Н.В. Лукашевич, А.А. Герасимова

ОПРЕДЕЛЕНИЕ УСТОЙЧИВЫХ СЛОВОСОЧЕТАНИЙ МЕТОДОМ АССОЦИАТИВНОГО ЭКСПЕРИМЕНТА

Федеральное государственное бюджетное образовательное учреждение высшего образования «Московский государственный университет имени М.В. Ломоносова»
119991, г. Москва, Ленинские горы, 1
Государственное научное бюджетное учреждения
«Академия наук Республики Татарстан»
420111, г. Казань, ул. Баумана, 20

Статья[1] посвящена исследованию косвенных признаков устойчивости словосочетания, которые можно обнаружить при помощи ассоциативных экспериментов. В результате сбора ассоциаций к словосочетаниям и их компонентам было установлено два типа нерегулярности устойчивых словосочетаний: во-первых, устойчивость проявляется в том, что слова-компоненты словосочетания становятся частотными ассоциациями друг к другу; во-вторых, для устойчивого словосочетания характерно относительно малое разнообразие ассоциаций, которое измеряется путем вычисления энтропии, а также малое количество совпадений ассоциаций к словосочетанию и к словам-компонентам. Использование перечисленных признаков нерегулярности позволяет с высокой точностью определить, является ли словосочетание устойчивым и нужно ли включать его в тезаурусы для автоматической обработки текстов.

Ключевые слова: устойчивые словосочетания; ассоциация; нерегулярность; тезаурус; русский язык.

1. Введение. В настоящее время одним из видов востребованных компьютерных ресурсов для автоматической обработки текстов, анализа документов являются так называемые тезаурусы, представляющие лексические или терминологические системы в форме семантических сетей, которые описывают формализованные отношения между значениями слов (словосочетаний). Известнейшим лексическим ресурсом такого рода является тезаурус английского

Лукашевич Наталья Валентиновна — доктор технических наук; ведущий научный сотрудник МГУ имени М.В. Ломоносова; ведущий научный сотрудник Академии наук Республики Татарстан (e-mail: louk_nat@mail.ru).

Герасимова Анастасия Алексеевна — студентка магистратуры отделения теоретической и прикладной лингвистики филологического факультета МГУ имени М.В. Ломоносова; младший научный сотрудник Академии наук Республики Татарстан (e-mail: anastasiagerasimova432@gmail.com).

[1] Исследование выполнено в рамках проекта РНФ № 16-18-02074.

языка WordNet [Miller, 1998]. Для русского языка существует аналог WordNet — тезаурус RuWordNet[2] [Loukachevitch et al., 2016]. В конкретных предметных областях создаются специализированные тезаурусы, например, медицинский тезаурус MESH[3] или тезаурус в авиационно-космической области NASA[4]. Тезаурус русского языка РуТез включает как лексическую информацию, так и термины широкой общественно-политической области (экономика, финансы, политика и др.) [Лукашевич, 2011]. Тезаурусы используются для различных приложений автоматической обработки текстов: автоматического определения семантического сходства слов и фрагментов текстов, автоматического реферирования текстов, информационного поиска и др.

Существенной проблемой при ведении таких тезаурусов является вопрос о том, какие словосочетания должны включаться в качестве отдельных единиц, каковы критерии включения. Например, стандарты по ведению информационно-поисковых тезаурусов обычно содержат разделы с описанием принципов включения многословных терминов в состав создаваемых тезаурусов [ANSI/NISO, 2005]. Критерии включения словосочетаний обсуждаются и для тезаурусов типа WordNet [Maziarz et al., 2015; Piasecki et al., 2015; Vincze, Almasi, 2014]. Изначально в синонимические ряды тезауруса WordNet (синсеты) предлагалось включать только «лексикализованные» выражения [Miller, 1998], однако границы лексикализации оказалось трудно определить [Agirre et al., 2006]. Авторы работы [Bentivogli, Pianta, 2004] обсуждают необходимость включить в итальянский WordNet и нелексикализованные (композициональные) выражения, называемые *фразеты*, которые имеют лексикализованные соответствия в других языках.

В качестве кандидатов на включение в тезаурусы часто обсуждаются устойчивые выражения, которые охватывают большое множество разнообразных типов словосочетаний: идиомы, составные существительные, имена собственные, фразовые глаголы, конструкции с вспомогательным глаголом и др. [Calzolari et al., 2002; Sag et al., 2002; Baldwin, Kim, 2010]. Необходимость включения в тезаурусы некоторых из перечисленных языковых выражений, например идиом, не вызывает сомнения. Для других типов устойчивых словосочетаний решение задачи о включении в ресурс менее очевидно. В частности, необходимо проанализировать, обладает ли словосочетание нерегулярностью какого-либо типа: лексической, синтаксической, семантической (например, некомпозициональностью значения)

[2] URL: http://ruwordnet.ru/ru/
[3] URL: https://www.ncbi.nlm.nih.gov/mesh
[4] URL: https://www.sti.nasa.gov/thesvol1.pdf

или статистической. Статистической нерегулярностью называют свойство словосочетания, когда частота совместной встречаемости его компонентов выше случайной [Church, Hanks, 1990; Manning, Schьtze, 1999; Farahmand, Nivre, 2015]. Этот вид нерегулярности проявляется еще и в том, что компонент словосочетания не может быть заменен на синоним без изменения значения словосочетания, если же такая замена возможна, то полученное словосочетание встречается значительно реже[5].

Наличие нерегулярности, т.е. неподчинение словосочетания каким-либо языковым правилам (синтаксическим, семантическим) как раз и делает необходимым внесение словосочетания в компьютерные словари. В данной статье нас будут интересовать словосочетания со статистической нерегулярностью, которые не имеют ни лексической, ни семантической нерегулярности, из-за чего бывает трудно решить, нужно ли включать их в компьютерные лингвистические ресурсы.

Цель статьи — с помощью соответствующих диагностик показать два типа проявления нерегулярности устойчивых словосочетаний. Во-первых, мы покажем, что нерегулярность проявляется в том, что компоненты устойчивых словосочетаний часто выступают в качестве ассоциаций друг к другу в ответах участников ассоциативных экспериментов. Во-вторых, оказывается, что устойчивые словосочетания демонстрируют специфику распределения ассоциаций: для таких языковых единиц характерно относительно низкое значение вариативности ассоциаций, а также малое количество пересечений ассоциаций к словосочетанию с ассоциациями к словам-компонентам, т.е. словосочетания имеют специфичные частотные ассоциации, которые отсутствуют у их компонентов.

Указанные свойства устойчивых словосочетаний были выявлены с помощью серии ассоциативных экспериментов на материале русского языка. Как известно, множество ассоциаций образует ассоциативное поле, которое отражает основные лингвистические параметры слова, в том числе его связь с другими слоями значений [Болотнова, 1994; Горошко, 2001].

Статья состоит из раздела 2, который обосновывается необходимость включения словосочетаний в тезаурусы, а также обсуждается понятие устойчивости и виды нерегулярности устойчивых словосочетаний. Раздел 3 посвящен описанию ассоциативных экспериментов. На основании результатов экспериментов выявлены два признака проявления устойчивости словосочетаний. Наконец, раздел 4 представляет собой заключение.

[5] В зарубежной традиции подобную диагностику принято называть «тестом на замещаемость» (*substitutionability test*). Подробнее см. [Sag et al., 2002; Farahmand et al., 2015; Farahmand, Henderson, 2016; Pearce, 2001; Senaldi et al., 2016].

2. Устойчивость словосочетаний в литературе. ***2.1. Понятие устойчивости словосочетаний.*** Имеется большое количество подходов к определению устойчивых словосочетаний. В работе [Архангельский, 1968] выделялись девять способов понимания устойчивости фразеологических единиц. В данном разделе мы рассмотрим два подхода к устойчивости, сформулированные в рамках отечественной традиции.

В модели «Смысл-Текст» устойчивость интерпретируется как высокая вероятность совместной встречаемости компонентов словосочетания [Мельчук, 1960]. И. Мельчук рассматривает *фраземы* (не свободные словосочетания) — словосочетания, составленные из двух или более лексем, связанных регулярным синтаксическим отношением, которые не могут быть заменены на квазисинонимы без изменения значения словосочетания. При этом классификация фразем проводится по двум параметрам: композициональность значения фраземы и природа ограничений, накладываемых на ее структуру (лексические или семантико-лексические ограничения). Соответственно, в работе [Mel'čuk, 2012] выделяются *идиомы*, характеризующиеся некомпозициональным значением и лексическими ограничениями, *коллокации*, значение которых композиционально, но на которые также накладываются лексические ограничения, и *клише*, композициональные выражения с семантико-лексическими ограничениями на структуру.

И. Мельчук выделяет три типа идиом: *полные идиомы*, значение которых не включает значения слов-компонентов; *полуидиомы*, в значение которых входит значение только одного из слов-компонентов и дополнительный элемент значения (*полуфразема* или *коллокация* в терминологии [Иорданская, Мельчук, 2007]); *квазиидиомы*, значение которых включает дополнительный компонент наряду со значениями слов-компонентов словосочетания. Коллокации представляют собой сочетание ключевого слова со словом, представляющим значение лексической функции от этого слова, и могут быть разделены на стандартные и нестандартные. Первые представляют собой результат применения стандартных лексических функций-параметров; для вторых — нестандартных, определяющих закрытое множество коллокаций.

В работе [Баранов, Добровольский, 2008] устойчивость рассматривается в структурном и узуальном аспектах. Структурный аспект характеризует внутреннее устройство языкового выражения, т.е. ограничения, которые накладываются на слова-компоненты. Авторы утверждают, что понимание устойчивости И. Мельчука ограничено и связано только с внутренним устройством словосочетания. Более того, сочетаемостная устойчивость в трактовке И. Мельчука харак-

терна только для словосочетаний, имеющих уникальные компоненты, например словосочетание *точить лясы* устойчиво только по словоформе *лясы*. Предлагаемая в [Баранов, Добровольский, 2008] структурная устойчивость проявляется в наличии ограничений на образование вариантов, морфологической дефектности компонентов словосочетания, а также его синтаксической непроницаемости. Сами авторы делают упор на узуальную устойчивость, которая проявляется в регулярном употреблении словосочетания носителями языка. Если словосочетание узуально не устойчиво, оно не является устойчивым вовсе. При этом узуальная устойчивость есть проявление более общей языковой категории *нерегулярности* — когда при наличии более общего правила используется правило менее общее.

2.2. Типы нерегулярности (идиосинкразии) словосочетаний. В зарубежной традиции устойчивые словосочетания также рассматриваются с точки зрения их нерегулярности разных типов [Calzolari et al., 2002; Sag et al., 2002; Baldwin et al., 2010]. В работе [Baldwin et al., 2010] выделяют лексическую нерегулярность в случае, когда компонент словосочетания не может употребляться ни в каком словосочетании, кроме данного. Синтаксическая нерегулярность проявляется, если словосочетание демонстрирует определенные ограничения на уровне синтаксиса (например, фиксированный порядок слов-компонентов словосочетания). Семантическая нерегулярность есть не что иное, как некомпозициональность значения языкового выражения. Если словосочетание демонстрирует хотя бы одно отклонение из перечисленных, оно представляет собой так называемое *лексикализованное выражение* [Sag et al., 2002; Baldwin et al., 2010].

Нерегулярность может быть и статистической: в этом случае вероятность появления слов-компонентов неслучайна [Church, Hanks, 1990; Manning, Schütze, 1999; Farahmand, Nivre, 2015]. Такое понимание устойчивости во многом совпадает с определением, данным И. Мельчуком. Более того, в работах [Sag et al., 2002; Farahmand et al., 2016] отмечается, что частотность таких словосочетаний выше, чем частотность словосочетаний, полученных в результате замены одного из слов-компонентов на синоним (пр. *прогноз погоды* vs. *предсказание погоды*). Другими словами, при статистической нерегулярности словосочетание демонстрирует узуальную устойчивость в терминах [Баранов, Добровольский, 2008].

Помимо всех названных типов нерегулярности хочется отметить также связи устойчивых словосочетаний с другими единицами лексикона. Этот вопрос особенно актуален в контексте тезаурусного представления множества языковых единиц. Рассмотрим в качестве примера английское устойчивое словосочетание *traffic lights* (‘светофор’). Это словосочетание часто используется в качестве примера

словосочетания с композициональным значением и статистической нерегулярностью: оно обладает высокой частотностью, и его компоненты плохо заменяются на синонимы [Sag et al., 2002].

Заметим, что денотат словосочетания *traffic lights* обладает большим количеством отношений с другими объектами действительности, и эти отношения не следуют из компонентного состава словосочетания: светофор — это объект дорожного строительства, он имеет сигнализирующие элементы, размещается на дорожных участках, используется для регулирования движения. Иными словами, в тезаурусе словосочетание *traffic lights* должно быть связано тезаурусными отношениями с концептами (FACILITIES ('объект строительства'), ROAD ('дорога'), SIGNALS ('сигналы'), REGULATION ('регулирование')) и эти отношения не выводятся из значений слов-компонентов *traffic* ('дорожное движение') и *lights* ('свет').

Рассмотренный пример не единичен и не уникален для английского языка. Словосочетание *пешеходный переход* связано тезаурусными отношениями с концептом безопасности, ресторанный дворик обычно расположен в торговом центре, следовательно, словосочетание *ресторанный дворик* имеет тезаурусное отношение с концептом ТОРГОВЫЙ ЦЕНТР. Таким образом, устойчивые словосочетания без явно выраженной лексической или семантической нерегулярности имеют особый набор семантических отношений с другими языковыми единицами, которые не следуют из их компонентов, обладают *реляционной нерегулярностью*.

2.3. Разметка словосочетаний по устойчивости. В некоторых трудах исследуется языковая интуиция носителей языка по определению устойчивости словосочетаний. Работа [Reddy et al., 2011] посвящена исследованию композициональности значения английских именных групп (например, *application form*, *parking lot*). Авторы предлагают модель предсказания устойчивых словосочетаний на основании суждений о том, в прямом ли значении использовались компоненты словосочетаний. Для построения такой компьютерной модели суждения о композициональности значения словосочетания сравнивались с суждениями о прямом/переносном значении компонентов. В опросе участвовало 30 респондентов. В результате было показано, что оба компонента играют большую роль в определении того, является ли словосочетание устойчивым.

В работе [Ramisch et al., 2016] описывается процедура построения базы данных, в которой содержится информация о суждениях по композициональности значения составных существительных для английского, французского и португальского языков. Респондентов просили оценить, насколько значение словосочетания выводимо из значений слов-компонентов. Авторы утверждают, что подобный ме-

тод косвенного аннотирования словосочетаний, в котором участвуют носители языка, не являющиеся специалистами в лингвистике, надежно предсказывает устойчивость. Тем не менее при таком подходе наблюдалась несогласованность ответов респондентов. Например, носители английского языка по-разному оценили словосочетание *dirty word* (‘бранное слово’) с точки зрения его композициональности/некомпозициональности. Это связано с тем, что в оценке модификатора *dirty* (‘грязный’) были существенные расхождения: для одних носителей прилагательное имеет переносное значение, для других — прямое, но редко употребляющееся.

В [Maziarz et al., 2015] предпринимается попытка формализации критериев внесения устойчивых словосочетаний в lWordNet] (тезаурус типа WordNet для польского языка). Авторы прибегли к экспертной оценке: 14 экспертов-лингвистов должны были классифицировать множество словосочетаний по группам «устойчивое», «неустойчивое», «не знаю». Удалось установить, что для достижения высокого уровня согласованности мнений экспертов при решении задачи включения словосочетания в тезаурус достаточно 5–7 лингвистов.

В наборе данных, представленных в [Farahmand et al., 2015], словосочетания аннотированы по двум признакам: некомпозициональность значения и статистическая нерегулярность словосочетания. В аннотации участвовали пять экспертов. Композициональность значения словосочетания оценивалась по бинарной шкале. Затем все словосочетания с композициональным значением размечались с точки зрения статистической нерегулярности: словосочетание считалось нерегулярным, если ни один из его компонентов не мог быть заменен на квазисинонимы без потери значения. Такое определение нерегулярности приводило к трудностям во время аннотации, поскольку некоторые интуитивно устойчивые словосочетания допускали замену компонента (ср. *floor space* и *floor area* ‘площадь помещения’). Тем не менее в результате разметки было выделено множество статистически устойчивых словосочетаний с композициональным значением: *cable car*, *food court*, *speed limit* и т.д. Полученную базу данных предлагается использовать для обучения систем автоматического распознавания устойчивых словосочетаний.

Как видно, все перечисленные методы разметки устойчивых словосочетаний с помощью экспертной оценки столкнулись с некоторыми проблемами. В данной статье мы хотим найти некоторые косвенные признаки нерегулярности устойчивых словосочетаний, проявляющиеся в языковом поведении носителей языка.

2.4. Идиоматичность словосочетаний и когнитивные эксперименты. В ряде работ рассматривались когнитивные аспекты пред-

ставления устойчивых словосочетаний. В работе [Sinclair, 1991] был сформулирован *принцип идиоматичности*, который заключается в том, что в арсенале говорящего имеются готовые единицы (*полуоформленные фразы* в переводе [Копотев, Стексова, 2016]), которые представляют собой устойчивое целое, хотя могут быть разложены на отдельные составляющие. Принцип идиоматичности находит отражение в физических характеристиках дискурса. Так, в работе [Erman, 2009] исследовалось, как устойчивость словосочетания влияет на длительность хезитационной паузы при выполнении задания на подбор пропущенного компонента словосочетания. Под устойчивостью понималось ограничение на замену компонентов (*restricted exchangeability*) — результат применения теста на статистическую нерегулярность.

Основная гипотеза работы [Erman, 2009] заключалась в том, что подобные устойчивые структуры хранятся в долговременной памяти говорящего в виде целостных единиц. В связи с этим активация в сознании одного компонента при наличии другого происходит с минимальными когнитивными усилиями, что проявляется в меньшей длительности паузы при подборе компонента устойчивого выражения. Гипотеза о целостности устойчивых словосочетаний подтверждается и в исследовании [Rohanian et. al., 2017]. С помощью методики регистрации движения глаз было показано, что характеристики движения взгляда при прочтении устойчивых словосочетаний значимо отличаются по сравнению с чтением свободных словосочетаний.

3. Ассоциативный эксперимент. Для обнаружения проявлений нерегулярности устойчивых словосочетаний была проведена серия ассоциативных экспериментов. Был выбран экспериментальный метод, поскольку существующие словари ассоциаций для русского языка не подходят для решения поставленной в исследовании задачи: данные русского ассоциативного словаря [Karaulov et al., 1994] устарели, а в проекте по сбору ассоциаций Sociation.org отсутствуют рассматриваемые типы словосочетаний.

При отборе материала для экспериментов, а именно устойчивых и неустойчивых словосочетаний, использовался состав словосочетаний в тезаурусе RuTez [Loukachevitch, Dobrov, 2014]. RuTez представляет собой лингвистическую онтологию, разработанную для решения задач автоматической обработки языка.

Лингвистическая онтология RuTez во многом сходна с тезаурусами типа WordNet. Так, понятия, включенные в тезаурусы, отражают значения реальных языковых единиц; репрезентация лексических значений устроена сходным образом. В то же время имеются и су-

щественные различия: в отличие от WordNet в тезаурусе РуТез одно и то же понятие может быть представлено текстовыми входами, которые относятся к разным частям речи; в онтологию РуТез включаются более разнообразные виды словосочетаний; в тезаурусах используются различающиеся наборы концептуальных отношений.

Создатели тезауруса РуТез разработали специальные правила внесения словосочетаний, которые выглядят композициональными, в тезаурус. Словосочетание включается в тезаурус в том случае, если у него имеются некоторые специфические отношения с другими словами или выражениями [Loukachevitch, Lashevich, 2016]. Выделяются случаи, когда словосочетание добавляется в онтологию:

1. Словосочетание является синонимом некоторого слова (*земельный участок — земля*) или словосочетание имеет частотное сокращение (*заработная плата — зарплата*);

2. Словосочетание состоит в отношении синонимии к другому словосочетанию, и это отношение не выводится из значений слов-компонентов словосочетания (*мобильный телефон*, *сотовый телефон*);

3. Словосочетание обобщает значения нескольких слов. Такие словосочетания как *транспортное происшествие* или *учебное заведение* на первый взгляд являются композициональными. Однако они несут важную функцию репрезентации знаний –словосочетания являются гиперонимами по отношению к некоторой группе понятий;

4. Словосочетание имеет тезаурусные отношения, которые не выводятся из значений слов-компонентов. Например, словосочетание *дорожное движение* связано отношениями с другими словосочетаниями, и подобные связи не могут быть выведены из значений слов-компонентов, например связь гипоним/гипероним (*левостороннее движение*, *одностороннее движение*), или ассоциативная связь концептов (*автомобильная авария* и *автомобильная пробка*).

Для эксперимента были отобраны именные группы, которые имеют высокую частотность в текстовой коллекции современных новостей (один миллион документов): 29 словосочетаний «прилагательное + существительное» и 32 словосочетания «существительное + существительное в родительном падеже», которые подразделяются на тезаурусные словосочетания (входят в состав тезауруса РуТез) и нетезаурусные. *Тезаурусная* группа включала 15 словосочетаний: *транспортное происшествие*; *учебное заведение*; *заработная плата*; *программное обеспечение* и т.д. Группа *нетезаурусных словосочетаний* включала такие словосочетания, как: *повышение эффективности*, *крупный размер*, *проект строительства* и т.д.

Для сбора ассоциаций была проведена серия лингвистических экспериментов. В первых двух экспериментах респондентам были представлены отобранные словосочетания (отдельно тезаурусные и нетезаурусные), в третьем — слова-компоненты этих словосочетаний. Респондентам было предложено привести однословную ассоциацию к каждому стимулу. При этом предлагалось не использовать имена собственные (в том числе названия изданий, сайтов, компаний), а также сокращения. Эксперимент проводился на базе сервиса Google Forms.

В результате 26 и 29 респондентов участвовали соответственно в сборе ассоциаций для тезаурусных и нетезаурусных словосочетаний 47 испытуемых приводили ассоциации к словам-компонентам. Несмотря на инструкцию, в некоторых случаях респонденты могли привести только многословную ассоциацию. Подобные случаи были единичными, поэтому они были учтены при обработке результатов. В табл. 1 для некоторых тезаурусных словосочетаний приведены примеры ассоциаций и их частотность.

Таблица 1

Примеры наиболее частотных ассоциаций для некоторых тезаурусных словосочетаний и их компонентов

Стимул	Ассоциации	Частотность
Слово 1: земельный	участок	38
	вопрос	2
Слово 2: участок	земля	11
	дача	11
	полицейский	4
	дорога	3
	дом	2
Словосочетание: земельный участок	дача	12
	дом	2
	надел	2
Слово 1: повышение	должность	8
	зарплата	7
	работа	6
Слово 2: цена	ценник	5
	стоимость	4
	высокая	3
	качество	2
Словосочетание: повышение цен	инфляция	12
	кризис	5
	нефть	2

При обработке ассоциаций для слов и словосочетания были посчитаны следующие характеристики (табл. 2 и 3 для тезаурусных и нетезаурусных словосочетаний соответственно):

1. Энтропия ассоциаций для слов-компонентов и словосочетаний, которая рассчитывалась следующим образом:

$$H = -\sum_{i=1}^{N} p_i \log_b p_i ,$$

где p_i – это вероятность получить некоторый конкретный ответ из множества ответов, которые давались респондентами. Вероятность ответа рассчитывается как отношение его количества по отношению к общему количеству ответов, данных респондентами. Энтропия соответствует разнообразию выбора ассоциаций, который приходится делать отвечающему при назывании слова-ассоциации. Следовательно, чем частотнее некоторые ответы в множестве ассоциативных ответов, тем меньше выбор и меньше энтропия;

2. Количество пересечений между множеством ассоциаций к словосочетанию и множеством ассоциаций к словам-компонентам словосочетания (колонки Ph1 и Ph2 в табл. 2 и 3);

3. Количество случаев, когда один компонент словосочетания выступает в качестве ассоциации к другому компоненту (столбцы A12 и A21 в табл. 2, 3).

Таблица 2

Результаты ассоциативных экспериментов для тезаурусных словосочетаний

Словосочетание	A12	A21	Ph1	Ph2	Энтропия
Транспортное происшествие	0	7	0	6	2,16
Учебное заведение	1	8	**1**	**1**	2,52
Программное обеспечение	**13**	**14**	6	1	2,77
Повышение цен	0	0	**0**	**0**	2,85
Земельный участок	**38**	**13**	0	14	2,89
Квадратный метр	**10**	**20**	0	1	3,22
Электронная почта	**6**	**12**	3	4	3,27
Дорожное движение	0	2	**0**	**0**	3,33
Заработная плата	**18**	**10**	8	2	3,42
Главный герой	5	1	**0**	**3**	3,56
Медицинская помощь	0	5	**0**	**4**	3,58
Торговый центр	**26**	0	1	0	3,62
Лента новостей	**18**	0	1	1	3,79
Мобильный телефон	**26**	**12**	3	7	3,81
Температура воздуха	4	0	6	1	3,81
Среднее значение	11	6,27	1,93	2,93	3,24

Таблица 3

Результаты ассоциативных экспериментов для нетезаурусных словосочетаний

Словосочетание	A12	A21	Ph1	Ph2	Энтропия
Финал лиги	0	0	**2**	**18**	2,16
Начало года	0	0	1	0	2,66
Ежедневный обзор	1	0	**3**	**14**	2,90
Пресс-служба администрации	0	0	**14**	**6**	2,97
Еженедельный обзор	0	0	**6**	**10**	3,19
Необходимый документ	1	0	**0**	**14**	3,64
Конец января	0	0	**2**	**7**	3,81
Должность главы	0	0	**5**	**2**	3,85
Новое поколение	0	4	1	8	3,90
Член совета	5	0	2	5	3,96
Повышение эффективности	0	1	6	6	3,98
Увеличение объема	3	0	7	4	4,02
Крупный размер	4	0	6	4	4,07
Миллион евро	0	0	5	4	4,11
...					
Особое внимание	0	2	0	5	4,63
Председатель комитета	5	0	5	3	4,63
Экономический форум	0	2	2	3	4,65
Интересный комментарий	0	0	2	6	4,69
Среднее значение	**1,22**	**0,36**	**2,80**	**6,36**	**4,07**

Результаты экспериментов показывают, во-первых, что среднее значение энтропии ассоциаций к словосочетаниям существенно выше для нетезаурусных словосочетаний (4,07) по сравнению с тезаурусными словосочетаниями (3,24). Иначе говоря, ассоциации к тезаурусным словосочетаниям более мотивированы. В то же время встречаются и несколько нетезаурусных словосочетаний, которые имеют низкое значение энтропии, например *пресс-служба администрации* (2,97). Это можно объяснить тем, что в языке имеется частотное более длинное словосочетание, частью которого является рассматриваемое словосочетание (*пресс-служба администрации президента*). В результате наиболее частым ответом является именно слово-продолжение исходного словосочетания (*президент*).

Во-вторых, мы видим, что слова-компоненты тезаурусных словосочетаний значительно чаще ассоциируются друг с другом, чем слова-компоненты нетезаурусных словосочетаний. Среднее количество ассоциаций такого типа для компонентов тезаурусных словосочетаний в 10 раз выше, чем для компонентов нетезаурусных словосочетаний. Более того, у восьми из 15 тезаурусных словосо-

четаний оба компонента могут выступать в качестве ассоциации к другому компоненту. Для нетезаурусных словосочетаний таких случаев не зафиксировано.

Мы предполагаем, что взаимные ассоциации компонентов друг к другу как раз и есть проявление того, что такие словосочетания хранятся в ментальном лексиконе как неделимые единицы, как и указывалось в работах [Sinclair, 1991; Ehrman, 2009]. По данному параметру устойчивыми являются словосочетания: *программное обеспечение*; *земельный участок*; *квадратный метр*; *электронная почта*; *заработная плата*; *мобильный телефон*; *лента новостей* и *торговый центр*. Назовем такие словосочетания ***ассоциативно устойчивыми***. Отметим, что именно эти словосочетания соответствуют вышеупомянутому тесту на замещаемость: они имеют высокую частотность в текстовом корпусе, и замена их компонентов на синонимы либо невозможна, либо соответствующие словосочетания встречаются значительно реже.

Кроме того, тезаурусные и нетезарусные словосочетания различаются по количеству совпадений ассоциаций к словосочетаниям с ассоциациями к их компонентам. В среднем у нетезаурусных словосочетаний таких совпадений в четыре раза больше, чем в случае тезаурусных словосочетаний.

Интересно, что словосочетания *дорожное движение* и *повышение цен*, которые кажутся полностью свободными, совсем не имеют пересечений по ассоциациям со своими словами-компонентами и вызывают свой собственный набор ассоциаций с относительно низкой энтропией. Например, *повышение цен* имеет наиболее частые ассоциации *инфляция* (12 ассоциаций из 25) и *кризис* (четыре ассоциации), которые не фигурируют в качестве ассоциаций к словам *повышение* и *цена*. Можно сказать, что такие словосочетания обладают ***реляционной нерегулярностью***, т.е. имеют отношения, которые отсутствуют у их слов-компонентов.

Отметим, что только одно нетезаурусное словосочетание оказалось по всем параметрам похоже на тезаурусное словосочетание, т.е. имело как низкую энтропию ассоциаций, так и малое количество совпадений в ассоциациях слов-компонентов и словосочетания: *начало года*. Самыми частотными ассоциациями для *начало года* являлись *январь* и *сентябрь*. Это связано с месяцами начала календарного и учебного года в России и, таким образом, с этим словосочетанием, действительно, связана важная информация, не замеченная экспертами тезауруса РуТез, но выявленная в ассоциативном эксперименте.

В случае тезаурусных словосочетаний большое количество пересечений между множеством ассоциаций к словосочетанию и мно-

жеством ассоциаций к словам-компонентам было обнаружено для словосочетаний *транспортное происшествие* и *температура воздуха*. Вероятно, это свидетельствует о том, что их статус как устойчивых словосочетаний должен быть пересмотрен.

Итак, низкий уровень энтропии и малое количество одинаковых ассоциаций для словосочетания и его компонентов в большинстве случаев свидетельствуют о том, что словосочетание следует считать устойчивым.

Для практического определения устойчивости словосочетания в ассоциативном эксперименте предлагается использовать *пороговое значение энтропии*, равное 0,8*MaxEntropy, где MaxEntropy — максимальная энтропия словосочетания в ассоциативном эксперименте (при условии равновероятности ответов респондентов). В проведенном нами ассоциативном эксперименте на тезаурусные словосочетания пороговое значение энтропии было равно 3,76, в ассоциативном эксперименте на нетезаурусные словосочетания — 3,89. Соответственно, мы предлагаем считать словосочетания, которые имеют энтропию ниже порогового значения, устойчивыми (включение их в тезаурус обязательно), а словосочетания с энтропией ассоциаций выше порогового значения — неустойчивыми. Таким образом, на основе этого критерия устойчивыми являются словосочетания: *учебное заведение*, *повышение цен*, *дорожное движение*, *главный герой*, *медицинская помощь*.

Итак, по результатам экспериментов на ассоциации мы обнаружили два признака устойчивости словосочетаний:

1. Компоненты устойчивого словосочетания часто ассоциируются друг с другом (словосочетание обладает *ассоциативной нерегулярностью*);

2. Для устойчивого словосочетания характерно малое значение энтропии ассоциаций к словосочетанию (меньше 0,8*MaxEntropy) в сочетании с малым количеством пересечений ассоциаций к словосочетанию и ассоциаций к словам-компонентам (совпадает менее 20% ассоциаций) (словосочетание обладает *реляционной нерегулярностью*).

Таким образом, подтвердилась наша гипотеза о том, что статистическую нерегулярность можно наблюдать с помощью закономерностей, которым подчиняются ассоциации к словосочетанию, а также к его компонентам. При этом три рассмотренных параметра: взаимная ассоциация слов-компонентов; энтропия ассоциаций к словосочетанию: совпадение ассоциаций, — позволяют отличить тезаурусные словосочетания от нетезаурусных с точностью больше 94%.

Чтобы проверить, сохранятся ли выявленные закономерности в случае меньшего количества собранных ассоциаций (соответ-

ственно, в случае меньшего количества респондентов), в каждом эксперименте были отобраны первые 15 ассоциаций. Оказалось, что и при таком количестве материала тезаурусные словосочетания демонстрируют ассоциативную нерегулярность. Что касается реляционной нерегулярности, наблюдались некоторые отклонения от выявленных ранее закономерностей: так, словосочетания *медицинская помощь* и *температура воздуха* имели энтропию ассоциаций выше 0,8*MaxEntropy, а два нетезаурусных словосочетания, *финал лиги* и *начало года*, имели как малую энтропию (ниже 0,8*MaxEntropy), так и малое число совпадений ассоциаций к словосочетанию и ассоциаций к компонентам. В результате при уменьшении объема ассоциаций до 15 оказалось, что тезаурусные словосочетания можно выделить с точностью 92%.

4. Заключение. В настоящей статье мы показали, что устойчивость словосочетания проявляется в наборе закономерностей, которым подчиняются множества ассоциаций носителей языка к словосочетанию и его компонентам. С помощью серии ассоциативных экспериментов мы продемонстрировали два признака устойчивости. Во-первых, словосочетание можно считать устойчивым, если компоненты словосочетания служат ассоциациями друг другу. Именно эти словосочетания имеют высокую частотность в текстовом корпусе и соответствуют тесту на замещаемость: замена их компонентов на синонимы либо невозможна, либо соответствующие словосочетания встречаются значительно реже.

Во-вторых, словосочетание устойчиво, если низкая энтропия ассоциаций к словосочетанию сочетается с малым количеством одинаковых ассоциаций к словосочетанию и словам-компонентам. Все три перечисленных фактора позволяют с высокой точностью установить, является ли словосочетание устойчивым. Отметим, что выявленные пороговые значения требуют проверки на большем объеме данных.

Результаты ассоциативных экспериментов позволяют нам сформулировать ряд рекомендаций касательно усовершенствования других методов разметки устойчивых и неустойчивых словосочетаний:

- при определении устойчивых словосочетаний методом краудсорсинга полезно спрашивать у респондентов ассоциации к словосочетаниям и их компонентам, чтобы выявить реляционную нерегулярность словосочетания;
- в случае принятия решений о включении словосочетания в тезаурус экспертами важно анализировать, имеют ли словосочетания такие семантические отношения, которые не выводятся из значений слов-компонентов.

Список литературы

Архангельский В.Л. Проблема устойчивости фразеологических единиц и их знаковые свойства // Проблемы устойчивости и вариантности фразеологических единиц. Материалы межвузовского симпозиума. Тула, 1968. С. 37–43.

Баранов А.Н., *Добровольский Д.О.* Аспекты теории фразеологии. М., 2008.

Болотнова Н.С. Лексическая структура художественного текста в ассоциативном аспекте. Томск, 1994.

Горошко Е.И. Интегративная модель свободного ассоциативного эксперимента. М., 2001.

Иорданская Л.Н., *Мельчук И.А.* Смысл и сочетаемость в словаре. М.: ЯСК, 2007.

Копотев М., *Стексова Т.* Исключение как правило: переходные единицы в грамматике и словаре. М.: ЯСК, 2016.

Лукашевич Н.В. Тезаурусы в задачах информационного поиска. М., 2011.

Мельчук И.А. О терминах «устойчивость» и «идиоматичность» // Вопросы языкознания. 1960. Т. 4. С. 73–80.

Agirre E., *Aldezabal I.*, *Pociello E.* Lexicalization and multiword expressions in the Basque wordnet // Proceedings of Third International WordNet Conference. 2006. P. 131–138.

Guidelines for the Construction, Format and Management of Monolingual Thesauri. ANSI/NISO. Baltimore, 2005.

Baldwin T., *Kim S.N.* Multiword expressions // Handbook of Natural Language Processing, Second Edition. Chapman and Hall/CRC, 2010. P. 267–292.

Bentivogli L., *Pianta E.* Extending wordnet with syntagmatic information // Proceedings of Second Global WordNet Conference. P. 47–53.

Calzolari N., *Fillmore C.J.*, *Grishman R.*, *Ide N.*, *Lenci A.*, *MacLeod C.*, *Zampolli A.* Towards Best Practice for Multiword Expressions in Computational Lexicons // Proceedings of LREC-2002. P. 1934–1940.

Church K.W., *Hanks P.* Word association norms, mutual information, and lexicography // Computational linguistics. 1990. Vol. 16. №. 1. P. 22–29.

Cordeiro S., *Ramisch C.*, *Idiart M.*, *Villavicencio A.* Predicting the Compositionality of Nominal Compounds: Giving Word Embeddings a Hard Time // Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Vol. 1: Long Papers). Association for Computational Linguistics, 2016a. P. 1986–1997.

Cordeiro S., *Ramisch C.*, *Villavicencio A.* Filtering and measuring the intrinsic quality of human compositionality judgments // ACL 2016. Berlin, 2016b. P. 32–37.

Farahmand M., *Henderson J.* Modeling the non-substitutability of multiword expressions with distributional semantics and a log-linear model // Proceedings of the 12th Workshop on Multiword Expressions. ACL 2016. Berlin, 2016. P. 61–66.

Farahmand M., *Smith A.*, *Nivre J.* A multiword expression data set: Annotating non-compositionality and conventionalization for English

noun compounds // Proceedings of NAACL-HLT. Association for Computational Linguistics, 2015. P. 29–33.
Farahmand M., *Nivre J.* Modeling the statictical idiosyncrasy of multiword expressions // Proceedings of the 11th Workshop on Multiword Expressions. 2015. P. 34–38.
Erman B. Cognitive processes as evidence of the idiom principle. International Journal of Corpus Linguistics. 2007. 12(1). P. 25–53.
Karaulov Yu., *Sorokin Yu.*, *Tarasov E.*, *Ufimtseva N.*, *Cherkasova G.* Russian Association Dictionary. 1994.
Loukachevitch N., *Dobrov B.* RuThes Linguistic Ontology vs.russian Wordnets // Proceedings of Global WordNet Conference GWC-2014. Tartu, 2014. P. 154–162.
Loukachevitch N., *Lashevich G.*, *Gerasimova A.*, *Ivanov V.*, *Dobrov B.* Creating Russian wordnet by conversion // Komp'juternaja Lingvistika i Intellektual'nye Tehnologii. 2016. V. 15. P. 405–415.
Loukachevitch N., *Lashevich G.* Multiword expressions in Russian Thesauri RuThes and RuWordNet // Proceedings of the AINL FRUCT 2016. Saint Petersburgh, 2016. P. 66–71.
Manning C.D., *Schьtze H.* Foundations of statistical natural language processing. MIT Press, 1999.
Maziarz M., *Szpakowicz S.*, *Piasecki M.* A procedural definition of multiword lexical units // Proceedings of Recent Advances in NLP Conference RANLP–2015. P. 427–435.
Mel'čuk I. Phraseology in the language, in the dictionary, and in the computer. Yearbook of Phraseology, 2012. Vol. 3(1). P. 31–56.
Miller G.A. Nouns in wordnet // WordNet: An electronic lexical database. P. 24–45.
Pearce D. Synonymy in collocation extraction // Proceedings of the workshop on WordNet and other lexical resources, Second Meeting of the North American Chapter of the Association for Computational Linguistics. P. 41–46.
Piasecki M., *Wendelberger M.*, *Maziarz M.* Extraction of the multi-word lexical units in the perspective of the wordnet expansion // RANLP-2015. P. 512–520.
Ramisch C., *Cordeiro S.*, *Zilio L.*, *Idiart M.*, *Villavicencio A.*, *Wilkens R.* How naked is the naked truth? A multilingual lexicon of nominal compound compositionality // Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Vol. 2: Short Papers). Association for Computational Linguistics, 2016. P. 114–133.
Reddy S., *McCarthy D.*, *Manandhar S.* An empirical study on compositionality in compound nouns // IJCNLP, 2011. P. 210–218.
Rohanian O., *Taslimipoor S.*, *Yaneva V.*, *Le An Ha.* Using Gaze Data to Predict Multiword Expressions. Proceedings of the International Conference "Recent Advances In Natural Language 2017".
Sinclair J. Corpus, concordance, collocation. Oxford, 1991.
Sag I.A., *Baldwin T.*, *Bond F.*, *Copestake A.*, *Flickinger D.* Multiword expressions: A Pain in the Neck for NLP. Proceedings of International Conference on Intelligent Text Processing and Computational Linguistics, CICLING-2002, Springer Berlin Heidelberg, 2002. P. 1–15.

Senaldi M., *Lebani G.E.*, *Lenci A.* Lexical Variability and Compositionality: Investigating Idiomaticity with Distributional Semantic Models // Proceedings of the 12th Workshop on Multiword Expressions. 2016. P. 21–31.
Vincze V., *Almasi A.* Non-lexicalized concepts in wordnets: A case study of English and Hungarian // Proceedings of Global WordNet, Conference GWC-2014. Global WordNet Association, 2014. P. 118–126.

Natalia V. Loukachevitch, Anastasia A. Gerasimova

DETECTING CONVENTIONALIZED MULTIWORD EXPRESSIONS USING WORD ASSOCIATION EXPERIMENT

Lomonosov Moscow State University
1 Leninskie Gory, Moscow, 119991
Tatarstan Academy of Sciences
20 Bauman St., Kazan, 420111

This paper deals with the types of idiosyncrasy of multiword expressions. In particular, we show that association experiments provide evidence about conventionalization of phrases and help decide on inclusion of such phrases in information-retrieval thesauri. Experimental data shows that conventionalized phrases can be revealed using two factors: 1) component words of a phrase have each other as frequent associations; 2) both entropy of phrase associations and intersection of component word and phrase associations are low. These factors allow predicting conventionalized phrases with high accuracy higher than the existing embedding models have.

Key words: multiword expressions; association; idiosyncrasy; thesaurus; Russian.

About the authors: *Natalia V. Loukachevitch* — Doctor of Technical Sciences; Leading Researcher at Lomonosov Moscow State University; Leading Research Fellow at Tatarstan Academy of Sciences (e-mail: louk_nat@mail.ru); *Anastasia A. Gerasimova* — MA student at the Department of Theoretical and Applied Linguistics of Faculty of Philology, Lomonosov Moscow State University; Junior Research Fellow at Tatarstan Academy of Sciences (e-mail: anastasiagerasimova432@gmail.com).

References

Agirre E., Aldezabal I., Pociello E. Lexicalization and multiword expressions in the Basque wordnet. *Proceedings of Third International WordNet Conference*. 2006, pp. 131–138.
Arkhangel'skii V.L. Problem of fixedness of phraseological units and their significant features. Problem yustoichivosti i variantnosti frazeologicheskikh edinits. *Materialy mezhvuzovskogo simpoziuma*. Tula, 1968, pp. 37–43.

Guidelines for the Construction, Format and Management of Monolingual Thesauri. ANSI/NISO. Baltimore, 2005.
Baldwin T., Kim S.N. Multiword expressions. *Handbook of Natural Language Processing*, second Edition. Chapman and Hall/CRC, 2010, pp. 267–292.
Baranov A.N., Dobrovol'skii D.O. *Aspekty teorii frazeologii* [The aspects of the theory of phraseology]. Moscow, 2008.
Bentivogli L., Pianta E. Extending wordnet with syntagmatic information. *Proceedings of Second Global WordNet Conference*, pp. 47–53.
Bolotnova N.S. *Leksicheskaya struktura khudozhestvennogo teksta v assotsiativnom aspekte* [The lexical structure of literary text in associative aspect]. Tomsk, 1994.
Calzolari N., Fillmore C.J., Grishman R., Ide N., Lenci A., MacLeod C., Zampolli A. Towards Best Practice for Multiword Expressions in Computational Lexicons. *Proceedings of LREC-2002*, pp. 1934–1940.
Church K.W., Hanks P. Word association norms, mutual information, and lexicography.*computational linguistics*. 1990. Vol. 16. №. 1, pp. 22–29.
Cordeiro S., Ramisch C., Idiart M., Villavicencio A. Predicting the Compositionality of Nominal Compounds: Giving Word Embeddings a Hard Time. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Vol. 1: Long Papers). *Association for Computational Linguistics*, 2016a, pp. 1986–1997.
Cordeiro S., Ramisch C., Villavicencio A. Filtering and measuring the intrinsic quality of human compositionality judgments. *ACL 2016*. Berlin, 2016b, pp. 32–37.
Erman B. Cognitive processes as evidence of the idiom principle. *International Journal of Corpus Linguistics*. 2007. 12(1), pp. 25–53.
Farahmand M., Henderson J. Modeling the non-substitutability of multiword expressions with distributional semantics and a log-linear model. *Proceedings of the 12th Workshop on Multiword Expressions. ACL 2016*. Berlin, 2016, pp. 61–66.
Farahmand M., Smith A., Nivre J. A multiword expression data set: Annotating non-compositionality and conventionalization for English noun compounds. *Proceedings of NAACL-HLT. Association for Computational Linguistics*, 2015, pp. 29–33.
Farahmand M., Nivre J. Modeling the statictical idiosyncrasy of multiword expressions. *Proceedings of the 11th Workshop on Multiword Expressions*. 2015, pp. 34–38.
Goroshko E.I. *Integrativnaja model' svobodnogo associativnogo jeksperimenta* [Integrated model of free associative experiment]. Moscow, 2001.
Iordanskaya L.N., Mel'čhuk I.A. *Smysl i sochetaemost' v slovare* [Sense and compatibility in lexicon]. Moscow, 2007.
Karaulov Yu., Sorokin Yu., Tarasov E., Ufimtseva N., Cherkasova G. *Russian Association Dictionary*. 1994.
Kopotev M., Steksova T. *Isklyuchenie kak pravilo*: *perekhodnye edinitsy v grammatike i slovare* [Exception as a rule. Transitional units in grammar and lexicon]. Moscow, 2016.
Lukashevich N.V. *Tezaurusy v zadachakh informatsionnogo poiska* [Thesauri in Information Retrieval Tasks]. Moscow, 2011.

Loukachevitch N., Dobrov B. RuThes Linguistic Ontology vs.russian Wordnets. *Proceedings of Global WordNet Conference GWC-2014*. Tartu, 2014, pp. 154–162.
Loukachevitch N., Lashevich G., Gerasimova A., Ivanov V., Dobrov B. Creating Russian wordnet by conversion. *Komp'juternaja Lingvistika i Intellektual'nye Tehnologii*. 2016. Vol. 15, pp. 405–415.
Loukachevitch N., Lashevich G. Multiword expressions in Russian Thesauri RuThes and RuWordNet. *Proceedings of the AINL FRUCT 2016*. Saint Petersburgh, 2016, pp. 66–71.
Manning C.D., Schьtze H. *Foundations of statistical natural language processing*. MIT Press, 1999.
Maziarz M., Szpakowicz S., Piasecki M. A procedural definition of multiword lexical units. *Proceedings of Recent Advances in NLP Conference RANLP–2015*, pp. 427–435.
Mel'čhuk I.A. O terminakh "ustoichivost'" i "idiomatichnost" [About the terms "Fixedness" and "ldiornaticity"]. *Voprosy yazykoznaniya*. 1960. Vol. 4, pp. 73–80.
Mel'čuk I. Phraseology in the language, in the dictionary, and in the computer. Yearbook of Phraseology, 2012. Vol. 3(1). P. 31–56.
Miller G.A. Nouns in wordnet. *WordNet*: *An electronic lexical database*, pp. 24–45.
Pearce D. Synonymy in collocation extraction. *Proceedings of the workshop on WordNet and other lexical resources*, *Second Meeting of the North American Chapter of the Association for Computational Linguistics*, pp. 41–46.
Piasecki M., Wendelberger M., Maziarz M. Extraction of the multi-word lexical units in the perspective of the wordnet expansion. *RANLP-2015*, pp. 512–520.
Ramisch C., Cordeiro S., Zilio L., Idiart M., Villavicencio A., Wilkens R. How naked is the naked truth? A multilingual lexicon of nominal compound compositionality. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Vol. 2: Short Papers). *Association for Computational Linguistics*, 2016, pp. 114–133.
Reddy S., McCarthy D., Manandhar S. An empirical study on compositionality in compound nouns. *IJCNLP*, 2011, pp. 210–218.
Rohanian O., Taslimipoor S., Yaneva V., Le An Ha. Using Gaze Data to Predict Multiword Expressions. *Proceedings of the International Conference "Recent Advances In Natural Language 2017"*.
Sinclair J. *Corpus*, *concordance*, *collocation*. Oxford, 1991.
Sag I.A., Baldwin T., Bond F., Copestake A., Flickinger D. Multiword expressions: A Pain in the Neck for NLP. *Proceedings of International Conference on Intelligent Text Processing and Computational Linguistics*, *CICLING-2002*, Springer Berlin Heidelberg, 2002, pp. 1–15.
Senaldi M., Lebani G.E., Lenci A. Lexical Variability and Compositionality: Investigating Idiomaticity with Distributional Semantic Models. *Proceedings of the 12th Workshop on Multiword Expressions*. 2016, pp. 21–31.
Vincze V., Almasi A. Non-lexicalized concepts in wordnets: A case study of English and Hungarian. *Proceedings of Global WordNet*, *Conference GWC-2014*. Global WordNet Association, 2014, pp. 118–126.