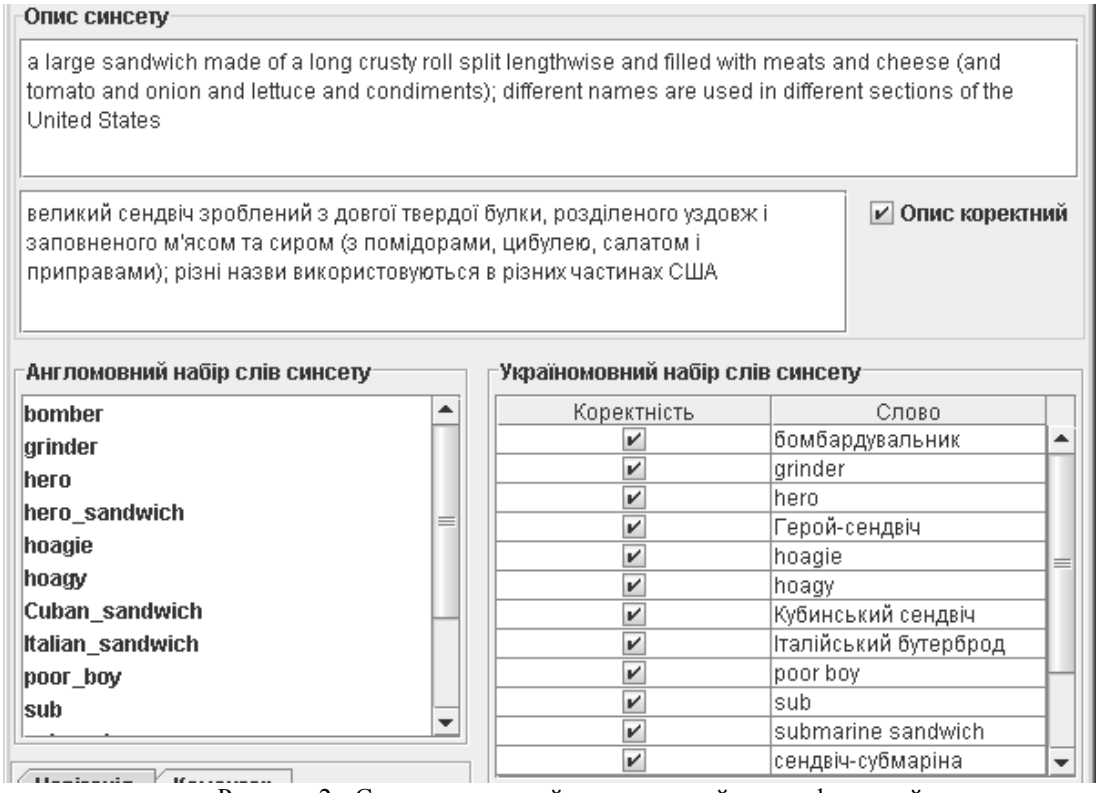


Рисунок 2 - Синсет с широкой англоязычной классификацией, которая не может быть адекватно передана в украинском языке

8. Количеством слов в синсете является одним из ключевых показателей перспективности использования онтологии, поэтому действия направленные на расширение

данных в концептах имеют одно из первостепенных значений. Если с автоматическим расширением синсетов все достаточно просто – необходимо использовать словари синонимов, как это описано в пункте 6.b, то с выявлением синсетов-кандидатов на расширение все несколько сложнее. С одной стороны, можно попытаться расширить все синсеты онтологии, но полностью автоматическое добавление синонимов может привести к падению качества синсетов за счет добавленных синонимов, имеющих другую смысловую нагрузку, поэтому работа по расширению должна проходить в автоматизированном режиме, а значит, потребует значительного расхода человеческого времени. Эффективнее всего такую работу проводить только для «бедных» синсетов, при этом «бедность» синсета напрямую не зависит от количества слов в нем, например, синсеты описывающие известных писателей могут содержать всего одно слово (собственно ФИО писателя) и по понятным причинам не являться «бедными». Мы предлагаем определять синсеты требующие пополнения по двум критериям:

a. Соотношение количества слов в украиноязычном и англоязычном синсетах: если украиноязычный содержит меньше слов, то это кандидат на пополнение. Данный критерий является спорным, исходя из соображений, изложенных в пункте 7, однако, в связи с относительной редкостью примеров, характерных для 7-го пункта, все же может использоваться достаточно эффективно.
b. Упор на словари синонимов. Сложность в использовании словарей синонимов в том, что одно и то же слово может иметь одновременно несколько наборов синонимов, причем определить без контекста нужный набор не представляется возможным. При наличии в синсете нескольких слов они могут играть роль контекста, в этом случае определение близкого по смыслу набора синонимов сводится к выбору множества максимальной мощности, полученного пересечением списка синонимов со списком слов синсета. Далее производится сравнение количества слов в выбранном наборе с количеством слов в синсете, по их соотношению определяется степень бедности синсета.

Сложнее ситуация если синсет содержит всего одно слово, которому соответствует несколько наборов синонимов разной мощности. В этом случае в качестве вероятностной оценки бедности синсета можно использовать среднее арифметическое мощностей множеств синонимов.

9. Наличие дублей в синсете довольно распространенная ситуация, которая решается следующим образом:
   a. Поиском дублей в списке слов синсета.
   b. Поиском дублей вида «слово1» или «1слово» и более сложных с помощью регулярных выражений вида «%слово%». При этом мы предполагаем, что список слов синсета содержит хотя бы одно слово в неискаженной форме, т.е. синсет может содержать такие слова: «слово1, 1слово, слово, 1слово1 …». Каждое из слов синсета мы пытаемся использовать в качестве основы регулярного выражения. Также при составлении регулярного выражения следует учитывать словосочетания, которые могут подпадать, например, под шаблон «слово%», но не являться дублем. Наиболее простой метод их учета – исключить шаблоны вида «слово %» и «% слово».
   c. Места в списке слов синсета также могут занимать словосочетания, выявление дублей среди которых также проводится за счет регулярных выражений, как правило, за счет замены пробела в словосочетании на «%». При этом также внимательно нужно относиться к появлению дополнительных слов в словосочетании, например, неаккуратно составленный шаблон может принять за дубль словосочетания вида «слово1 слово2» и «слово1 слово3 слово2».
10. Для различения бессмысленных последовательностей символов и слов украинского языка необходимо использовать лексико-грамматический словарь, при этом следует учитывать возможность вхождения в синсет латинских и английских слов, как это описано в пункте 7. Также следует предусмотреть разбор словосочетаний на составляющие для анализа по словарю. Разбор сложных словосочетаний и анализ их осмысленности можно рассматривать как подзадачу пункта 2.

### *Оценка связей*

Качество онтологии зависит от двух параметров: качества синсетов и качества связей. Как показано выше, большинство аспектов оценки качества синсетов не представляют особой сложности и могут быть, как минимум частично, автоматизированы. Оценка качества связей представляет собой более сложную процедуру и завязана на понимание взаимосвязей между объектами реального мира, которое не может быть выведено автоматически. В данном разделе мы рассмотрим две основные проблемы, которые были обнаружены при создании онтологии UWN, и постараемся указать пути их решения.

**Проблема полноты связей**

Первой и, наверное, основной проблемой является неполнота набора связей в онтологии, причина которой кроется в том, что семантические связи, как правило, не могут быть получены автоматическим путем из какого-либо источника и создаются вручную. Таким образом, задача создания графа связей между синсетами упирается в создание максимально детальной классификации всех понятий онтологии и установление связей между ними (см. рисунок 3).

Даже исходный набор семантических связей в WordNet 3.0 имеет лакуны, например, нами было обнаружено отсутствие обратных связей в 794 случаях. Выявленные нами недостающие связи можно разделить на два класса:

- прилагательное – прилагательное,
- глагол – глагол.

Например, связь вида *(94448, a, ^, 118066, a)* должна иметь обратную связь вида *(118066, a, ^, 94448, a)*, а связи вида *(1432601, v, *, 1831531, v)* должна соответствовать связь *(1831531, v, >, 1432601, v)*. Типы связей, в некоторых случаях не имеющие обратных, перечислены в таблице 1.а, а соответствующие им типы обратных связей перечислены в таблице 1.б. Мы предполагаем, что комплексное исследование связей способно выявить лакуны и в связях других типов. Однако если пропущенные связи могут быть выведены путем исследования структуры онтологии, опираясь на свойства имеющихся в наличии связей, то связи, присутствующие между объектами реального мира, но отсутствующие в онтологии, весьма плохо поддаются автоматической генерации.

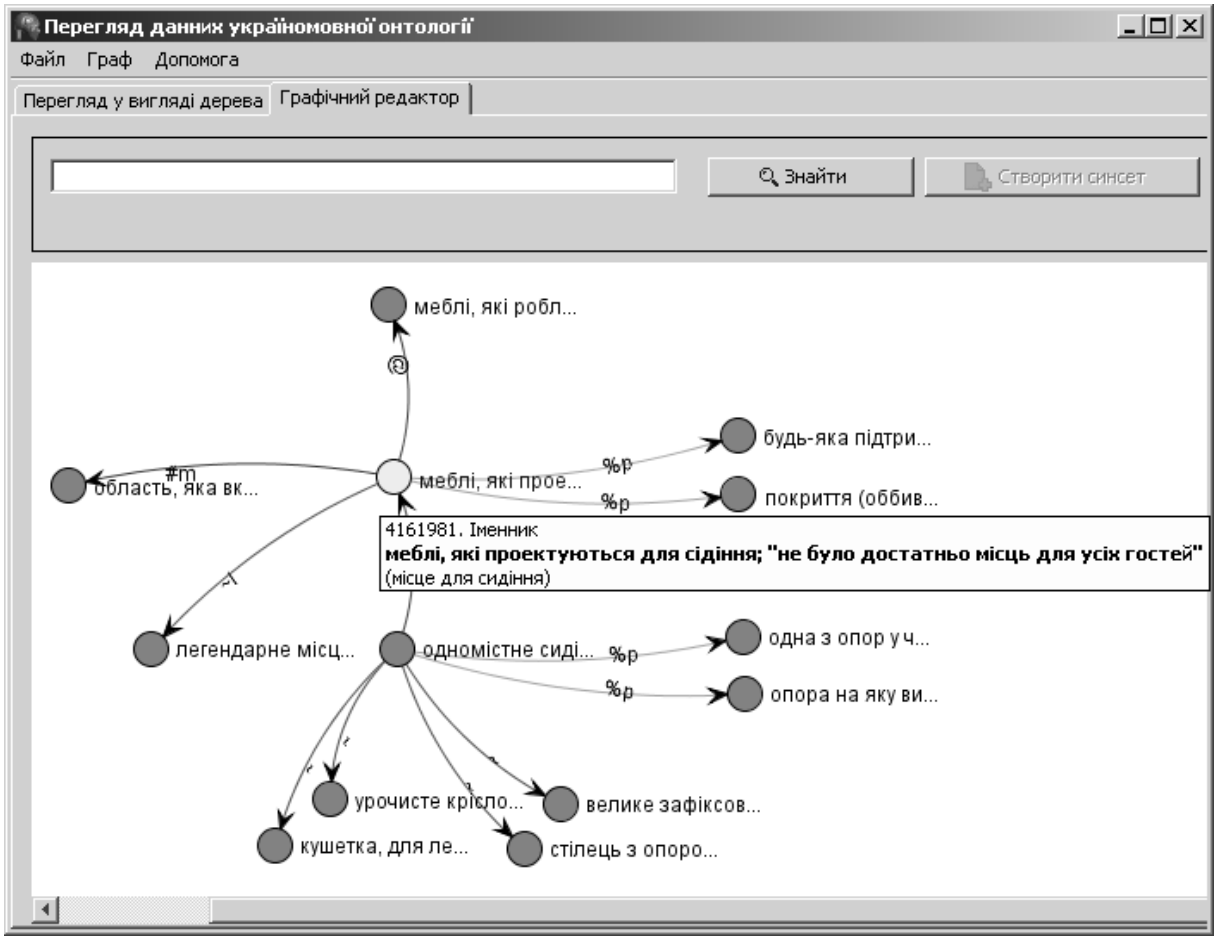


Рисунок 3 - Фрагмент онтологии вокруг синсета «місце для сидіння»

Таблица 1- Список пропущенных связей

| SYNSET1 | RELATION | SYNSET2 |
|---|---|---|
| V | * | V |
| V | > | V |
| V | ^ | V |
| A | ^ | A |

*а* – типы связей, у которых пропущены обратные

| SYNSET1 | RELATION | SYNSET2 |
|---|---|---|
| V | > | V |
| V | * | V |
| V | ^ | V |
| A | ^ | A |

*б* – обратные связи, соответствующие связям в таблице *а*

Одним из примеров отсутствующих связей является смысловая связь между синсетом «ОПЕК» и синсетом «нафта». На рисунке 4, слева желтым цветом изображен синсет «ОПЕК» и все его связи в онтологии (снизу голубыми связями показаны страны входящие в ОПЕК, а сверху черными связями показаны синсеты «нафтовий картель» и «міжнародна організація», в которые синсет «ОПЕК» входит как подвид). В правой части рисунка желтым цветом изображен синсет «нафта» и все его связи в онтологии (снизу голубой и синей связями показаны элемент нефти «вуглець» и подвид нефтепродуктов «залишкова нафта», выше находятся синсеты «скам'яніле паливо» и «олія», подвидами которых является нефть). Важнейшим выводом, который можно получить, проанализировав данную систему связей, является то, что в онтологии синсет «ОПЕК» никак не связан с синсетом «нафта», хотя в реальном мире это два очень тесно связанных понятия. Более того, если мы построим кратчайший путь меж этими синсетами, то он пройдет через 7 дополнительных концептов. Детальный анализ пути показывает, что он идет от нефти через травы, из которых делают разные типы масел, до трав которые растут на территории США, затем переходит на синсет, описывающий США как страну, далее, через Организацию американских государств (ОАГ), выходит на международные организации как таковые и оттуда спускается к ОПЕК. Соответственно, для любых методов оценки семантической связности и близости синсеты «ОПЕК» и «нафта» будут определяться как слабосвязанные, хотя в реальном мире это не так.

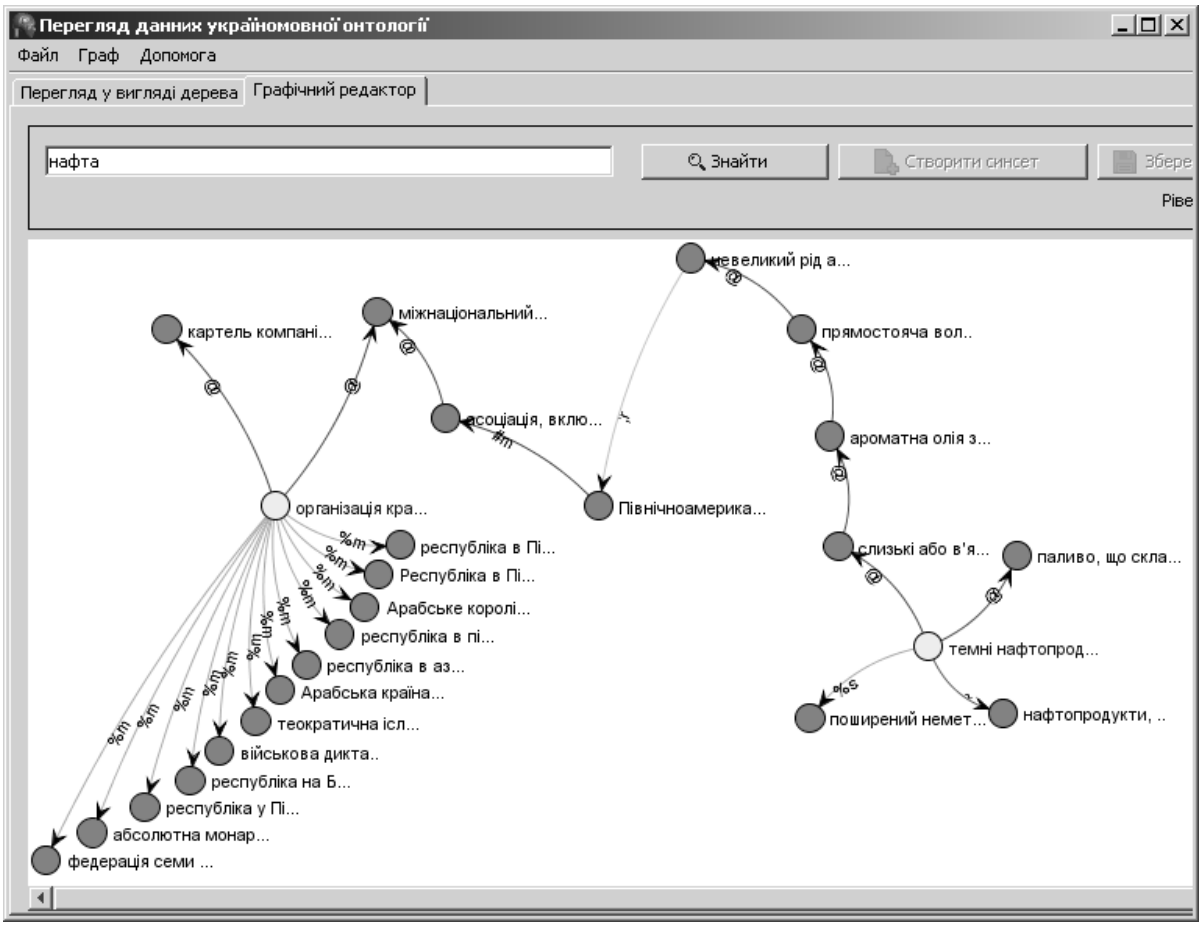


Рисунок 4 - Фрагмент онтологии, показывающий кратчайший путь меж синсетом «ОПЕК» (слева) и синсетом «нафта» (справа)

Для решения проблемы отсутствующих связей, в качестве базового пути, можно предложить использование метода Леска [15], который разработан для оценки близости двух понятий путем оценки схожести их словарных определений. Если в качестве словарных определений использовать глоссарии синсетов, то в случаях, когда их схожесть по методу Леска будет выше некоторого порогового значения, можно создавать в онтологии новые связи. Очевидно, что семантика создаваемых связей будет различной, а значит, разумным будет ввести в онтологию некий новый тип связи, смысловая характеристика которой в каждом конкретном случае должна быть установлена человеком-экспертом. Учитывая специфику онтологии также необходимо внести некоторые модификации в метод Леска, учитывающие не только взаимные пересечения глоссариев, но и случаи вхождения слов одного синсета в глоссарий другого. В результате обработки онтологии предложенным алгоритмом получим большой набор новых связей, типизация и проверка корректности которых должна осуществляться экспертами. Общую эффективность метода можно оценить без привлечения экспертов, для этого необходимо сравнить результаты работы некоторого подмножества методов определения семантической связности [13] на онтологии до внесения дополнительных связей и после.

**Смысловая структура связей**

Вторая ключевая проблема связана с самим методом построения онтологии UWN. Поскольку UWN создавалась путем адаптации исходной англоязычной онтологии, то связи между понятиями отображают семантику английского языка. Существует мнение, что такое отображение является полностью корректным, т.к. любой язык описывает объективную реальность, а значит, разные языки описывают разными терминами одни и те же объекты, соответственно, структура связей между объектами не должна изменяться при переходе с языка на язык. Например, птица, не зависимо от языка, будет иметь крылья, хвост, клюв, перья и т.д., поэтому связи синсета «птица» с синсетами, описывающими эти понятия, останутся неизменными при смене языка онтологии.

Как показывает практика, изложенный выше принцип работает далеко не всегда, поскольку язык используется не как средство для отображения объектов внешнего мира в символы, а, скорее, предоставляет возможности символьного описания объектов некого общего ментального поля народа, т.е. описывает понятия не внешнего, а внутреннего мира, существующего внутри сознания. Таким образом, любой реально существующий объект должен пройти процесс опознания и изучения в сознании индивида и лишь затем получить некоторое слово-описание в языке. Поэтому описания объектов в языке имеют именно те характеристики, которыми их наделило сознание. В качестве доказательства можно привести большое количество объектов реального мира, которые имеют разное количество наименований в различных языках, причем каждое из таких наименований описывает определенные характеристики объекта, которыми его наделило сознание носителей языка. Подход к языку как к средству описания внутренних объектов сознания позволяет объяснить отличия в количестве понятий, описывающих один объект в разных языках. Чем более важными являются различные характеристики объекта для носителей языка, тем большее количество терминов они вводят для его описания. В качестве примера можно привести слово снег, для обозначения разных видов которого в эскимосских языках используется более 20 слов.

Поскольку язык отображает ментальное поле народа, то отличие структуры разных языков напрямую зависит от степени близости народов. Поэтому при создании онтологии путем адаптации, кроме изменения смыслов синсетов для придания им более характерных для локального языка смыслов, также необходимо изменение сети связей между концептами для отображения реальной картины взаимодействия понятий в смысловом поле языка. На данном этапе развития компьютерной лингвистики справится, хотя бы частично, с этой задачей может только группа квалифицированных экспертов-лингвистов.

## *Заключение*

Приведенное в статье исследование базируется на данных, полученных в процессе построения украиноязычной онтологии UWN путем адаптации знаний лексико-семантической базы знаний WordNet. Одним из основных результатов работы является описанная модель классификации ошибок знаний и методов их автоматического исправления. Как видно из приведенного материала, причиной возникновения большинства некачественных концептов является человеческий фактор, поэтому построение детальной классификации ошибок является необходимым шагом на пути к построению автоматической системы контроля качества. Предложенные в статье способы автоматического устранения ошибок могут быть использованы в системе автоматизированного повышения качества знаний.

Исправление структуры семантической сети для достижения максимально точного отображения смысловой картины языка на узлы онтологии является более сложной задачей,

работа в этой области – направление наших дальнейших исследований.

### *Список литературы*


1. Никоненко А.А. Обзор баз знаний онтологического типа / А.А. Никоненко // Искусственный интеллект. – 2009. - № 4.- С. 208-219.
2. The Top-Down Strategy for Building EuroWordNet: Vocabulary Coverage, Base Concepts and Top Ontology / [Alonge A., Bertagna F., Bloksma L. et al.]. – 1998.
3. BalkaNet: A Multilingual Semantic Network for Balkan Languages / [Stamou S., Oflazer K., Pala K. et al.] // In Proceedings of the 1st Global WordNet Conference, (Mysore, India). - 2002.
4. Leuf B. The Wiki way: quick collaboration on the Web / Leuf B., Cunningham W. - Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 2001.
5. Altangerel Chagnaa. Extracting Features for Verifying WordNet / Altangerel Chagnaa, Cheol-Young Ock and Ho-Seop Choe // Lecture Notes in Computer Science. – 2007. - Volume 4798. – P. 605-610.
6. George A. Miller WordNet: A Lexical Database for English / A. George // Communications of the ACM. – 1995. - Vol. 38, No. 11. – P. 39-41.
7. Никоненко А.О. Проект UWN: Методологія створення універсальної онтологічної бази знань української мови / А.О. Никоненко // Тези міжнародної наукової конференції [MegaLing'2011 «Горизонти прикладної лінгвістики та лінгвістичних технологій»], (Партеніт, Крим, Україна). – Партеніт, 2011. – С. 57-58.
8. Никоненко А.О. Проект UWN: Досвід створення універсальної онлайн онтології української мови / А.О. Никоненко // Тези міжнародної наукової конференції [ISDMCI'2011 «Интеллектуальные системы принятия решений и проблемы вычислительного интеллекта»],(Євпаторія, Крим, Україна). – Євпаторія, 2011. – С. 92-96.
9. Piek Vossen(ed). EuroWordNet: A Multilingual Database with Lexical Semantic / Piek Vossen // Networks Kluwer Academic Publishers. - Dordrecht, 1998.
10. Никоненко А.О. Проект UWN: Методи спільного редагування онлайн онтологій / А.О. Никоненко // Тези міжнародної науково-практичної конференції [«Інформаційні технології та комп'ютерна інженерія»], (Харків, Україна, 2011). – Х. – 2011. – С. 43-44.
11. Gómez Pérez A. Ontological Engineering / A. Gómez Pérez, M. Fernández López, Chorcho O. // Springer Verlag. – London, UK, 2004.
12. Климова М.В. Розробка методу та моделі верифікації знань в онтологічних системах / М.В. Климова // Східно-Європейський журнал передових технологій. – 2009. – № 4/8 (40). – с. 32-36.
13. Yang, D. Measuring Semantic Similarity in the Taxonomy of WordNet / Yang, D. and D. M. W. Powers // Twenty-Eighth Australasian Computer Science Conference (ACSC2005), (Newcastle, Australia, ACS). - 2005.
14. Веб-ресурс: http://en.wikipedia.org/wiki/Part-of-speech_tagging
15. Lesk M. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. / M. Lesk // In SIGDOC '86: Proceedings of the 5th annual international conference on Systems documentation, (New York, NY, USA. ACM). – NY, 1986. – P. 24–26.


*Надійшла до редакції 31.03.2012*

**А. А. Никоненко**
Київський національний університет імені Тараса Шевченка

**A. A. Nykonenko**
Kyiv National Taras Shevchenko University

**Підходи до верифікації знань в лінгвістичних онтологіях.**

Статтю присвячено процесу наповнення даними лінгвістичної онтології та вирішенню проблем, що виникають при створенні знань. Основну увагу приділено практичним рекомендаціям з виявлення та автоматичного виправлення помилок в онтології.

***Ключові слова: онтологічні бази знань, верифікація знань, оцінка якості концептів, оцінка якості зв'язків, автоматичне виправлення помилок***

**Linguistic Ontologies Verification Approaches.**

The article is about ontology creation process and knowledge generating problems. At first it presents some facts about approaches to assessing the quality of the knowledge. Then it throws light on the practical recommendations for identifying and correcting errors automatically.

***Keywords: ontological knowledge bases, knowledge verification, concepts quality estimation, relations quality estimation, automatic error correction***