Прикладная лингвистика и лингводидактика

УДК 81’322.2 + 81’374 **DOI: 10.14529/ling160403**
ББК Ш11

УНИФИКАЦИЯ ПЕРЕВОДНЫХ ЭКВИВАЛЕНТОВ МНОГОКОМПОНЕНТНОЙ ЛЕКСИКИ ПРИ КОЛЛЕКТИВНОЙ РАЗРАБОТКЕ ДВУЯЗЫЧНЫХ ЛЕКСИКОНОВ

О.И. Бабина, Е.С. Тамгина
Южно-Уральский государственный университет, г.Челябинск

Рассматривается вопрос коллективной разработки двуязычного лексикона, представляющего собой базу знаний для системы автоматической обработки текстов. Пилотное исследование подтвердило тезис о неизбежной вариативности переводов одних и тех же лексических единиц, предлагаемых различными переводчиками. Для соблюдения автоматической системой принципа единства терминологии предлагается введение в процесс компиляции автоматического словаря этапа унификации переводных эквивалентов многокомпонентной лексики. Авторами разработана формализованная процедура для выполнения данного этапа. Экспериментальная проверка работы предложенной процедуры на ограниченной выборке целевых терминов позволила сделать вывод о возможности свести к минимуму количество переводных эквивалентов термина в составе более длинных вхождений лексикона. Корректировка вхождений лексикона в результате выполнения предложенной процедуры позволяет унифицировать лексические манифестации вхождений и в перспективе обеспечить более высокое качество работы автоматической системы.

Ключевые слова: многокомпонентная лексика, коллективная разработка лексикона, переводной лексикон, управление терминологией, автоматическая обработка текстов.

Введение

В современных условиях необходимости работы с большими данными автоматизация различных процессов человеческой деятельности является неотъемлемой частью действительности. Это относится и к необходимости автоматизации различных аспектов деятельности переводчика посредством построения систем перевода, а также лингвистических ресурсов, представляющих собой базу знаний для соответствующих автоматических систем.

Лексический компонент речевой способности человека представляется в лексиконе [3]. Лексикон является отражением наиболее подвижной части естественного языка. При этом именно лексикон выполняет ключевую роль в организации речевой деятельности, так как, прежде всего, с помощью единиц лексикона осуществляется семантическое наполнение высказывания. В связи с этим, вполне закономерно, что компьютерный лексикон[1] (репрезентирующий модель лексикона человека) представляет собой неотъемлемую часть базы знаний систем автоматической обработки текстов (АОТ).

Автоматизация построения лексикона наталкивается на ряд проблем, обусловленных природой единиц лексикона: подвижностью естественного языка, многозначностью лексических единиц, возможностью синонимичных номинаций и т. д. Все эти проблемы приходится решать лингвисту при осуществлении лексикографических работ.

При построении компьютерных лексиконов для многоязычной обработки текста ключевыми задачами является сбор словника и поиск эквивалентов лексических единиц на различных языках [1].

Первой задачей, требующей внимания и разработки индивидуальных стратегий для ее решения, является отбор словника. Для ее решения необходимо определить предметную область, для которой разрабатывается система АОТ, и использовать собранный (самостоятельно или кем-либо ранее) корпус текстов (в современной действительности лексикографические работы в подавляющем большинстве основываются на корпусном подходе). Особенно важным является определение допустимых единиц лексикона. В традиционной лексикографии, как правило, единицами выступают отдельные слова. Однако еще в работах З. Харриса указывалось на возможность рассматривать некоторые аспекты значения как функцию дистрибутивных отношений [12]. Многие исследователи сегодня, отмечая важность словосочетания для разрешения лексической многозначности [7, 13, 15, 17, 18], признают необходимость разработки лексиконов, включающих единицы-словосочетания, что находит широкое применение в

[1] Далее в статье будем использовать термин «лексикон» в узком смысле, подразумевая «компьютерный лексикон»

практике составления машиночитаемых словарей для систем АОТ.

В связи с применением дистрибутивного принципа, отдельной задачей является автоматизация отбора многокомпонентной лексики. Помощь в этом оказывает специализированное программное обеспечение, поддерживающее функции составления конкордансов, а также частотных списков n-грам (например, «Программа автоматизированного составления и обработки словников» [4, с. 52–63], программный комплекс LingAssistant [2] и др.). Однако более «продвинутым» шагом к построению словника, включающего многокомпонентную лексику, является решение задачи автоматического отбора словосочетаний по определенному параметру (например, именных групп, устойчивых словосочетаний, терминов и т. п.). В этой области ведутся многочисленные разработки [6, 8, 11, 16], использующие, чаще всего, статистические метрики или гибридный подход для идентификации словосочетаний и определения степени их устойчивости.

При решении второй задачи – поиске эквивалентов лексических единиц на различных языках – разработчики также прибегают к различным способам автоматизации поиска эквивалентов: автоматизация выравнивания с последующим извлечением эквивалентов из параллельных корпусов или, руководствуясь дистрибутивной гипотезой, извлечение лексики для переводного словаря из псевдопараллельных корпусов [9, 10, 14]. Однако автоматические методы не позволяют достигнуть абсолютной точности полученных переводных эквивалентов, что с неизбежностью ведет к необходимости задействовать ручной труд переводчика, как минимум, для верификации составленных переводов. При этом в процедуре верификации могут быть задействованы разнообразные лингвистические ресурсы: электронные словари, памяти переводов, тезаурусы, системы машинного перевода, корпусы текстов, системы информационного поиск и т. д.

Ручной труд при поиске эквивалентов позволяет достигать наиболее высокого качества лексикона, однако, требует огромных затрат времени и усилий. Поэтому разработка качественного лексикографического ресурса для реальных приложений неподъемна для одного человека, и осуществляется в результате совместного труда коллектива лингвистов. Это, в свою очередь, является еще одним источником проблемы: в силу вариативности языковых средств даже в пределах узкой предметной области возможны различные *корректные* решения при выборе иноязычных эквивалентов одних и тех же наименований как одним человеком, так и, тем более, различными людьми, задействованными в разработке лексикона. Это влечет за собой проблему соблюдения единства терминологии и требует специальных усилий для осуществления управления терминологией.

1. Методология

1.1. Мотивация

В нашем исследовании мы принимаем в качестве исходного положения необходимость включения многокомпонентной лексики в лексикон для многоязычных систем АОТ; это, в свою очередь, ведет к допущению, что одни и те же лексемы могут встречаться в составе различных вхождений-словосочетаний лексикона. При этом использование синонимичных номинаций, допустимых даже в узкой предметной области, потенциально ведет к проблеме нарушения принципа единства терминологии в лексиконе, и как итог, в результатах работы системы АОТ, включающей данный лексикон в свою базу знаний. Особенно остро эта проблема встает при разработке лексикографического ресурса коллективом лингвистов.

Действительно, пилотное исследование подтвердило тезис о нарушении единообразия при коллективном переводе одних и тех же компонентов в составе различных многокомпонентных терминов. Мы рассмотрели лексикон именной терминологии, извлеченной из русскоязычного корпуса научно-технических текстов по машиностроению, с ее переводом, выполненным согласно методике, предложенной в [5] студентами во время прохождения переводческой практики. Данная методика верификации переводных эквивалентов профессиональной лексики предполагает использование документов сети Интернет в качестве лингвистического корпуса текстов. Из полученных пар переводных эквивалентов были отобраны именные группы, включающие в свой состав лексический компонент «выбросы вредных веществ», который встретился в восьми вхождениях лексикона и был переведен в пределах этих именных групп пятью разными способами. Все предложенные эквиваленты проходят проверку по корпусу в Интернет с положительным результатом. Однако, в случае если данный компонент является ключевым в некотором тексте, вероятность встретить его в различных комбинациях (в том числе, таких, которые представлены в лексиконе в более длинных многокомпонентных сочетаниях) достаточно высока. А значит система АОТ, опираясь на данные лексикона, будет предлагать (в том числе, в пределах одного и того же текста) различные эквиваленты в зависимости от того, в комбинации с какими другими лексическими единицами встретился данный лексический компонент. Для преодоления этого недостатка требуется механизм управления терминологией при подготовке лексикона.

1.2. Формализованная процедура верификации переводных эквивалентов для вхождений лексикона с общим лексическим компонентом

Принимая все достоинства методики поиска переводных эквивалентов для специальной лексики, предложенной в [5], мы отмечаем, что в данной

методике не рассматривается вопрос о соблюдении единства терминологии в случае коллективной работы над двуязычным лексиконом. При этом не вызывает сомнений, что разработка полноценных лексикографических ресурсов не может осуществляться одним человеком – такие ресурсы создаются в результате коллективной работы лингвистов.

В связи с этим мы полагаем, что при коллективной работе над лексиконом следует ввести этап верификации переводов, целью которой будет унификация предлагаемых для вхождений лексикона переводов.

Предлагаемая нами процедура верификации перевода для компонента именной группы t, встречающегося в составе нескольких многокомпонентных именных групп и допускающего множественную манифестацию на языке перевода, включает следующие этапы:

1. Отбор из двуязычного лексикона именных групп с общим лексическим компонентом и соответствующими переводами.
2. Построение иерархии именных групп (от более длинных к более коротким).
3. Формирование перечня кандидатов для перевода общего лексического компонента.
4. Проверка отобранных кандидатов и, при необходимости, корректировка переводов в лексиконе.

Этап отбора включает поиск именных вхождений лексикона, содержащих целевую подстроку t на языке оригинала. Все вхождения, удовлетворяющие этому условию, и соответствующие им переводы заносятся во множество V верифицируемых именных групп. Элементы этого множества $e^i \in V$ представляют собой манифестацию вхождения лексикона в форме эквивалентных лексических единиц на двух языках: $e^i = (luS^i, luT^i)$, $i = \overline{1, M}$, где luS^i – i-я лексическая единица на языке оригинала, содержащая в качестве своего компонента термин t, luT^i – переводной эквивалент для luS^i, M – количество вхождений лексикона, содержащих целевой термин t.

Построение иерархии – достаточно тривиальная процедура, которая проводится за 2 шага: 1) формирование гнезд вложенных друг в друга лексических единиц из лексикона (гнездо включает максимально длинную лексическую единицу из лексикона, а также все имеющиеся в лексиконе вхождения, полностью вложенные в данную лексическую единицу; при этом наиболее длинное из многокомпонентных вхождений является репрезентантом этого гнезда); 2) распределение образованных гнезд вхождений на классы-множества C_n по количеству лексических компонентов $n = \overline{1, N}$ в самой длинной лексической единице на языке оригинала в гнезде. В итоге определяется вектор из множеств $C = (C_N, \dots, C_1)$, где C_n включает множество сформированных гнезд, каждое из которых представлено репрезентантами этого гнезда: $C_n = \{e_k^n = (luS_k^n, luT_k^n)\}, k = \overline{1, K_n}$, где K_n – количество гнезд, вошедших в состав C_n. Множества в векторе C упорядочены по убыванию количества компонентов n в элементах этого множества.

Формирование перечня кандидатов осуществляется на основе извлечения предложенных способов перевода целевого термина t из всех переводных эквивалентов luT^i, $i = \overline{1, M}$ единиц лексикона. Итогом данного этапа является формирование множества $CTE = \{cte_j\}, j = \overline{1, P}$, где cte_j – j-й переводной эквивалент для целевого термина t из P допустимых эквивалентов, извлеченных из лексикона.

Проверка и корректировка отобранных кандидатов включает несколько шагов.

На первом шаге последовательно выполняем подстановку всех кандидатов cte_j в каждый переводной эквивалент luT_k^N вхождений-репрезентантов из класса C_N. Трансформированные таким образом переводные эквиваленты проверяются на корректность в корпусе текстов в сети Интернет. В результате, возможно выяснить, сколько вхождений допускает при переводе замену на каждый из кандидатов множества CTE. При этом все переводы, допустимые для вхождения-репрезентанта гнезда, автоматически принимаются совместимыми и со всеми иерархически подчиненными вхождениями этого гнезда. В случае, если ни одного допустимого эквивалента для репрезентанта не найдено, в качестве репрезентанта принимается следующее (в порядке убывания количества компонентов) вхождение в гнезде, и процедура проверки допустимости перевода повторяется для нового репрезентанта гнезда. Полученные количества (вхождений, для которых данный кандидат является допустимым переводом), а также показатели частотности употребления каждого из допустимых кандидатов для перевода вхождений, можно рассматривать как весовые коэффициенты. Далее, ранжировав список кандидатов cte_j по убыванию значений весовых коэффициентов, мы получаем упорядоченный набор кандидатов для перевода целевого термина t. В итоговый набор терминов, которые должны сохраниться в лексиконе после прохождения процедуры верификации, последовательно включаются варианты перевода из ранжированного списка до тех пор, пока в нем не окажется минимальное необходимое количество переводных эквивалентов luT_k^N, достаточное, чтобы перевести все вхождения-репрезентанты из класса C_N. В предельном случае, минимальный набор переводных эквивалентов будет содержать единственный вариант перевода, допустимый для всех вхождений в пределах класса C_N. Этот вариант перевода ассоциируется со всем вхождениями из гнезд класса C_N. Если для перевода вхождения недостаточно одного варианта, то для оставшихся вхождений, не ассоциированных с наилучшим

переводом, предпринимается попытка ассоциировать эти вхождения со вторым по рангу переводом. И т.д.

Следующие шаг – последовательное рассмотрение остальных множеств C_n в порядке убывания n. Гипотеза состоит в том, что найденный на предыдущем шаге минимальный набор переводных эквивалентов достаточен для перевода всех вхождений множества C_n. Кандидаты на перевод cte_j рассматриваются в порядке, определенном ранжированием на предыдущем шаге, начиная с первого по рангу. Последовательно для каждого luT_k^n осуществляется подстановка очередного варианта перевода cte_j и проверка полученного варианта перевода в Интернет. Если наилучший перевод термина t допустим для репрезентанта гнезда, то он ассоциируется со всеми лексическими единицами этого гнезда. Если не допустим, производится попытка подстановки следующего по рангу варианта перевода с последующей проверкой результата, и так далее до тех пор, пока не будет найден допустимый эквивалент для данного гнезда. Тогда этот эквивалент ассоциируется со всеми вхождениями в гнезде. В случае, если найденный допустимый эквивалент не входил в минимальный набор переводных эквивалентов, он добавляется в набор.

В итоге выполнения предыдущего шага каждое вхождение из лексикона, включающее целевой термин t, ассоциировано с одним вариантом его перевода, причем количество этих вариантов сокращается до минимально необходимого для обслуживания данной предметной области. Этот минимальный набор допустимых эквивалентов на последнем шаге процедуры принимается за перечень эквивалентов собственно целевого термина t, который может сохраняться в лексиконе.

2. Результаты

Следуя представленному алгоритму, для словосочетаний с целевым термином «*выбросы вредных веществ*» были составлены следующие классы с гнездами (для краткости указаны лишь вхождения на языке оригинала; элементы, составляющие гнездо более чем из одного элемента, заключены в круглые скобки; репрезентант гнезда указан в круглых скобках первым):

$C_1 =$ {*выбросы вредных веществе*}

$C_2 =$ {*выброс вредных веществ автотракторными двигателями, повышенный выброс вредных веществ, норма выбросов вредных веществ*}

$C_3 =$ {(*снижение выбросов вредных веществ с отработавшими газами дизеля, выброс вредных веществ с отработавшими газами, снижение выбросов вредных веществ*), *зависимость удельных выбросов вредных веществ*}

Множество кандидатов переводов включает:

$CTE =$ {*emission of hazardous substances*, *emission of harmful substances, hazardous substances emission*, *harmful substances emission, release of harmful agents*}

Подстановка кандидатов из CTE в репрезентант первого элемента множества C_3 при проверке по корпусу не дает ни одного точного совпадения. Замена репрезентанта в данном гнезде на второй элемент по списку дает совпадения с 3 элементами из множества CTE, причем при проверке в поисковике Google перевод *emission of harmful substances* в составе репрезентанта употребляется в более чем трехстах документах, в противовес единичным случаям употребления переводных эквивалентов *emission of hazardous substances* и *harmful substances emission.* Наиболее частотный из указанных переводов является допустимым и для второго элемента множества C_3.

Проверяя найденный частотный кандидат *emission of harmful substances* для элементов множества C_2, находим, что он является допустимым для первых двух вхождений из списка и имеет наибольший вес по сравнению с другими эквивалентами. Однако для вхождения *норма выбросов вредных веществ* единственным верифицируемым по корпусу переводом является *harmful substances emission standard*. Заметим, что кандидаты перевода *emission of harmful substances* и *harmful substances emission* являются трансформационными вариантами друг друга. При этом использование второго варианта значительно реже и строго контекстуально предопределено. Поэтому, используя принцип частотности (по результатам анализа поисковых выдач по запросам к поисковой системе Google), мы можем заключить, что вероятность встретить в текстах исследуемой предметной области вариант *emission of harmful substances* более высока, поэтому данный кандидат более предпочтителен в качестве перевода для собственно целевого термина *выбросы вредных веществе* (множество C_1) в лексиконе.

Таким образом, следуя нашей формализованной процедуре, для перевода лексической единицы *выбросы вредных веществе* в пределах рассматриваемой предметной области достаточно сохранение в лексиконе для системы АОТ единственного эквивалента *emission of harmful substances* как для перевода целевого термина, так и практически для всех вхождений (за исключением одного трансформационного варианта в одном из контекстов), в которые данная единица входит в качестве компонента.

Для подтверждения этого вывода нами проведен семантический анализ первых страниц поисковых выдач по запросам к системе Google, сформированным по элементам множества CTE. Контекстный анализ показывает, что кандидат *release of harmful agents* используется лишь в текстах медико-биологической направленности. Эквиваленты, содержащие в себе компонент *hazardous*, обозначают риск немедленного пагубного эффекта на здоровье и, часто, связаны с радиоак-

тивной угрозой. Эти данные, а также результаты дефиниционного анализа по толковому словарю[2] позволяют заключить, что указанные эквиваленты гипонимичны более общему термину *emission of harmful substances*, что обусловливает возможность использования последнего в более широком спектре контекстов.

Аналогичное исследование мы провели для еще трех целевых терминов из лексикона: «массо-геометрическая характеристика» (варианты перевода в исходном лексиконе: *mass and geometric characteristics, mass and geometrical characteristics, mass-geometric characteristics, mass geometry characteristics, mass and geometric features*), «пневмозатвор» (варианты перевода: *pneumoshutter, pneumatic seal, pressure lock, pneumatic valve*), «механизм поворота» (варианты перевода: *steering system, steering mechanism, steering device, pivot mechanism, mechanism of rotation, steering group*). Итоги выполнения этапов процедуры позволили определить следующие допустимые в исследуемой предметной области пары эквивалентов для указанных целевых терминов и вхождений, в состав которых они входят:

массо-геометрическая характеристика: mass and geometric characteristics
пневмозатвор: pressure lock
механизм поворота: steering system

Таким образом, в каждом случае нам удалось свести количество переводных эквивалентов к одному допустимому термину на языке перевода. При этом результаты контекстно-семантического анализа кандидатов согласуются с итогами отбора оптимального переводного эквивалента, полученными в ходе исполнения шагов предложенной нами формализованной процедуры. Выполненная в соответствии с полученными унифицированными переводами корректировка лексикона позволяет гарантировать единообразие переводов при поиске переводных эквивалентов целевых терминов в лексиконе.

3. Обсуждение

Предложенная процедура верификации коллективных переводов формальна и допускает автоматизацию. В нашем исследовании этапы, связанные с проверкой корректности переводов по корпусу текстов в Интернет, выполнялись вручную для обеспечения высокого качества конечного результата. Однако, учитывая применение частотного принципа при отборе и принципиальную возможность автоматического парсинга страниц выдачи поисковика с целью выявления наличия точных совпадений в аннотациях найденных страниц, мы полагаем, что и этот этап может быть автоматизирован путем введения механизма весовых коэффициентов и пороговых значений для оценки результатов поисковой выдачи. Представляется, что такие коэффициенты должны иметь относительный характер. При этом, видимо, следует быть готовым к тому, что автоматическая проверка не будут гарантировать абсолютно точный результат. Однако, вопрос вызывает лишь степень неточности и ее допустимость в угоду возможности в значительной степени ускорить этап верификации переводных эквивалентов лексикона. Этот вопрос требует более детальной разработки и экспериментальной проверки.

Полезным побочным эффектом предложенной нами процедуры является выявление лакун в лексиконе в результате формирования гнезд вложенных друг в друга лексических единиц. Отсутствие в лексиконе некоторых грамматически «оправданных» компонентов более длинных единиц может стать стимулом для дальнейшей корректировки лексикона в случае, если отсутствующий компонент обладает достаточной значимостью. Безусловно, критерии значимости должны быть четко определены. Думается, что с общей стратегией исследования согласуется использование в качестве таких критериев метрик из теории информационного поиска.

В силу того, что выборка для экспериментальной проверки работы формализованной процедуры, была достаточно небольшой, нам не удалось найти терминологию, которая не только допускала, но и требовала бы различных переводов для одних и тех же компонентов разных вхождений. Видимо, если такие случаи и есть, они связаны с омонимией терминов. Однако в пределах одной предметной области такое явление значительно более редкое, чем в общеупотребительном языке. В случае наличия омонимии возможная проблема, которая нам видится, это принятие решения о том, насколько необходимо оставлять все омонимичные переводы в лексиконе для целевого термина (то есть термина в отсутствие разрешающих его многозначность контекстов). Тезис о том, что лишь частотных характеристик достаточно для принятия такого решения, представляется сомнительным. В частности, при проверке эквивалентов целевого термина «массо-геометрические характеристики» наиболее частотные варианты перевода *mass and geometrical characteristics* и *mass and geometric characteristics* при проверке по корпусу оказались примерно сопоставимы по употребительности. При этом, очевидно, что различие в двух переводах заключается лишь в использовании морфологического варианта слова *geometric(al)*. Сохранение обоих эквивалентов в лексиконе избыточно, и, скорее, приведет к большей путанице и нарушению единства терминологии при использовании в системах АОТ. В данном случае, мы решили этот вопрос, выполнив диахронический анализ употребления двух вариантов с помощью онлайн сервиса Google Books Ngram

[2] Нами использовался Collins Dictionary Online: http://www.collinsdictionary.com/

Viewer[3] и выявив устойчивый рост в употребительности варианта *geometric*, в то время как для варианта *geometrical* с 1960-х гг. прошлого века наблюдается спад. В итоге, в современном языке частотность употребления варианта *geometric* приблизительное вдвое выше варианта *geometrical.* Этими соображениями, в итоге, был обусловлен окончательный выбор эквивалента для указанного целевого термина в пользу *mass and geometric characteristics.*

Заключение

В ходе проведенного исследования нами была предложена формализованная процедура верификации переводных эквивалентов двуязычного лексикона для решения проблемы обеспечения соблюдения принципа единства терминологии, которая возникает при коллективной работе над составлением переводного словаря. Экспериментальная проверка работы формализованной процедуры показывает, что для произвольно отобранных лексических единиц языка оригинала возможно сведение синонимичных номинаций их переводных эквивалентов до (часто) единственной лексической единицы на языке перевода. Дополнительный этап корректировки компьютерного лексикона в соответствии с предложенной процедурой позволяет обеспечить более качественную работу системы автоматической обработки текстов, включающих такой лексикон в свою базу знаний.

Литература

1. Бабина, О.И. Построение базы лингвистических знаний для многоязычных систем автоматической обработки текстов / О.И. Бабина // Лингвистика в контексте культуры: материалы V международной научно-практической конференции (Челябинск, 28–30 ноября 2012 г.) / под общ. ред. Е.В. Харченко. – Челябинск: Издательский центр ЮУрГУ, 2012. – С. 19–23.

2. Бабина, О.И. Построение модели извлечения информации из технических текстов: дис. ... канд. филол. наук / О.И. Бабина. – Челябинск, 2006. – 235 с.

3. Залевская, А.А. Слово в лексиконе человека: психолингвистическое исследование / А.А. Залевская. – Воронеж: Изд-во Воронеж. гос. ун-та, 1990. – 208 с.

4. Хроленко, А.Т. Современные информационные технологии для гуманитария: практическое руководство / А.Т. Хроленко, А.В. Денисов. – М.: Флинта : Наука, 2007. – 128 с.

5. Шереметьева, С.О. К вопросу об электронных ресурсах профессиональной лексики / С.О. Шереметьева, П.Г. Осминин, Е.С. Щербаков // Вестник ЮУрГУ. Серия «Лингвистика». – 2014. – Т. 11, № 1. – С. 57–63.

6. A Contrastive Approach to Multi-Word Term Extraction from Domain Corpora / F. Bonin, F. Dell'Orletta, G. Venturi, S. Montemagni // Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010. – Valetta, 2010. – P. 3222–3229.

7. Improving Word Translation Disambiguation by Capturing Multiword Expressions with Dictionaries / L. Bungum, B. Gambäck, A. Lynum, E. Marsi // Proceeding of the 9th Workshop on Multiword Expressions, NAACL-HLT 2013. – Atlanta, 2013. – P. 21–30.

8. Chen, J. A Multi-Word Term Extraction System / J. Chen, C.-H. Yeh, R. Chau // Proceedings of the 9th Pacific Rim International Conference on Artificial Intelligence. – Guilin, 2006. – P. 1160–1165. – (Volume 4099 of the series Lecture Notes in Computer Science).

9. Fung, P. A Statistical View on Bilingual Lexicon Extraction: From Parallel Corpora to Non-Parallel Corpora / P. Fung // Lecture Notes in Computer Science. – 2002. – Vol. 1529. – P. 1–17.

10. Garera, N. Improving Translation Lexicon Induction from Monolingual Corpora via Dependency Contexts and Part-of-Speech Equivalences / N. Garera, C. Callison-Burch, D. Yarowsky // Proceedings of the Thirteenth Conference on Computational Natural Language Learning. – Boulder, 2009. – P. 129–137.

11. Hadni, M. Multi-word Term Extraction based on New Hybrid Approach for Arabic Language / M. Hadni, A. Lachkar, S. el Alaoui Ouatik // Proceedings of the Second International Conference on Computational Science and Engineering (CSE-2014). – Dubai, 2014. – P. 109–120.

12. Harris, Z.S. Distributional Structure / Z.S. Harris // Word. – 1954. – Vol. 10, No. 2–3. – P. 146–162.

13. Finlayson, M.A. Detecting Multi-Word Expressions Improves Word Sense Disambiguation / M.A. Finlayson, N. Kulkarni // Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World (MWE 2011). – Stroudsburg, 2011. – P. 20–24.

14. Rapp, R. Automatic Identification of Word Translations from Unrelated English and German Corpora / R. Rapp // Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics. – Stroudsburg, 1999. – P. 519–526.

15. Multiword Expressions: A Pain in the Neck for NLP / I.A. Sag, T. Baldwin, F. Bond, A.A. Copestake // Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing (CICLing 2002). – Mexico City, 2002. – P. 1–15.

16. Sheremetyeva, S. On Extracting Multiword NP Terminology for MT / S. Sheremetyeva // Proceedings of the 13th Annual Conference of the EAMT. – Barcelona, 2009. – P. 205–212.

17. Shudo, K. A Comprehensive Dictionary of Multiword Expressions / K. Shudo, A. Kurahone,

[3] https://books.google.com/ngrams

T. Tanabe // Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL–HLT 2011). – Vol. 1. – Portland, 2011. – P. 161–170.

18. Váradi, T. 2006. Multiword Units in an MT Lexicon / T. Váradi // Proceedings of the Workshop on Multi-Word Expressions in a Multilingual Context. – Trento, 2006. – P. 73–78.

Бабина Ольга Ивановна, кандидат филологических наук, доцент, доцент кафедры лингвистики и перевода, Южно-Уральский государственный университет (Челябинск), babinaoi@susu.ru

Тамгина Екатерина Сергеевна, студент Института лингвистики и международных коммуникаций, Южно-Уральский государственный университет (Челябинск), rinka1811@mail.ru

Поступила в редакцию 31 октября 2016 г.

DOI: 10.14529/ling160403

UNIFICATION OF MULTI-WORD EXPRESSION TRANSLATIONS DURING COLLABORATIVE COMPILATION OF BILINGUAL LEXICONS

O.I. Babina, *babinaoi@susu.ru*
E.S. Tamgina, *rinka1811@mail.ru*
South Ural State University, Chelyabinsk, Russian Federation

This paper addresses the collaborative compilation of a bilingual lexicon used as knowledge base for a natural language processing (NLP) system. A pilot experiment confirmed the inevitable variability of corpus-relevant translation equivalents offered by different translators for the same lexical units. The authors suggest the introduction of an additional step to unify translation equivalents for multi-word units in the lexicon compilation process to ensure terminological consistency of the NLP-system. A formalized procedure has been developed to perform this step. An experimental evaluation of the developed procedure based on a sample of target terms allowed the authors to conclude that it is possible to minimize the quantity of translation equivalents for components of multi-word expressions. Lexicon entry correction using the results of the suggested procedure enables unification of the entries' lexical manifestations and, in the long term, can provide for higher quality NLP-systems.

Keywords: multi-word expressions, collaborative lexicon compilation, bilingual lexicon, term management, natural language processing.

References

1. Babina O.I. *Postroenie bazy lingvisticheskih znaniy dlya mnogoyazyichnyih sistem avtomaticheskoy obrabotki tekstov* [Building a Linguistic Knowledge Base for a Multilingual Natural Language Processing System], Lingvistika v kontekste kultury: materialy V mezhdunarodnoy nauchno-prakticheskoy konferentsii [Linguistics in Cultural Context: Proceedings of the 5th International Conference]. Chelyabinsk, South Ural St. Univ. Publ., 2012, pp. 19–23.

2. Babina O.I. *Postroenie modeli izvlecheniya informatsii iz tehnicheskih tekstov* [Constructing a Model for Information Extraction from Technical Texts]: PhD thesis, Chelyabinsk, Russia, 2006, 235 p.

3. Zalevskaya A. A. *Slovo v leksikone cheloveka: psiholingvisticheskoe issledovanie* [The Word in a Human Lexicon: Philological Study], Voronezh, Izdatelstvo Voronezhskogo Gosudarstvennogo Universiteta, 1990, 208 p.

4. Khrolenko A.T., Denisov A.V. *Sovremennyie informatsionnyie tehnologii dlya gumanitariya: prakticheskoe rukovodstvo* [Modern Information Technology for a Humanities-Minded Person], Moscow, Flinta, Nauka, 2007, 128 p.

5. Sheremeteva S.O., Osminin P.G., Scherbakov E.S. On Electronic Recourses for Professional Lexicon. *Bulletin of the South Ural State University. Ser. Linguistics*, 2014, vol. 11, no. 1, pp. 57–63.

6. Bonin F., Dell'Orletta F., Venturi G., Montemagni S. A Contrastive Approach to Multi-Word Term Extraction from Domain Corpora. *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010*, Valletta, Malta, 2010, pp. 3222–3229.

7. Bungum L., Gambäck B., Lynum A., Marsi E. Improving Word Translation Disambiguation by Capturing Multiword Expressions with Dictionaries, *Proceeding of the NAACL-HLT 9th Workshop on Multiword Expressions,* Atlanta, Georgia, USA, 2013, pp. 21–30.

8. Chen J., Yeh C.-H., Chau R. A Multi-Word Term Extraction System, *Proceedings of the 9th Pacific Rim International Conference on Artificial Intelligence*, Guilin, China, 2006, August, Vol. 4099 of the series Lecture Notes in Computer Science, pp. 1160–1165.

9. Fung P. A Statistical View on Bilingual Lexicon Extraction: From Parallel Corpora to Non-Parallel Corpora, *Lecture Notes in Computer Science*, 2002, vol. 1529, pp. 1–17.

10. Garera N., Callison-Burch C., Yarowsky D. Improving Translation Lexicon Induction from Monolingual Corpora via Dependency Contexts and Part-of-Speech Equivalences, Proceedings of the Thirteenth Conference on Computational Natural Language Learning, Boulder, Colorado, USA, 2009, pp. 129–137.

11. Hadni M, Lachkar A., Alaoui Ouatik S. el. Multi-Word Term Extraction based on New Hybrid Approach for Arabic Language, Proceedings of the Second International Conference on Computational Science and Engineering (CSE-2014), Dubai, UAE, 2014, April, pp. 109–120.

12. Harris Z. S. Distributional Structure, Word, 1954, Vol. 10, no. 2–3, pp. 146–162.

13. Finlayson M.A., Kulkarni N. Detecting Multi-Word Expressions Improves Word Sense Disambiguation, Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World (MWE 2011), Stroudsburg, PA, USA, 2011, pp. 20–24.

14. Rapp R. Automatic Identification of Word Translations from Unrelated English and German Corpora, Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics, Stroudsburg, PA, USA, 1999, pp. 519–526.

15. Multiword Expressions: A Pain in the Neck for NLP, Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing (CICLing 2002), Mexico City, Mexico, 2002, pp. 1–15.

16. Sheremetyeva S. On Extracting Multiword NP Terminology for MT, Proceedings of the 13th Annual Conference of the EAMT, Barcelona, Spain, 2009, May, pp. 205–212.

17. Shudo K., Kurahone A., Tanabe T. A Comprehensive Dictionary of Multiword Expressions, Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL–HLT 2011), Portland, Oregon, USA, 2011, pp. 161–170.

18. Váradi T. Multiword Units in an MT Lexicon, Proceedings of the Workshop on Multi-Word Expressions in a Multilingual Context, Trento, Italy, 2006, pp. 73–78.

Olga I. Babina, Candidate of Philology (PhD), Associate Professor, Associate Professor of the Department of Linguistics and Translation, South Ural State University (Chelyabinsk), babinaoi@susu.ru

Ekaterina S. Tamgina, undergraduate student of the Institute of Linguistics and International Communication, South Ural State University (Chelyabinsk), rinka1811@mail.ru

Received 31 October 2016

ОБРАЗЕЦ ЦИТИРОВАНИЯ

Бабина, О.И. Унификация переводных эквивалентов многокомпонентной лексики при коллективной разработке двуязычных лексиконов / О.И. Бабина, Е.С. Тамгина // Вестник ЮУрГУ. Серия «Лингвистика». – 2016. – Т. 13, № 4. – С. 15–22. DOI: 10.14529/ling160403

FOR CITATION

Babina O.I., Tamgina E.S. Unification of Multi-Word Expression Translations During Collaborative Compilation of Bilingual Lexicons. *Bulletin of the South Ural State University. Ser. Linguistics*. 2016, vol. 13, no. 4, pp. 15–22. (in Russ.). DOI: 10.14529/ling160403