НОВОСИБИРСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
СИБИРСКОЕ ОТДЕЛЕНИЕ РОССИЙСКОЙ АКАДЕМИИ НАУК

МАТЕРИАЛЫ
55-Й МЕЖДУНАРОДНОЙ
НАУЧНОЙ СТУДЕНЧЕСКОЙ КОНФЕРЕНЦИИ

МНСК-2017

17–20 апреля 2017 г.

ПРИКЛАДНАЯ ЛИНГВИСТИКА

Новосибирск
2017

УДК 81.33
ББК 81.1я431

Материалы 55-й Международной научной студенческой конференции МНСК-2017: Прикладная лингвистика / Новосиб. гос. ун-т. – Новосибирск : ИПЦ НГУ, 2017. – 46 с.

ISBN 978-5-4437-0643-6

ISBN 978-5-4437-0643-6

NOVOSIBIRSK STATE UNIVERSITY
SIBERIAN BRANCH OF THE RUSSIAN ACADEMY OF SCIENCES

PROCEEDINGS
OF THE 55th INTERNATIONAL STUDENTS
SCIENTIFIC CONFERENCE

ISSC-2017

April, 17–20, 2017

APPLIED LINGUISTICS

Novosibirsk, Russian Federation
2017

Proceedings of the 55th International Students Scientific Conference. Applied linguistics / Novosibirsk State University. – Novosibirsk, Russian Federation. 2017. – 46 pp.

ISBN 978-5-4437-0643-6

Section scientific supervisor – Dr. Philol., Prof. Timofeeva M. K.

Section head – Cand. Ped., Assoc. Prof. Snytnikova N. I.

Responsible secretary – Cand. Phys. Math, Assoc. Prof. Stukacheva M. V.

Section scientific committee

Cand. Psychol. Mozheykina L. B.
Dr. Phys. Math., Prof. Saveliev L. Ya.
Cand. Biol., Dr. Philos., Prof. Savostianov A. N.

ISBN 978-5-4437-0643-6

УДК 004.912

Применение методов машинного обучения в лингвистике на примере анализа темы высказывания

Бакаров А. А.
Новосибирский государственный университет

В наше время одним из наиболее популярных источников информации являются интернет-форумы, играющие роль доступных и обширных справочников по любым темам. Впрочем, они не лишены и одного существенного недостатка: большое число сообщений не относится к теме раздела, и пользователю приходится прилагать большие усилия, чтобы фильтровать демагогии, флейм, троллинг. Однако данная проблема могла бы быть устранена платформой, которая бы автоматически определяла тему высказывания и классифицировала сообщения в зависимости от их темы. Задача по созданию такой платформы относится как к сфере лингвистических наук (Natural Language Processing), так и к сфере машинного обучения (Machine Learning).

В простейшем случае классификации имеется два класса сообщений в зависимости от принадлежности сообщения к теме, и классификатором может выступать любой эффективный для данной задачи алгоритм машинного обучения с учителем. Но для того чтобы сообщения можно было подавать на вход классификатору, необходимо преобразовать текстовые данные в векторный вид. Метод преобразования основывается на технологии Word2Vec, которая, в отличие от классического статистического подхода к обработке текстов, учитывает семантические связи между словами. Такие семантические вектора описывают расстояние от лексемы до центра заранее сформированного и размеченного тематического кластера. Зная векторные репрезентации слов, можно получить векторное представление отдельно взятого сообщения и уже использовать эти данные в качестве параметров модели классификатора. Классовая принадлежность сообщения будет определяться как функция мягкого максимума косинусов углов между вектором сообщения и векторными вхождениям каждого класса. Остается разметить обучающую выборку и задать гиперпараметры выбранного классификатора.

Научный руководитель – Степанов П. А.

УДК 81’33

Комбинированный подход к автоматическому определению частеречной принадлежности слова

Бручес Е. П.
Новосибирский государственный университет

В данной работе описывается комбинированный подход для однозначного определения части речи у словоформы. Как следствие, подзадачей является снятие морфологической омонимии у словоформы, которой изначально может быть приписано более одной части речи.

Для решения поставленной задачи были изучены и реализованы несколько алгоритмов, а именно статистический метод и нейронные сети, которые различались входными признаки. Затем методы были объединены, тем самым была достигнута максимальная точность в решении указанных задач, по сравнению с результатами каждого отдельного алгоритма.

Для обучения и тестирования была использована часть Национального корпуса русского языка (НКРЯ) со снятой омонимией.

Одним из признаков, который мы использовали в морфологическом анализе, в качестве элементов контекста являются нормализующие подстановки, т. е. правила приведения словоформы к ее базовой форме. Это было обусловлено тем фактом, что русский язык является флективным, т. е. основным средством выражения грамматических значений служат окончания. Кроме того, предложенный метод позволяет существенно сократить размер словаря. Данный признак мы использовали в двух подходах.

1. Статистический метод. На корпусе текстов были подсчитаны вероятности того, что при наличии у слова подстановки *Subst_Word* и в контексте подстановки *Subst_Context* данное слово имеет часть речи *PartOfSpeech_Word*. Проверка описанного метода показала, что статистические данные имеются только для 75 826 слов из 100 000, выделенных для тестирования, и для этой части были получены следующие результаты: часть речи для каждого слова определялась с точностью 0,95; точность снятия омонимии – 0,93.

2. Нейронная сеть. В данной нейронной сети в качестве входного слоя использовался слой WordEmbedding, который подсчитывает векторы слов. Размер входного вектора был равен трем, в каждую позицию которого был записан индекс позиции подстановки слов в словаре (размер словаря – 20 810). Неизвестному слову присваивался индекс равный размеру словаря +1. Определение части речи – 0,95, разрешение омонимии – 0,92.

Также были реализованы дополнительно несколько нейронных сетей, которые различались входными признаками.

1. В качестве признаков были выбраны части речи слова. Входной вектор имеет размер 51 (17 частей речи * 3 слова). Данная нейронная сеть определяла часть речи для каждого слова с точностью 0,95 и разрешала морфологическую омонимию с точностью 0,91.

2. Признак: позиция словоформы в словаре. Эта нейронная сеть имеет ту же архитектуру, что и сеть, принимающая на вход нормализующие подстановки с тем отличием, что словарной единицей выступает теперь словоформа. Часть речи для каждого слова определялась с точностью 0,92; точность снятия омонимии – 0,91.

3. В качестве входных признаков использовались векторы слов, которые были заранее рассчитаны на НКРЯ. Размер каждого вектора для слова равен 300, соответственно размер входного вектора был равен 900 (300 * 3 слова). В случае омонимичной словоформы векторы усреднялись. Описанная нейронная сеть определяла часть речи для каждого слова с точностью 0,68 и разрешала морфологическую омонимию с точностью 0,57.

Для повышения точности решения обеих задач нами было принято решение объединить описанные выше методы. Окончательный результат принимался следующим образом. Для каждой части речи складывались вероятности результатов каждого из подхода, умноженные на некоторый коэффициент. Часть речи с максимальной суммой выбиралась как верная. После проведения нескольких экспериментов были выбраны следующие коэффициенты: для результатов нейронной сети по частям речи коэффициент был выбран равный 5,0, для нейронной сети по векторам слов – 4,0, для остальных алгоритмов – 1,0. Ансамбль такого вида решил задачу определения части речи с точностью 0,98, задачу снятия частеречной омонимии – 0,97.

Из полученных результатов видно, что комбинирование различных подходов и признаков помогает решить поставленную задачу с гораздо более высокой точностью, чем отдельно взятые методы. Также в работе было показано, насколько релевантен тот или иной признак в определении части речи словоформы.

К сожалению, все еще остается велика доля ошибок в определении служебных частей речи. Данная проблема является одним из приоритетных направлений для дальнейшей работы.

Научный руководитель – канд. физ.-мат. наук Свиридов К. С.

УДК 81.33

Речевое манипулирование в публикациях пользователей социальных сетей

Буглов Г. О.
Новосибирский государственный университет

Начиная с 90-х гг. XX в., наблюдается **стабильный рост** пользователей сети Интернет, причем растет не только количество уникальных пользователей, но и процент населения земли, использующий сеть Интернет. Вместе с этим расширяется и аудитория социальных сетей.

В России наблюдаются те же тенденции: в 2008 г. всего 25,4 % населения России являлись пользователями Интернета, а к 2016 г. этот же показатель вырос почти трижды – 70,4 %. Одной из причин такого бурного роста является распространение смартфонов – сегодня 56 млн россиян в возрасте от 16 лет пользуются Интернетом на мобильных устройствах – смартфонах и планшетах (46,6 % от всей аудитории).

Вместе с увеличением количества пользователей сети Интернет растет и количество интернет-рекламы. Особенно отчетливо это заметно на фоне динамики сегментов маркетинговых коммуникаций: в 2015 г. только интернет-реклама имела положительную динамику (15 %), в то время как объем долей телевидения, радио, прессы, наружной рекламы и прочего уверенно падал (от –14 % до –29 %).

Целью работы является определение термина «речевая манипуляция» в контексте рекламы в социальных сетях, определение места рекламы в социальных сетях в различных классификациях, построение классификации приемов речевого манипулирования в интернет-рекламе.

В ходе работы были выделены основные признаки термина «речевая манипуляция», проанализированы различные психологические и лингвистические подходы к его определению.

Существует множество классификаций приемов речевого манипулирования, однако большинство из них учитывает только приемы, использующиеся преимущественно в новостных заметках или текстах, каким-либо образом связанных с политической ситуацией. Их использование не даст каких-либо удовлетворительных результатов при анализе языка рекламы. В связи с этим некоторые классификации были модифицированы и дополнены для составления классификации непосредственно приемов, использующихся в языке рекламы.

В качестве материала для анализа были выбраны тексты с рекламным содержанием из социальной сети Instagram, содержащие в себе речевые манипуляции.

За основу взят способ деления методов речевого манипулирования в зависимости от их отнесенности к уровням языка. Таким образом, можно выделить методы, относящиеся к лексике, грамматике и синтаксису.

Научные руководители – канд. психол. наук, доц. Можейкина Л. Б.

УДК 81’33

Языковая личность автора в поэтическом дискурсе (на материале произведений Д. Томаса)

Исамбетова Л. В.
Кемеровский государственный университет

Понятие «языковая личность» означает совокупность способностей и характеристик человека, обусловливающих создание и восприятие им речевых произведений. Исследование данного феномена является одним из актуальных направлений в лингвистике; действительно, благодаря внимательному изучению речи отдельных носителей языка можно добиться успеха в описании национальных особенностей коммуникации в целом. Известно, что в дискурсе ярко проявляются образованность личности, ее отношение к выбору языковых средств и целый ряд других факторов, что неизбежно находит отражение в особенностях индивидуальной вербальной репрезентации. Наглядный пример тому – драматические произведения, где портретную характеристику персонажей можно составить преимущественно по их речи.

Поэтический дискурс представляет собой особое художественное поле для изучения проблемы языковой личности, поскольку лирика – наиболее субъективный литературный жанр, в котором динамика глубинных переживаний зачастую раскрывается на фоне почти полного отсутствия внешнего действия. В то же время автор с помощью фигуры лирического героя может создавать целые галереи субъектов переживания.

Вместе с тем, в художественных произведениях такого рода нередко присутствует имплицитная информация о происхождении автора, его творческом кредо, жизненной философии, психологии и социально-нравственных установках – все это может быть зашифровано в сложной системе образности, передаваться богатым репертуаром лингвостилистических средств, а также через звуковую аранжировку дискурсивных отрезков. В связи с этим, изучение языковой личности требует разностороннего подхода на стыке таких наук, как психология, литературоведение, социология, лингвокультурология и др.

В настоящем исследовании анализу подлежит поэзия Д. Томаса, которая интерпретируется как речевой акт, производимый под влиянием языковой личности автора и национального контекста. Рассмотрение поэтического дискурса осуществляется в соответствии с концепцией Ю. Н. Караулова по трем уровням: вербально-семантическому, логико-когнитивному и деятельностно-коммуникативному, что позволяет сделать наблюдение относительно влияния языковой личности автора на особенности структурно-содержательной организации лирических произведений.

Кроме того, предметом изучения являются экстралингвистические факторы: культурно-историческое пространство первой половины 19-го в.,

социальное положение поэта и особенности его психологического склада, сведения о которых почерпнуты из ряда биографических очерков. Особый интерес представляет тот факт, что Д. Томас активно выступал с чтением своих стихов, а также вел радиопередачу на BBC, благодаря чему известна авторская исполнительская манера. На основе изученных составляющих формируется образ языковой личности Д. Томаса в корреляции с английской поэзией в целом.

Так, на первом уровне анализа очерчивается лексикон поэта, который демонстрирует наличие основных семантических сетей: *Animals*, *Childhood*, *Death*, *Food*, *Human*, *Religion*, *Nature.* Ср. первую строфу стихотворения «*Fern Hill*»:

Now as I was young and easy under the apple boughs
About the lilting house and happy as the grass was green,
The night above the dingle starry,
Time let me hail and climb
Golden in the heydays of his eyes,
And honoured among wagons I was prince of the apple towns
And once below a time I lordly had the trees and leaves
Trail with daisies and barley
Down the rivers of the windfall light.

В данном дискурсивном фрагменте отмечены как единичные примеры конституентов вышеназванных полей (более ярко представленных в других стихотворениях) типа *Childhood – young*, так и многокомпонентная семантическая сеть *Nature – apple boughs*, *grass*, *night*, *trees*, *leaves*, *daisies*, *barley*, *rivers*, которая, в свою очередь, инкорпорирует значимую тематическую группу *The Farm*, так как стихотворение описывает ландшафты конкретной фермы в Кармартеншире, на которой писатель провел свое детство. Далее в стихотворении происходит аккумуляция единиц, связанных с описанием жизни на ферме: *stables*, *hey*, *ricks*, *fields*, *farm*. Примечательно, что ряд существительных имеют форму множественного числа, что порождает образную картину плодородного изобилия. Созданию благостного настроения способствует выбор автором положительно окрашенных слов: *happy*, *easy*, *heydays*. Наглядно представлена концепция-оппозиция «жизнь – смерть», эксплицируемая антонимами *light – night.*

В заключение отметим, что проблема статуса языковой личности как лингвокультурного феномена коррелирует практически со всеми дисциплинами, связанными с изучением антропосферы, что требует от исследователей комплексного, многоуровневого подхода к технологии ее реконструирования и более или менее достоверной интерпретации.

Научный руководитель – канд. филол. наук, доц. Омеличкина С. В.

УДК 81'33

Подход к разработке терминологических словарей по компьютерной лингвистике для поддержки решения задач автоматической обработки текста

Каршакевич А. О.
Новосибирский государственный университет

На сегодняшний день количество текстов, написанных на естественном языке, стремительно растет, вследствие чего существует необходимость внедрения автоматических средств для работы с ними. Решением такой задачи занимается компьютерная лингвистика (КЛ), научная дисциплина, моделирующая функционирование языка с помощью различных технических средств. КЛ является молодой перспективной наукой со сложным внутренним устройством. Необходимо исследовать и усовершенствовать ее понятийный аппарат. Определение терминологии КЛ, а также создание словаря ее терминов помогло бы решить эту проблему.

Разработка компьютерных словарей является одним из разделов КЛ. Сфера их применения включает в себя решение задач извлечения знаний, автоматического реферирования, машинного перевода, информационного поиска, например, для расширения поискового запроса. Также с их помощью реализуется автоматический анализ текста на различных уровнях языка. **Целью** данной работы является создание двух компьютерных словарей по КЛ для решения задач автоматической обработки текстов. Они могут использоваться различными программами, в частности при классификации текстовых документов с применением методов машинного обучения.

Источником для формирования словников послужили корпусы текстов, составленные по материалам научных конференций «Диалог», «RCDL», и научных электронных библиотек «eLIBRARY.RU» и «КиберЛенинка», поскольку эти ресурсы отражают реальное функционирование лексических единиц в российской КЛ. Так как данные словари предназначены для классификации текстов, полученный корпус был размечен, согласно классам, которыми являются разделы КЛ. Она делится на базовые теоретические и прикладные направления: Моделирование языка и языковой деятельности (с разделами Автоматическая обработка текста, Речевые технологии, Формализация описаний языковых средств и свойств речевых произведений) и Создание прикладных систем. Полученные коллекции документов не содержат нескольких похожих по тематике текстов в одном классе и включают в себя тексты сопоставимой длины.

В результате исследования разработаны два словаря по КЛ. Первый словарь состоит из ключевых слов, которые указывались авторами текстов обучающего корпуса или которые сопровождали тексты в электронных

библиотеках. Второй словарь был составлен экспертным методом с опорой на знание предмета, направлений КЛ и основ терминографии. При выборе лексических единиц, образующих словник, учитывались следующие критерии: частотность в обучающем корпусе, актуальность употребления термина в современных научных работах по КЛ, участие термина в формировании понятийного аппарата исследуемой предметной области. КЛ представляет собой сложный синтез других наук, таких как лингвистика, математика, искусственный интеллект, в ее семантическое поле входят термины, относящиеся к этим научным дисциплинам. Следовательно, для повышения качества классификации текстов необходимо включать в обучающий словарь не только лексемы, входящие в понятийное поле КЛ, но и общенаучные термины, термины смежных дисциплин и термины «метаязыка». К последним принадлежат термины уровней языковой системы и представлений этих уровней. Примерами терминов «метаязыка» являются слова *аббревиатура*, *парадигма*.

Работа была выполнена с использованием программного средства Клан (Соколова Е. Г., Загорулько Ю. А., Кононенко И. С. Опыт систематизации знаний и интернет-ресурсов для портала знаний по компьютерной лингвистике // Компьютерная лингвистика и интеллектуальные технологии: По мат. ежегод. Междунар. конф. «Диалог 2009». М. : РГГУ, 2009. Вып. 8(15). С. 465–470), созданного для разработки узкоспециальных словарей. В программе задействовано несколько модулей, которые автоматически извлекают из текста слова и словокомплексы и сопровождают их статистическими данными – частотой и весом. Лингвист вручную отбирает лексемы и словосочетания, отправляя нерелевантные для классификации слова в словарь стоп-терминов и словарь стоп-словокомплексов. Теоретической опорой для разработки словарей послужили «Русско-английский тезаурус по КЛ» (URL: http://uniserv.iis.nsk.su/thes/) и «Портал знаний по КЛ» (URL: http://uniserv.iis.nsk.su/cl/). Полученные словари планируется использовать для решения задач извлечения знаний.

Научный руководитель – Боровикова О. И.

УДК 81-33

Исследование восприятия рекламных роликов товаров для детей посредством семантического эксперимента

Козловская Е. А.
Омский государственный университет им. Ф. М. Достоевского

Работа выполнена при финансовой поддержке гранта РГНФ 15-04-00325а «Детство в дискурсивном пространстве региона: комплексный анализ институциональных и персональных коммуникаций с участием ребенка».

Семантическое исследование позволяет выяснить особенности восприятия гетерогенных составляющих полимодальных рекламных текстов. Эксперимент проходил в 3 этапа – сначала реципиенты воспринимали полный полимодальный текст рекламного ролика, на втором этапе – вербальную составляющую, на третьем – видеосоставляющую. Для проведения исследования были выбраны 2 рекламных ролика детских товаров – подгузники «Либеро» и йогурт «Агуша». Для получения данных эксперимента было задействовано 196 человек. Для проведения исследования была разработана анкета, которая содержала следующее задание: «Посмотрите предлагаемый рекламный ролик и укажите, какие характеристики товара Вам запомнились? (Не пересматривайте ролик снова, только один раз!)». После получения результатов ответы испытуемых были распределены по семантическим группам.

В ходе доэкспериментального комплексного психолингвистического исследования был сделан предварительный вывод о том, что вербальная и невербальная составляющая полимодального рекламного текста являются одинаково значимы в процессе воздействия на реципиента, так как рекламный ролик воспринимается в совокупности этих составляющих.

Семантический эксперимент позволил сделать выводы о способности реципиентов выделять из общего потока информации (причем и только вербальной, и только визуальной) актуальные смыслы.

При восприятии полного текста полимодального ролика рекламы детских товаров вербальная часть оказывала большее влияние на реципиентов. Соответственно, на данном этапе мы сделали промежуточный вывод о том, что вербальная составляющая полимодального текста рекламного ролика детских товаров является основной.

На втором этапе, при отсутствии видеоряда, основная часть информации достигала сознания реципиента. Точное цитирование частей звучащего текста встречалось чаще, но оно не скрадывало воспроизведение реципиентами прочей информации о товаре. Соответственно, мы по-прежнему можем думать, что вербальная часть в

тексте полимодального рекламного ролика обладает бо̀льшей значимостью, чем невербальная, и в случае потери первой, мы потеряем значительную часть информации о товаре.

Третей этап эксперимента представлял собой восприятие реципиентами только видео, без аудио составляющей. Большому количеству испытуемых удалось указать характеристики, присущие рекламируемым товарам, причем некоторые сделали это практически дословно, выделив такие же лексемы, как и те реципиенты, которые слушали только звук и смотрели полный видеоролик. Все семантические поля, присутствующие в тексте видеоролика, были восприняты реципиентами. А это позволяет утверждать, что для выявления характеристик товара, о которых идет речь в рекламном ролике, вербальная часть не является необходимым компонентом.

До получения результатов данного этапа логично было предполагать, что лидирующей семантической группой должна оказаться группа «Описание увиденного», так как при отсутствии информации, получаемой вербальным путем, испытуемым проще было бы ограничиться указанием на то, что они увидели в ролике. Но при обработке результатов эксперимента на данном этапе выяснилось, что указанная семантическая группа присутствует в ответах реципиентов всех групп, но не является основополагающей ни в одной из них.

Все же большинство реципиентов предпочли именно сложный путь декодирования информации и выделили характеристики, действительно присваиваемые продукту в вербальной части, им недоступной. Даже эмоциональные оттенки, заложенные в тексте, которые обычно выражаются в звучащей речи, были считаны реципиентами.

Соответственно, точно так же, как при потере визуальной части не произошла потеря смысла, так и при отсутствии вербальной части реципиенты по-прежнему правильно воспринимали рекламный ролик.

Это означает, что для правильного восприятия рекламного видеоролика детских товаров необходимым является только один канал восприятия. А значит, что невербальный и вербальный компонент рекламного ролика являются равновеликими, и невозможно выделить наиболее важную его составляющую. Логично, что максимальное количество информации реципиент может усвоить, задействуя одновременно 2 канала восприятия, а не один, а значит, инкорпорируя вербальную и визуальную часть, мы получим идеальный цельный полимодальный текст, обладающий наибольшей силой влияния – двигатель торговли.

Научный руководитель – д-р филол. наук, проф. Бутакова Л. О.

УДК 81'33

Возможности применения корпусных технологий в лингвоэкспертологии

Кочергина К. С.
Томский государственный университет

В лингвистической экспертной практике последних лет все более активно применяются экспериментальные методы, в том числе используется корпусный подход.

Возможности применения лингвистических корпусов в экспертной деятельности рассматриваются в работах А. Н. Баранова, А. А. Котова, З. И. Минеевой и др. Отдельные исследования (И. В. Тубалова, Ю. А. Эмер) посвящены изучению специфики конфликтного текста на материале Национального корпуса русского языка (далее: НКРЯ).

Вместе с тем, необходимо отметить, что имеющиеся корпуса современного русского языка не в полной мере отвечают требованиям анализа конфликтных текстов с позиции лингвистической экспертизы.

Самый известный проект – НКРЯ – представляет собой коллекцию текстов (в электронном формате), репрезентирующих русский язык в его современном состоянии. Масштабный характер этого корпуса достигается с помощью системы подкорпусов, ориентированных на определенные исследовательские цели и максимально полно представляющих все типы текстов. Несмотря на наличие в НКРЯ одиннадцати различных подкорпусов, ни в одном из них не представлены реальные конфликтные тексты. Возможность присутствия в НКРЯ потенциально конфликтных текстов – так называемых конфликтогенных – не отрицается, но это никак в нем не отражено.

Кроме НКРЯ существуют специализированные лингвистические корпуса, например, Томский диалектный корпус (Е. А. Юрина). Использование специализированных корпусов помогает решать узкоспециальные задачи и осуществлять результативный и достаточный по объему поиск лексических единиц, характерных для определенной области.

В данной работе предлагается создание коллекции конфликтных (и конфликтогенных) текстов как отдельного специализированного корпуса для реализации корпусного подхода в экспертной деятельности. В качестве перспективы возможен вариант разработки подкорпуса конфликтных текстов в составе НКРЯ.

Обозначим основные характеристики, которыми, на наш взгляд, должен обладать корпус конфликтных текстов.

1. Состав корпуса конфликтных текстов. В корпус войдут как устные, так и письменные тексты различного объема (от высказывания в одно слово до полных текстов), отражающие реальную коммуникацию

начала XXI в., регистрирующие коммуникативный конфликт, вовлеченные в правовую сферу.

2. Структура корпуса конфликтных текстов. В рамках корпуса будут выделены традиционные подкорпуса устной и письменной речи. Также возможно выделение подкорпуса так называемой «электронной» речи (Л. А. Капанадзе) – для текстов, размещенных в сети Интернет – и мультимедийного подкорпуса (Е. А. Гришина) – для текстов, сопряженных со зрительным и звуковым рядом (например, рекламные ролики).

3. Метаразметка корпуса конфликтных текстов. Конфликтные тексты разнородны по стилю, жанру, источнику, темам, поэтому наиболее актуальной представляется метаразметка текстов по группам в соответствии с категориями дел: защита чести и достоинства, клевета, оскорбление, экстремистская деятельность, нарушение авторских прав и др. Также возможна разметка корпуса в соответствии с классификацией конфликтных текстов, предложенной О. Н. Матвеевой: тексты-неудачи, тексты-злоупотребления, тексты-манипуляторы.

Дальнейшая работа требует детального изучения специфики конфликтных текстов для определения параметров их разметки, разработки структуры корпуса, сбора и обработки материала.

Использование подобного корпуса позволит:

1) вычленять значения слов, не зафиксированных в словарях, например, *гуки*;

2) определять реальные, актуальные, новые значения слов, их оттенки, употребляемые в «живой» речи и не зафиксированные в словарях;

3) проводить экспериментальные исследования в области семантики и подтверждать полученными данными экспертные выводы;

4) опираться при проведении экспертизы на прецеденты в аналогичных ситуациях.

Корпус конфликтных текстов будет представлять интерес для лингвистов-экспертов, лингвистов-исследователей (юрислингвистов), изучающих языковые явления на материале конфликтных текстов, и лексикографов.

Подобный корпус послужит инструментом для эффективного и быстрого поиска контекстов и словоупотреблений, даст материал для решения практических вопросов лингвоэкспертологии и юрислингвистики, а также позволит объективировать результаты исследования.

Научный руководитель – д-р филол. наук, проф. Демешкина Т. А.

УДК 81'33

Автоматическое обнаружение и исправление ошибок использования паронимов в текстах на русском языке с помощью алгоритмов, основанных на искусственных нейронных сетях

Лукаш А. В.
Новосибирский государственный университет

Выявление и исправление смысловых ошибок в текстах представляет собой актуальную и трудновыполнимую задачу для современных систем текстового редактирования и проверки правописания. Среди лексико-семантических ошибок, в частности, рассматриваются сложные опечатки (малапропизмы), при которых происходит замена знаменательного слова на сходное по звучанию, но не отвечающее контекстной семантике. К числу малапропизмов относятся паронимические ошибки: слово, значащееся в словарях, употребляется вместо формально похожего на него с идентичными грамматическими характеристиками (род, число, падеж, лицо) и синтаксической функцией в предложении. Под паронимами понимаются пары слов одного или омонимичного корня и одинаковой части речи, но с различиями в аффиксах (в префиксах и суффиксах), находящимися в фиксированных рамках. Так выделяют буквенные паронимы (однобуквенные: *адресант – адресат*) и морфемные (с малым числом различных служебных морфем: *этический – этичный*). Примерами ошибок в употреблении паронимов разной частеречной принадлежности могут стать следующие словосочетания: ***благодарственный*** *человек* вместо ***благодарный*** *человек*; ***тягостно*** *вздыхать* вместо ***тяжко*** *вздыхать*; ***одеть*** *пальто (на себя)* вместо ***надеть*** *пальто*.

Тема определения некорректного употребления паронимов на данный момент недостаточно исследована, основные экспериментальные методы связаны с применением бинарного классификатора или статистических показателей о встречаемости слов в сочетаниях. Однако признаки внешнего сходства малоинформативны для улучшения критерия паронимии, поскольку требуется привлечение семантики и одновременно соблюдение принципа контекстной инвариантности, при котором в случае подстановки одного паронима на место другого (в той же грамматической форме) правильность контекста не будет морфологически нарушена, измениться может только исходный смысл предложения. Вышеперечисленное и было учтено в данной работе, целью которой стала реализация программы с применением алгоритмов искусственных нейронных сетей для выявления ошибочных замен паронимов в русских текстах.

Решение поставленной задачи состояло из нескольких этапов.

1. Формирование по материалам основных словарей паронимов (Ю. А. Бельчикова, М. С. Панюшевой, О. В. Вишняковой,

В. П. Григорьева, Н. А. Кожевниковой, З. Ю. Петровой, В. И. Красных и др.) сводного списка буквенных и морфемных паронимов, включающего их исходные части речи, некоторые формы словообразовательных гнезд с зафиксированными в источниках контекстами употреблений пар (*поступок – проступок*), троек (*гуманный – гуманистический – гуманитарный*), четверок (*цветистый – цветной – цветовой – цветочный*) паронимов, приводимых к машиночитаемому формату.

2. Подготовка и конвертация объемного разножанрового корпуса текстов для наиболее эффективного обучения алгоритмов на данных с встречающимися паронимами, подсчет частоты вхождения каждого их которых проводился параллельно.

3. Разметка собранного корпуса: частичная лемматизация; удаление знаков препинания и стоп-слов (предлогов, частиц, союзов, междометий, местоимений и пр.) или тегирование отдельных групп (аббревиатур, имен собственных и др.); предобработка сокращений, вводных слов и сочетаний; специальный стемминг слов с опечатками.

4. Выбор гиперпараметров и реализация нейросетевой структуры, базирующейся на наборе алгоритмов word2vec для расчета векторных представлений слов (максимизации косинусной близости между векторами слов, встречающихся в похожих контекстах, и минимизация иначе) и сбора данных о совместном появлении слов в предложениях.

5. Тестирование нескольких программных моделей и определение итоговой с общим принципом работы, заключающемся в подборе слов-кандидатов на замену неверному пароними в полученном на вход тексте на базе существующих в словаре лемм и с учетом части речи и формы входного слова (например, ***исполнительского*** *органа* → ***исполнительного*** *органа*).

Приведенный алгоритм с модификациями в дальнейшем может быть основой и для обнаружения омофонов – слов, разных по значению и написанию, но одинаковых по звучанию (*притворить – претворить*, *компания – кампания*), а также совершенствования исправления обычных орфографических опечаток. В результате данной исследовательской работы программа станет одним из модулей функционирующей системы проверки орфографии, грамматики, пунктуации и стилистики текстов на русском языке.

Научный руководитель – канд. физ.-мат. наук Павловский Е. Н.

УДК 376.42

Особенности восприятия логопедических тестов у детей с ограниченными возможностями здоровья аутистического спектра с использованием методики айтрекинг

Макуха А. С.
Новосибирский государственный университет

Восприятие – психофизиологический процесс формирования перцептивного образа. Это сложный процесс, для изучения которого необходимо использовать комплекс методов. Исследование же восприятия у детей с расстройством аутистического спектра (РАС) затруднено в связи с психическими и поведенческими особенностями этой группы. Если ребенок без нарушений ментального развития может вербализовать результаты процесса восприятия, то у аутиста такое задание вызовет большие трудности. В связи с этим при исследовании восприятия у детей с РАС необходимо использовать такую методику, которая позволяла бы эффективно работать с данной выборкой и предоставляла бы достоверные и точные данные об особенностях описанного процесса. Такой методикой является неинвазивная технология айтрекинг.

Что касается логопедических тестов, то для разработки методик обучения детей с РАС важно понимать, каким образом они воспринимают предъявляемый материал для диагностики речевых способностей, поскольку знание подобных процессов позволяет разработать методики, наиболее эффективные для коррекции развития детей с РАС.

Мы исследовали особенности восприятия логопедических тестов детьми с РАС.

Для проведения данного исследования были выбраны следующие методики:

1. Материал для диагностики уровня речевого развития на основе речевых карт для детей 4–7 лет с общим недоразвитием речи, составленных Н. В. Нищевой. Он представляет собой систему тестов, оценивающих речевую способность ребенка с разных аспектов: фонематическое восприятие, лексико-грамматический строй речи, способность продуцировать связную речь и др.

2. Айтрекинг – технология, позволяющая отслеживать и записывать любые перемещения взгляда человека и на основании полученных данных анализировать особенности восприятия человеком предъявляемых стимулов. В данном исследовании использовался такой экспериментальный инструмент, как айтрекинговые очки. Нами была выбрана именно эта методика, поскольку она представляется наиболее удобной как для выбранных условий, так и для групп испытуемых, а именно: а) процесс сбора данных при работе с детьми легко можно превратить в интересную

игру с использованием необычного устройства; б) при работе с группой РАС необходимо обеспечить ребенку максимальный комфорт, при этом свести к минимуму физический контакт, болезненно переносимый детьми с данным расстройством; из большого количества экспериментальных средств методика айтрекинг лучше других отвечает предъявленным требованиям.

Участниками исследования являются 12 детей в возрасте 5–6 лет без отклонений в развитии и 12 детей в возрасте от 3 до 7 лет с диагнозом РАС.

Исследование проводится в наиболее привычных для детей условиях, т. е. в кабинете у логопеда / психолога в присутствии специалиста. Каждый ребенок по одному приглашается в кабинет, на него надевается устройство для сбора информации, после чего проводится обычное занятие с использованием выбранных методик.

В результате исследования мы получили данные, которые демонстрируют различия в процессах восприятия тестов у детей в выбранных группах. Например, у детей с РАС наблюдаются сложности с произвольной фокусировкой внимания на заданном предмете. Эти особенности находят отражение в уровне развития речи испытуемых. Так, у детей с РАС имеются затруднения в восприятии и выполнении заданий на фонематический анализ и синтез. В частности, затруднено или полностью отсутствует понимание таких понятий, как «первый звук в слове», «последний звук в слове». Контрольная группа детей выполняет данный блок заданий на 100 %. При исследовании лексики у детей с РАС наблюдаются затруднения при подборе обобщающих понятий для предъявленной группы предметов (пр. одежда, деревья). У контрольной группы детей такие проблемы возникают сравнительно реже. Кроме того, большинство испытуемых из группы РАС не смогли выполнить задания, связанные с продуцированием связной речи в связи с нарушениями восприятия предъявляемых материалов (пр. составить рассказ по серии сюжетных картинок). В контрольной группе такие проблемы выявлены только в 1 из 12 случаев. Что касается блока грамматических заданий, значительных различий в показателях представленных групп не выявлено.

Научный руководитель – канд. биол. наук Брак И. В.
Консультант – канд. психол. наук Можейкина Л. Б.

УДК 81'33+159.99

Психоэмоциональные реакции при восприятии речевых манипуляций в политических текстах

Ожерельева А. А.
Новосибирский государственный университет

Знания, получаемые из средств массовой информации (СМИ), играют значительную роль в жизни людей. Политика СМИ предполагает использование речевых средств, имплицитно воздействующих на адресата. В связи с этим возникает необходимость проверки достоверности предлагаемой информации. Актуальность работы заключается в выявлении речевых манипуляций в тексте и их отражения в сознании индивида. Мы исследовали показатели отражения ассоциативных и оценочных реакций человека, воспринимающего текст СМИ.

Объект исследования – языковые средства имплицитного воздействия на общественное сознание (на материалах новостного портала http://www.bbc.co.uk/russian). Предмет – ассоциативные и оценочные отклики респондентов при восприятии текстов СМИ.

Гипотеза исследования: мы предположили, что 1) ассоциативная реакция (результат ассоциативного теста, или АТ) на манипулятивные тексты будет эмоционально окрашенной, в отличие от реакции на нейтральные тексты; 2) оценочная реакция (результат семантического дифференциала, или СД) будет совпадать с ассоциативной по эмоциональной окраске вне зависимости от воспринятого текста.

Мы использовали метод интент-анализа для выбора текстов четырех статей новостного портала http://www.bbc.co.uk/russian, содержащих наиболее распространенные способы лингвистического манипулирования сознанием, опираясь на работу предыдущих лет [1]. Следующим шагом было составление презентации из 25 слайдов, тексты 15 из которых содержали приемы имплицитного воздействия на читателя, 10 – не содержали. Последние, так называемые нейтральные, тексты, были взяты с интернет-портала http://tvkultura.ru/news; критерием выбора служило отсутствие в тексте эмоциональной заряженности и манипулятивных приемов. В эксперименте принимали участие 40 испытуемых в возрасте от 18 до 23 лет. Для дальнейшего хода исследования мы применяли методы свободного письменного АТ и СД с использованием 10 стимулов, выбранных из ранее предложенных участникам текстов. По результатам АТ и СД мы называли стимул положительным (отрицательным), если количество положительных (отрицательных) оценок превосходит количество противоположных по значению минимум на 10 единиц; в ином случае – нейтральным.

Исходя из полученных по проведении эксперимента результатов, мы сделали следующие выводы:

1. Наблюдается тенденция к более нейтральным ассоциациям на стимулы из эмоционально неокрашенных текстов непосредственно после их прочтения респондентами при однозначно положительном оценочном отклике. Это объясняется тем, что эмоциональная окраска недавно воспринятого текста влияет на ассоциативное поле респондентов.

2. Оценочный отклик не совпадает с ассоциативным по эмоциональной окраске. Это можно объяснить тем, что на результаты АТ оказывают непосредственное влияние как эмоциональная окраска, так и приемы вербального манипулирования в только что воспринятых респондентами текстах; например, после прочтения отрывков, содержащих речевые манипуляции, испытуемые проявляют склонность к более негативному оцениванию стимулов, имеющих прямое отношение к России. Также нами было замечено, что разница между ассоциативными и оценочными реакциями на стимулы, использовавшиеся в отрывках, содержащих речевые манипуляции, выражена более явно, чем на стимулы, использовавшиеся в нейтральных текстах; различие между ассоциативным и оценочным откликами проявляется сильнее после восприятия манипулятивных текстов.

В будущем наши выводы и дальнейшие предположения будут проверены с помощью методики регистрации движения взгляда Eye-Tracking. Использование данного метода позволит найти связь психофизиологических, ассоциативных и оценочных реакций на стимулы, содержащиеся в предложенных участникам текстах.

---

1. Ожерельева А. А. Интент-анализ имплицитных способов воздействия на читателя новостного портала // Сборник тезисов МНСК-2016. Новосибирск, 2016.

Научный руководитель – канд. психол. наук, доц. Можейкина Л. Б.

УДК 81'322.2

Распознавание интенций покупателей в сообщениях социальных сетей (на примере сети «В контакте»)

Пименов И. С.
Новосибирский государственный университет

Значительное количество торговых сделок заключается посредством социальных сетей, которые одновременно характеризуются и высоким уровнем активности пользователей, и отсутствием единой системы организованной регуляции экономических отношений. Оптимизация инфраструктуры виртуального рынка возможна, например, за счет разработки программных средств, позволяющих связать продавца и покупателя, разделенных информационным шумом, заинтересованных в сделке по одному и тому же товару или услуге.

Задача автоматического обнаружения интенций (выраженных эксплицитно или имплицитно фактов желания совершить какое-либо действие) относится к проблемам поиска и извлечения информации из текстов и является актуальной в области компьютерной лингвистики.

Как правило, различают два основных подхода к решению данной задачи: 1) *экспертный*: систему правил (шаблонов) извлечения фактов строит эксперт, опираясь как на собственный опыт, так и на статистики, полученные путем автоматической обработки данных; 2) *машинное обучение*: правила строятся автоматически на базе обучающих коллекций, размеченных экспертом. В настоящей работе реализован первый подход.

Совокупность шаблонов интенций может быть представлена в виде ориентированного ациклического графа, позволяющего группировать схожие маркеры (слова и словосочетания, указывающие на присутствие интенции в тексте) на основании общности грамматических характеристик (и, как следствие, синтаксической функции), а также оптимизировать процедуру поиска, устраняя проблемы частичной заполненности факультативных позиций в шаблоне.

При построении графов были использованы данные N-граммного анализа коллекции текстов, содержащих интенции. Коллекция состояла из 1203 сообщений (текстов) социальной сети «В контакте», отобранных экспертом с помощью API LeadScanner. Под N-граммой понималась цепочка X из N подряд следующих элементов текста (лемм, граммем). В силу того, что сообщения представляли собой короткие тексты, рассматривались цепочки ограниченной длины ($N \leq 4$).

Процедура построения графов включала следующие действия.

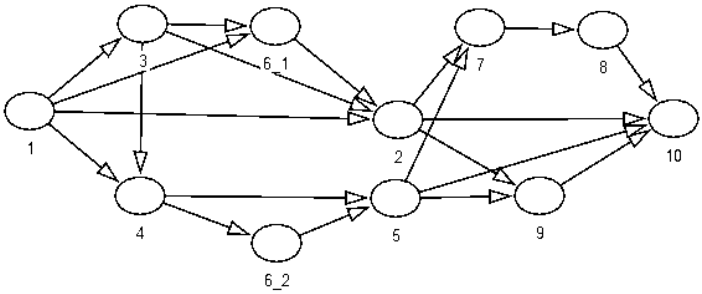
1. Нормализацию и морфологический анализ коллекции текстов (использованы модули PyMorphy2 версии 3.5.2.).

2. Формирование цепочек N-грамм с вычислением их абсолютной F_a и тестовой F_t частоты встречаемости в коллекции.

3. Фильтрацию N-грамм по критерию «маркерности». Цепочка X является маркером, если выполняется условие: $(C > P_1)$ & $(F_t > P_2)$, где $C = F_t / F_a$. Пороги P_1 и P_2 были установлены экспериментально.

4. Исследование цепочек на вариативность, а именно: возможность ограниченных по длине вставок в любую позицию цепочки. Это позволило устанавливать для каждой вершины графа возможные связи с последующим вершинами, которые должны быть включены в граф.

На рисунке приведен общий вид одного из построенных в ходе работы графов.



Пример графа интенции с входными маркерами «подсказать», «посоветовать», «порекомендовать»

1 – глагол, императив, множественное число («подсказать», «посоветовать», «порекомендовать», «предложить»);

2 – существительное, аккузатив (непосредственный объект интенции);

3 – формула вежливости («пожалуйста» и его слэнговые варианты);

4 – существительное, аккузатив («контакты», «номер», «координаты»);

5 – существительное, генетив (непосредственный объект интенции);

6_1, 6_2 – прилагательное, аккузатив («хороший», «толковый», «качественный»);

7 – предлог («в»);

8 – существительное, имя собственное, предложный падеж;

9 – существительное, имя собственное, именительный падеж;

10 – конец сообщения.

Считается, что интенция содержится в тесте, если в ходе его анализа из вершины 1 может быть достигнута вершина 10.

Программа поиска интенций реализована на языке Python. Предполагается ее апробация с получением оценок полноты и точности на тестовой коллекции текстов, включающей как тексты, содержащие интенции, так и не содержащие их.

Научный руководитель – канд. физ.-мат. наук Саломатина Н. В.

УДК 8. 81

Англоязычные песни как отражение грамматических, лексических и культурных особенностей изучаемого языка

Полекова Ю. А.
Красноярский государственный педагогический университет
им. В. П. Астафьева

В связи с растущей информатизацией в мире, развитием международных связей во всех сферах общества необходимо знание иностранного языка. Поэтому становится актуальным поиск новых эффективных способов и приемов изучения английского языка в «не англоязычной» среде. Прием использования англоязычной песни в изучении языка очень актуален, так как в ней находят отражение грамматические, лексические и культурные особенности изучаемого языка.

Целью данной статьи является анализ текстов англоязычных песен популярных исполнителей разных времен, начиная с 80-х гг. прошлого века и до наших дней. Тексты англоязычных песен, взятые из интернет-ресурсов – материал для исследования, в качестве изучения предмета рассматриваются лингвистические, грамматические и культурные особенности англоязычного песенного текста.

Безусловно, с точки зрения лингвистического уровня, музыкальная песня – это вид аутентичного текста, написанный носителями языка для носителей этого языка. Такие текстовые материалы специально не обработаны и отражают естественное языковое употребление. Из этого следует, что аутентичный текст напрямую отражает специфику и культуру изучаемого языка.

В рассмотренных песенных текстах можно наблюдать следующие лексические особенности: контракцию слов и просторечные формы, характерные для разговорной формы, которые авторы используют для сохранения ритма песен. Например: *wanna* (want to), *gimme* (give me), *dontcha* (don't you), *ain't* (am not/are, not/is), *kinda* (kind of), *ya* (you) т. д.

Кроме того, преимущественно для сохранения ритма песен, используется пропуск окончаний в словах, отражающийся в орфографии. Таким образом, при исполнении песни наблюдается элизия – выпадение одного или нескольких звуков: *talkin' – talking*, *wit' – with*, *'cause – because*.

В песенные тексты очень часто вводятся элементы разговорной и сленговой лексики в новых значениях, типичные для разговорного стиля: *pretty* (quite), *blue* (sad, depressed), *mad* (angry), *to dig* (to understand), *babe* (darling), *a drag* (something), *sweetheart* (beloved).

Грамматические правила, в большинстве своем, при написании песен выполняются, но можно заметить некоторые явления, которые все-таки не вписываются в правила классической грамматики. Например:

дублирование подлежащего: *my father he's a tailor*; использование двойного отрицания: *girl ain't no way*; использование элементов эллиптических форм: а) пропуск подлежащего: *will soon be gone*; б) пропуск вспомогательного глагола: *single mama you doing out there*.

Непосредственное отражение культуры в англоязычной песне можно увидеть в пословицах: *no pain no gain* – эквивалент русской пословицы (без труда не вытащишь рыбку из пруда), *a bad beginning makes a bad ending* – плохое начало ведет к плохому концу; а также устойчивые выражения и фразовые глаголы: *the valley of the fools* – долина дураков, *don`t let me down* – не разочаровывай меня, *wanna be with you* – хотеть быть с тобой, *to count out* – исключить из, *white lies* – невинная ложь, *hold on to smb* – держаться за кого-либо и др.

Также в текстах песен продуктивное употребление получает использование слов интернациональной лексики: *stratosphere, boulevard, alcohol* и т. д., понятной и доступной большинству слушателей на разных языках.

Таким образом можно сделать вывод, что для сохранения ритма песни авторы текстов используют самые разнообразные приемы: преимущественное употребление получают контракция слов, просторечные формы и элизия, а также очень часто вводят элементы разговорной и сленговой лексики в новых значениях. Нарушения грамматических правил могут характеризоваться дублированием подлежащего, использованием двойного отрицания и элементов эллиптических форм (пропуск подлежащего и вспомогательного глагола). Все эти явления типичны для разговорного стиля, который напрямую относится не только к лексическим и грамматическим особенностям, но и к культурной составляющей изучаемого английского языка. Культурную картину языка в песнях также отражают пословицы, интернациональная лексика, устойчивые выражения и фразовые глаголы.

---

1. Лингво-лаборатория «Амальгама» [сайт]. URL: http://www.amalgama-lab.com.
2. AZLyrics [сайт]. URL: http://www.azlyrics.com/

Научный руководитель – канд. филол. наук, доц. Колесова Н. В.

УДК 81’32

Особенности глагольной категории вида в русском языке в свете семантики Монтегю

Рыжков А. М.
Новосибирский государственный университет

Семантика Монтегю – это теория, в рамках которой американский математик Ричард Монтегю попытался применить методы, разработанные в различных областях математики (такие как модальная логика, интенсиональная логика, теория типов, лямбда-исчисление), для анализа семантики естественного языка. Кульминацией его исследований в этой области стал труд «The Proper Treatment of Quantification in Ordinary English», изданный в 1973 г. [5]. Ричард Монтегю рассмотрел фрагмент английского языка и построил правила перевода предложений из данного фрагмента на созданный им формальный язык.

Хотя теория Монтегю создавалась на материале английского языка, исследователь предполагал возможность ее применения к другим естественным языкам. Однако при использовании семантики Монтегю для анализа конкретного языка требуется учитывать некоторые его особенности. В качестве примера такой работы может быть упомянуто исследование немецкого именного словообразования в рамках теории Монтегю [4].

Единственным найденным авторами трудом по применению данной теории к русскому языку является монография И. А. Герасимовой «Формальная грамматика и интенсиональная логика» [1], которая представляет собой введение в теорию Монтегю на примере небольшого фрагмента русского языка.

Основная цель данной работы – анализ особенностей русского языка, играющих важную роль при применении к нему семантики Монтегю. В роли объекта исследования выступает глагольная категория вида в русском языке, которая указывает на характер протекания или распределения во времени действия, обозначаемого соответствующим глаголом. При определении смысла предложения русского языка крайне важно учитывать значение указанной категории. Основной инструмент исследования – интервальная семантика, использованная Д. Даути при анализе продолжительных времен глагола в английском языке [2]. Следует заметить, что английская глагольная система не имеет категории вида: характер протекания или распределения действия во времени выражается средствами категории с соответствующим названием. Такое пересечение во множестве выражаемых значений для глагольной категории времени в английском языке и глагольной категории вида в русском дает возможность предположить, что интервальная семантика может быть

использована при рассмотрении последней. Следовательно, основная цель работы – выяснить, насколько интервальная семантика применима для анализа глагольной категории вида в русском языке в рамках теории Монтегю.

Сведения о языковых особенностях, играющих важную роль при рассмотрении фрагментов естественного языка через призму семантики Монтегю, необходимы для возможного применения указанной теории в таких областях, как машинный перевод, извлечение информации из текста и др. Примером попытки использования теории Монтегю на практике являются исследования по созданию системы машинного перевода «Rosetta», проведенные в Голландии в 80-е гг. XX в. На сегодняшний день аппарат, созданный Монтегю, в прикладных сферах используется не слишком активно, однако представляется, что он имеет определенный потенциал и может быть успешно применен в указанных выше областях.

1. Герасимова И. А., Формальная грамматика и интенсиональная логика. М. : ИФ РАН, 2000. 156 с.

2. Dowty D. R. Word Meaning and Montague Grammar. Dodrecht : D. Reidel Publishing Company, 1979. 415 p.

3. Dowty D. R. et al. Introduction to Montague Semantics. Dodrecht : D. Reidel Publishing Company, 1989. 315 p.

4. Fanselow G. Zur Syntax und Semantik der Nominalkomposition. Tübingen : Max Niemeyer Verlag, 1981. 244 p.

5. Montague R. The Proper Treatment of Quantification in Ordinary English // J. Hintikka, J. Moravcsik, P. Suppes (eds.): Approaches to Natural Language. Dordrecht, 1973. P. 221–242.

Научный руководитель – канд. физ.-мат. наук, доц. Стукачев А. И.

УДК 004.912

Сравнительный анализ моделей автоматизированного определения метроритмических характеристик русских поэтических текстов

Савватеева Т. А.
Новосибирский государственный университет

При разработке алгоритмов автоматизации анализа поэтических текстов одной из основных является задача определения поэтического размера стихотворения. С этой целью, прежде всего, необходимо провести расстановку ударений во всех словах. Акцентуированные словоформы имеются в Словаре А. А. Зализняка, но требуется алгоритмическое решение проблем выявления нужной формы омографов («*зАмок*» или «*замОк*»?) и нарушения стандартной безударности служебных слов («*уронили мишку нА пол*»). Одним из способов решения этих проблем является метод «по аналогии». Этот метод заключается в сравнении строк и строф с неоднозначным ударением анализируемого стиха со строками и строфами, в которых ударение в словах расставляется однозначно и последующем выборе ударения, обеспечивающего единство метрической характеристики для всего стихотворения.

Но даже когда ударения правильно расставлены, автоматическое определение размера стиха остается нетривиальной задачей из-за возможного наличия неполных стоп, цезур и т. п.

Также важной характеристикой поэтического текста является рифма. Для автоматического определения типа рифмовки возможны два метода: использование уже существующих генераторов рифм, либо грамматический разбор строк или окончаний строк.

Простейший алгоритм без учета перечисленных выше проблем определения размера был предложен в работе [1]. Именно этот алгоритм и используется в разрабатываемой в ИВТ СО РАН системе автоматизированного анализа поэтических текстов http://poem.ict.nsc.ru. Однако в работе [2] был предложен намного более подробный алгоритм, учитывающий некоторые возможные нарушения стандартного размера. Реализация и сравнение эффективности этих двух алгоритмов и является основной задачей данной работы.

В рамках работы была написана программа на языке программирования Python 3.5, основанная на алгоритме, описанном в статье [2]. Основными пунктами работы программы являются:

1) акцентуация слов;
2) выделение рифмованных строк;
3) разбиение слов на слоги;
4) представление строк в виде ритмических схем и проверка схем на соответствие условиям, заданным в [2].

Для расстановки ударений используется база данных MySQL, составленная на основе словаря А. А. Зализняка, устранение неоднозначности ударения происходит методом «по аналогии», описанном выше.

Для выделения рифмованных строк в статье [2] предлагается использовать «Большой словарь рифм» [3]. Однако использование сторонних ресурсов при работе программы, во-первых, будет требовать установки дополнительного программного обеспечения, во-вторых, значительно замедлит работу программы. Поэтому при реализации предложенного алгоритма была создана дополнительная база данных MySQL, созданная на основе «Большого словаря рифм».

Ритмические схемы стиха (пункт 4) составляются на основе пунктов 1–3. По ритмическим схемам подсчитывается общее число ударных слогов и общее число слогов без учета клаузулы. На основе перечисленных выше параметров и вида рифмовки и определяется тип стихотворения.

Главными отличиями алгоритмов, описанных в [1] и [2] являются:

1) подход к определению метрической характеристики: в [1] используются числовые векторы, в [2] – соотношения слогов в ритмической схеме стиха;

2) подход к определению типа рифмовки: в [1] используется модуль фонетического разбора слов, в [2] – база данных.

При этом в статье [2] рассматривается большее разнообразие возможных типов поэтических текстов, чем в статье [1].

---

1. Козьмин А. В. Автоматический анализ стиха в системе Starling / Компьютерная лингвистика и интеллектуальные технологии // Труды международной конференции «Диалог 2006» (Бекасово, 31 мая – 4 июня 2006 г.). М. : Издательский центр РГГУ, 2006. С. 265–268.

2. Бойков В. Н., Каряева М. С., Соколов В. А., Пильщиков А. И. Об автоматической спецификации стиха в информационно-аналитической системе // CEUR Workshop Proceedings. 2015. V. 1536. P. 144–151.

3. Большой словарь рифм. URL: http://rifmovnik.ru/docs.htm.

Научный руководитель – д-р техн. наук, доц. Барахнин В. Б.

УДК 81’33

Анализ технологии создания системы извлечения именованных сущностей из текстов с использованием Томита-парсера

Судоплатова С. Н.
Новосибирский государственный университет

В настоящее время большое количество данных представлено в виде текстов на естественном языке, так называемый *plain text*. Эти массивы информации могут быть использованы для получения семантических структур различного типа: субъект-объектные отношения, действия, признаки и др. Решение таких задач, безусловно, требует большого количества ресурсов, поэтому возникла задача автоматического извлечения информации из текстов.

Одной из наиболее актуальных задач является **извлечение именованных сущностей**. Именованной сущностью считается слово или словосочетание, предназначенное для обозначения конкретного, вполне определенного предмета или явления, выделяющее этот предмет или явление из ряда однотипных. К именованным сущностям относятся, например, названия организаций, имена собственные и т. п. Извлечение именованных сущностей применяется в таких областях, как информационный поиск, автоматизированный сбор новостей, сбор информации об изменениях данных, касающихся каких-либо сущностей (например, об изменении адресов организаций) и т. д.

Томита-парсер – это созданное компанией Яндекс программное обеспечение для извлечения фактов из текстов. Оно использует подход, основанный на правилах, применяя для извлечения фактов написанные пользователем контекстно-свободные грамматики. **Целью** моего исследования был анализ технологии создания системы извлечения именованных сущностей из текстов с использованием данного ПО. В частности, необходимо было разработать контекстно-свободную грамматику для извлечения названий организаций с помощью Томита-парсера и оценить качество ее работы.

Для работы над проектом были предоставлены текстовые данные, извлеченные с сайтов организаций. На этом же тексте предварительно была проведена ручная разметка именованных сущностей. Тем самым, имелась возможность сравнить результаты работы грамматики с результатами ручной разметки и оценить полноту и точность извлечения фактов.

Для поиска названий организаций в правилах созданной грамматики используется контекст, а также графематические и морфологические особенности наименований учреждений, в частности:

- дескрипторы – слова, называющие организации (например, *магазин*, *аптека*, *школа*, формы собственности – *ООО, ЗАО, ИП* и др.);

- контексты – клише или выражения, которые часто употребляются с названиями организаций (*объявлять распродажу, заниматься разработкой*);
- особенности написания названий организаций: написание с заглавной буквы, CamelCase (*КаратСервис*), особые символы (*Gifts & Goods*) и др.;
- словообразовательные особенности названий, например, наличие специфических аффиксоидов (*-сервис-, -спец-, -строй, -трейд* и др.).

Оценки качества получились следующими:

Полнота = 0,4;

Точность = 0,736;

F-мера = 0,519.

На качество извлечения сущностей повлияли многие факторы, в том числе особенности текстовых данных (наличие в выборке большого количества названий, стоящих вне контекста и не имеющих при себе дескриптора, грамматически неправильные части текста), особенности работы Томита-парсера (токенизация, качество автоматического морфологического анализа), ошибки во вручную размеченной выборке, на которой подсчитывались ошибки. Безусловно, и сами правила грамматики не совершенны и могут дорабатываться и перерабатываться.

Таким образом, была создана грамматика для извлечения названий организаций с помощью Томита-парсера, оценено качество ее работы и намечены возможные действия по ее улучшению. Извлечение названий организаций – довольно сложная задача, ведь описать правила для всех видов названий в любых контекстах или без них невозможно, а также не всегда можно успешно отличать в тексте названия организаций от названий, например, товаров или от других имен собственных. Но в целом детерминированный подход к извлечению именованных сущностей, который был реализован с помощью Томита-парсера, может использоваться и давать неплохие результаты.

Научный руководитель – канд. филол. наук Домрачев М. А.

УДК 81´33

Лингвистическое сопровождение компьютерных игр на примере игры «Assassin`s Creed 3»

Суздальницкий Я. А.
Бурятский государственный университет, г. Улан-Удэ

Видеоигры представляют собой богатый источник материала для лингвистических исследований, при проведении которых необходимо учитывать их специфику как дискурсивных явлений, относящихся к сфере компьютерной коммуникации. Многие разработчики переводят свое программное обеспечение на другие языки для расширения аудитории, пользующейся их продукцией.

Актуальность данной работы определяется, во-первых, современной ролью компьютерных игр. Изучение контента компьютерной игры (причем ее разных жанров) представляется актуальным для целого ряда наук, в том числе и лингвистики (в ситуации локализации игры). Язык интерфеиса и сопроводительных документов, использующихся при создании игры, интересен с точки зрения прагмалингвистики и теории функциональных стилей. Текст компьютерной игры представляет собой совершенно новый материал, который необходимо ввести в научный оборот.

Говоря о переводе компьютерных игр, мы неизбежно используем термин «локализация». Что же такое локализация компьютерных игр? Локализация компьютерной игры – это перевод оригинальной версии игры на другой язык и адаптация ее к культуре другой страны. Процесс локализации состоит из некоторых этапов: анализ материалов, перевод текстов игры, перерисовка текстур и графики, дублирование и закадровый перевод видеороликов, тестирование локализованной версии игры. Кроме этого, необходимо наличие игрового опыта в играх схожего жанра, досконального знание сюжета локализуемой игры и взаимодействия персонажей. В процессе локализации компьютерных игр переводчики могут столкнуться с некоторыми сложностями. Причинами таких сложностей могут послужить различные нюансы, такие как короткие сроки, постоянные проверки. Вследствие подобных сложностей возникают ошибки в переводе, которые не позволяют получить полное эстетическое удовольствие от игры.

Для проведения анализа была использована игра «Assassin`s Creed 3». Игра представлена в жанре Action. Если учесть тот факт, что игра насыщена историческими фактами и событиями, то коммуникативная цель игры – обучающая. Сюжет игры берет свое начало в Северной Америке в 18 в. Игра повествует нам о многовековой войне между тамплиерами и ассасинами. Главный герой и он же автор – Дезмонд Майлс, потомок ассасинов. С помощью Анимуса – машины, которая считывает память

предков из ДНК – он переживает заново события далёкого прошлого. Его цель – узнать, где его предки спрятали «Частицы Эдема», и достать их раньше тамплиеров, а также предотвратить конец света.

Локализацией игры занималась компания Ubisoft. Стоит отметить, что перевод игры был сделан качественно и профессионально. Однако, переводчики, занимавшиеся локализацией игры, не смогли до конца адаптировать ее под русскую аудиторию, что привело к тому, что большая часть сюжета остается непонятной без использования внешних источников.

В первую очередь хочется отметить исторические события: Франко-Индейская война, Бостонское чаепитие, Стычка у Грейт-Медоуз, Война за независимость. В ходе работы не было найдено ошибок при переводе. Впрочем, это не имеет никакого смысла, если нет пояснений по поводу этих исторических событий. Желая понять сюжет сполна, игроку приходится дополнительно искать информацию. Учитывая тематику игры, легко объяснимо наличие в ней исторических имен: Бенджамин Франклин, Эдвард Брэддок, Чарльз Ли, Джордж Вашингтон, Самюэл Адамс. Перевод был выполнен методом транскрипции и транслитерации, и он был сделан корректно. Каждое имя несет в себе небольшое пояснение, однако, этого пояснения недостаточно, чтобы полностью окунуться в атмосферу Америки. Кроме этого, в лингвистическом сопровождении игры есть небольшие проблемы при озвучке. Дикторы говорят раньше, чем это начинают делать персонажи.

Анализ настоящей работы показал, что компьютерная игра действительно является богатым ресурсом для лингвистических исследований, в частности, локализации. Лингвистическое сопровождение является неотъемлемой частью компьютерных игр. Именно поэтому процесс локализации требует качественного перевода и культурной адаптации к особенностям определенной страны, региона или группы населения. Изучив перевод конкретной компьютерной игры, мы столкнулись с большим количеством ошибок и неточностей, которые влияют как на сам процесс игры, так и на эмоциональную составляющую игрока. В процессе изучения данной проблемы, нами были предложены некоторые решения. Во-первых, переводом лингвистического сопровождения игры должны заниматься профессиональные переводчики-локализаторы, имеющие представление об игре. Во-вторых, при переводе игры, имеющей исторический характер, стоит обращать внимание непосредственно на исторические события и имена. То есть локализаторы могут сделать сноски, которые будут содержать дополнительную необходимую информацию, или же в игре может быть представлен тематический глоссарий.

Научный руководитель – канд. филол. наук, доц. Самбуева В. Б.

УДК 81'33

Проверка орфографии методами машинного обучения

Фомин В. В.
Новосибирский государственный университет

Спелл-чекинг, т. е. автоматическая проверка орфографии, – распространенная частная проблема компьютерной лингвистики. Около 10–15 % запросов к поисковым системам содержат ошибки. Поскольку ошибки в запросе заметно ухудшают качество выдачи, их автоматическое исправление представляется важной задачей.

Как и многие задачи компьютерной лингвистики, автоматическая проверка орфографии может осуществляться с помощью методов машинного обучения – обширной междисциплинарной области, связанной с созданием обучаемых (а не эксплицитно программируемых) алгоритмов.

Потребность в спелл-чекере возникла в том числе при разработке мобильного приложения «2ГИС», содержащего электронные карты со справочниками. Запросы к приложению по состоянию на февраль 2017 г. подвергаются некоторой предварительной обработке, в том числе и простому спелл-чекингу. Тем не менее, работа такого спелл-чекинга не всегда удовлетворительна. Более точный алгоритм исправления ошибок заметно повысил бы удобство работы с приложением. При разработке такого спелл-чекера нужно учитывать специфику конкретной задачи: ввод с клавиатуры смартфона и картографическая тематика со специфической лексикой (напр. названия улиц), не всегда встречающейся в словарях.

Создание спелл-чекера для геоинформационного справочника можно разделить на следующие подзадачи: обнаружение опечатки, определение ее типа, ее исправление, составление и разметка корпуса, составление словаря и оценка качества алгоритма.

Обнаружение опечатки является примером задачи бинарной классификации; такие задачи хорошо решаются с помощью алгоритмов машинного обучения – логистической регрессии и нейронных сетей. При этом в качестве признаков, отличающих правильные строки от неправильных, можно использовать частоты n-графов (сочетаний букв длины n): например, слово «*вркзал*» можно опознать как содержащее опечатку, так как сочетание «*врк*» гораздо менее частотно, чем «*вок*». Кроме того, для этой подзадачи важно использовать словарь (хотя слово, содержащееся в словаре, может тем не менее в данном контексте содержать опечатку, и наоборот).

Определение типа опечатки относится к задачам многоклассовой классификации; эта проблема тесно связана со следующей, поскольку подход к исправлению опечатки зависит от ее типа.

К **исправлению опечатки** возможны следующие подходы:

– с помощью правил (rule-based);
– с помощью словаря;
– с помощью n-грамм (сочетаний из n слов);
– с помощью морфологического анализа.

Два главных типа опечаток – фонетические (типа *яблако*) и клавиатурные (типа *яблгко)* – решаются с помощью поиска по словарю, однако для каждого из этих двух типов поиск устроен несколько по-разному. При поиске в словаре слов, похожих на *яблако*, важно учитывать близость слов в произношении, тогда как для второго случая важно скорее расстояние между буквами на клавиатуре.

Данные о частотности n-грамм важны для исправления опечаток, при которых ошибочное написание слова выглядит как другое существующее слово, например, *еда для кита* вместо *еда для кота*. Поскольку поиск по словарю в таком случае неприменим, опечатку можно исправить благодаря тому, что сочетание *еда для кота* более частотно.

Морфологический анализ полезен для исправления опечаток, при которых слово находится в неправильной форме, например, *еда для коту*. Найти правильный вариант можно с учетом того факта, что предлог *для* требует родительного, а не дательного падежа.

Для **составления корпуса** нужна база запросов к приложению «2ГИС», прошедшая ручную разметку: для каждого запроса указывается, содержится ли в нем опечатка, и как этот запрос должен выглядеть с исправленной орфографией. Такой корпус нужен для обучения алгоритмов, а также для оценки качества уже обученных программ.

Для **оценки качества** применяются общепринятые показатели, самым простым из которых является процент правильных ответов (так называемый accuracy). Однако такой показатель обладает некоторыми недостатками. (Если ошибку содержит один запрос из пяти, то алгоритм может просто каждый раз утверждать «ошибки нет», и в 80 % случаев он будет прав). Поэтому используются также другие, более информативные показатели, такие как precision и recall – способность не считать целью то, что ею не является, и способность не пропускать цель, а также f1-мера, которая в равной степени учитывает оба этих показателя.

Научный руководитель – Бондаренко И. Ю.

УДК 81’33

Воспроизводимость моделей аргументации на примере текстов online-петиций

Фомичева А. В.
Новосибирский государственный университет

Автороведческая экспертиза предполагает установление автора текста главным образом на основе лексики и синтаксиса. Однако в некоторых случаях в рассматриваемом документе содержится весьма малое количество авторского текста, которого недостаточно для сбора статистики, на основе которой устанавливается авторство. В таком случае необходимы другие инструменты, одним из которых является теория аргументации. В русском языкознании работы, посвященные теории аргументации, рассматривают ее как аспект теоретический. **Новизна** данного исследования заключается именно в практическом применении теории аргументации. **Объект** исследования – теория аргументации. **Предмет** исследования особенности аргументирования своей позиции отдельным человеком как показатель его авторского стиля.

Чтобы использовать теорию аргументации для установления авторства, нужно доказать, что каждый человек использует определенные, характерные для него модели аргументации, и разные люди используют разные модели. Соответственно **цель** нашего исследования – найти в тексте модели аргументации и доказать их воспроизводимость у одного автора. Нами была проанализирована семантика аргументации (содержание аргументов и тезисов может сильно меняется в зависимости от темы диалога), синтаксика аргументации (последовательность тезисов, аргументов и контраргументов) и прагматика аргументации (принципы и правила, обеспечивающие уместность, действенность и успешность аргументации). Прагматика аргументации – наиболее индивидуальная часть, так как имеет непосредственную связь с действительностью и говорящим, а также она не ограничена определенным малым количеством конструкций, следовательно имеет больше вариантов выражения, и соответственно более индивидуальна [1].

В качестве материала исследования были выбраны авторские online-петиции (от лат. *petition* – «обращение, ходатайство, прошение») – разновидность электронного текста, получившая широкое распространение в сфере современной массмедийной коммуникации благодаря доступности создания и распространения в сети, оперативности и эффективности в достижении результата, экономии усилий, времени и материальных затрат [3].

Нами были проанализированы три петиции одного автора [4] (этот же автор недавно опубликовал новую петицию, однако она практически полностью повторяет две предыдущие, поэтому на ее основе анализ для

выявления моделей аргументации не проводился). В ходе исследования нами были найдены следующие модели аргументации:

– навязывание вывода через интерпретацию причинно-следственных связей;

– обращение к авторитету;

– перечисление аргументов с помощью вводных слов «*во-первых*», «во-вторых»;

– придание своим словам убедительности с помощью отсылок на федеральные законы;

– противопоставление («*надо – имеется*»);

– ссылка на необоснованное утверждение (автор называет это общеизвестным);

– ссылка на предыдущий опыт;

– «эмоциональная» аргументация (т. е. использование экспрессивно окрашенных предложений, вызывающих у читателей определенные эмоции относительно явлений, объектов, личностей, описываемых в предложениях).

При этом, в нескольких текстах повторяются модели 1, 3, 4, 7, 8, что может послужить доказательством того, что все эти тексты написал один и тот же человек.

Таким образом, на основе проведенной нами работы, мы можем сделать вывод, что человек действительно склонен повторять одни и те же модели аргументации в разных текстах.

---

1. Баранов А. Н. Лингвистическая теория аргументации (когнитивный подход): автореф. дис. ... д-ра филол. наук. М., 1990.

2. Василенко Л. Ю. Лингвокогнитивный анализ аргументации в тексте судебного решения: автореф. дис. … канд. филол. наук. М., 2011.

3. Кардович И. К., Коробова Е. В., Миронова Д. А. Аргументация как основополагающая категория дискурса // Современные исследования социальных проблем (электронный научный журнал). 2016. № 3(59). С. 224–237.

4. Online-сервис для составления и продвижения петиций. URL: www.change.org/ru (дата обращения: 20.07.2015).

5. Online-сервис для составления и продвижения петиций, автор Людмила Городскова. URL: www.change.org/u/39290607

Научный руководитель – Абрамкина А. Е.

УДК 88’33

Анализ технологии создания систем автоматического распознавания устной речи с большим словарем средствами CMU Sphinx

Яковенко О. С.
Новосибирский государственный университет

Современные поисковые системы используют такой инструментарий, как голосовой поиск – преобразователь поисковых запросов, произносимых пользователем, в текст исходного поискового запроса. Самый известный в настоящее время *голосовой поиск* – это известный нам «окей гугл». Его работа заключается в том, что пользователь произносит ключевую фразу «окей гугл» и далее с микрофона устройства производится распознавание речи.

На данный момент большинство подобных служб используют готовое программное обеспечение Google или Яндекса, и немногие имеют собственный голосовой поиск. Цель этой работы – разработать систему распознавания речи на основе инструментария CMU Sphinx. Это программное обеспечение в первую очередь будет удобно для водителей, пользующимся навигационными приложениями. А актуальность данной проблемы заключается в необходимости разработки системы распознавания речи с учетом специфики запросов (территориальные обозначения).

Таким образом, задачи данной работы – это ознакомиться с инструментарием CMU Sphinx, провести некоторые эксперименты по построению систем распознавания из уже имеющихся корпусов речи и построить свою систему распознавания на небольшом собственном корпусе речи.

Основным инструментом для построения систем распознавания послужила программа CMU Sphinx, созданный командой из университета Карнеги (Carnegie Mellon University). Было проведено 5 экспериментов для исследования работы инструментария и для пробного обучения и адаптации акустических моделей для поставленной задачи. Акустические модели строятся на основе корпуса следующей структуры:

Обязательные компоненты:

1) аудиофайлы;

2) файл транскрипции, где в каждой строке записана транскрипция аудиофайлов, а после транскрипции в скобках указана ссылка на файл.

Производные компоненты:

3) языковая модель – квантитативная модель языка, построенная на информации о сочетаемости слов в текстах (транскрипциях);

4) акустический словарь – текстовый файл, в котором хранится информация о фонетическом представлении каждого слова, которое встречалось в транскрипции;

5) словарь филлеров описывает фонемную природу слов-филлеров (отсутствие звука, мычание, кашель и прочие шумы) по аналогии с акустическим словарем;

6) список используемых фонем, что встречались в акустическом словаре и в списке филлеров.

Среди самых важных **операций** с вышеперечисленными данными можно выделить:

1) *распознавание* из файла или микрофона с помощью языковой модели, акустической модели и акустического словаря. Используется, как правило, языковая и акустическая модель по умолчанию, доступные для скачивания на сайте разработчиков CMU Sphinx;

2) о*бучение* акустической модели – создание акустической модели на основе имеющегося корпуса аудиозаписей и транскрипций к ним;

3) а*даптация* акустической модели – «надстройка» имеющейся акустической модели новым лексиконом, употреблением слов в конкретном контексте и синтаксических конструкциях;

4) д*екодирование* – проверка точности распознавания той или иной акустической модели на том или ином аудио корпусе.

Аудио корпус для последнего эксперимента строился на основе некоторого количества поисковых запросов справочника «2гис» для городов Москва, Новосибирск и Санкт-Петербург. Читались и записывались запросы молодыми людьми возрастом от 18-ти лет до 21-го года. Общий размер корпуса составил 1 час 42 минуты, из них 1 час 32 минуты – тренировочного и 10 минут тестового.

Результаты показали, что при таком размере словаря (697 единиц) недостаточно материала и имелось недостаточно повторений каждого из слов. Хорошо распознавались в основном такая общая лексика, как слова для обозначения географических объектов: «улица», «дом», «проспект», «район», а также некоторые простые числительные: «один», «пять». Следует также отметить, что не слишком хорошо показала себя адаптация имеющейся (по умолчанию) акустической модели, что может быть обусловлено большим размером словаря, который затрудняет выбор программы при соотнесении фонем и слов.

На основе вышеописанных результатов можно сделать следующие выводы:

1) для частных задач следует обучать с нуля акустическую модель, а не адаптировать имеющиеся модели по умолчанию;

2) для построения акустических моделей с большим словарем следует искать способы улучшения точности путем следующих методов: увеличения корпуса, построения грамматики или иных способов.

Научный руководитель – Бондаренко И. Ю.

УДК 81'33

Антитеза «женщина и война» в фильме «...А зори здесь тихие»: языковые средства выражения

Янь Цинвэнь
Новосибирский государственный университет

Предметом данного исследования служат вербальные актуализации концептов «женщина» и «война» в фильме «...А зори здесь тихие» режиссера Станислава Ростоцкого. Цель исследования – показать антитезу данных концептов, используя для этого их языковое выражение.

Основной конфликт в фильме «...А зори здесь тихие» базируется на противопоставлении концептов «женщина» и «война», именно через эту антитезу достигается трагизм произведения. Антитеза «женщины» и «войны» является неодномерной. В ней реализуются два противопоставления: 1) мужского и женского начал и 2) войны и мира.

Для удобства анализа антитезы концептов «женщина» и «война», будем рассматривать языковые средства ее выражения, разнесенные по двум группам: 1) лексические средства выражения; 2) морфологические (грамматические) средства выражения.

Лексические средства выражения

Лексические средства выражения антитезы чаще всего в кинотексте встречается в речи персонажей. Показателен эпизод, в котором Рита после ранения говорит Васкову: «*Ну зачем так? Все же понятно...* ***война*****.** Васков же возражает: *«Пока* ***война*** *– понятно. А потом, когда* ***мир*** *будет? Будет понятно? Что ответить, когда спросят: что же это вы, мужики,* ***мам*** *наших от пуль сберечь не могли?»*. В этой реплике мы можем увидеть все четыре члена противопоставления – с одной стороны, «война», «пули» и «мужики», которые должны иметь с этим дело; с другой – «мир» и «мамы», которых надо «сберечь». Здесь Васков указывает и на женское созидательное начало, и на ответственность мужчин за безопасность нынешних и следующих поколений, указывает на то, что ответственность мужчин за гибель женщин он не может списать даже на войну.

В произведении встречается большое количество подобного рода лексических противопоставлений рассматриваемых концептов. Также мы встречаем противопоставлением мужского и женского начал, в первую очередь в приложении к военной дисциплине:

Мария: «*Тихо-то как стало»* [после убытия зенитчиков].

Васков: «*А все из-за вас* (женщин)*! Выселить, выселить весь женский пол из прифронтовой полосы. В Соловки!»*

В данной фразе концепт «женщина» реализован непосредственно, с использованием словосочетания *женский пол*, а концепт «война» – опосредованно, через выражение *прифронтовая полоса*. Явный

антагонизм концептов выражается в необходимости пространственного отделения женщин от места, где идет война. О мужчинах здесь напрямую не говорится, но подразумевается, что женщины мешают им нести службу, разрушают дисциплину.

Морфологические (грамматические) средства выражения

Второй уровень, реализующий противопоставление – морфологический уровень. На этом уровне противопоставление войны и мира получает грамматическое выражение через категории рода.

Во всем произведении мы встречаем половую унификацию в обращениях персонажей друг к другу, например, «боец Комелькова», «младший сержант Осянина». В таких случаях антитеза усматривается в непривычном грамматическом использовании слов мужского рода для женщин. Интересна грамматическая интерпретация половой унификации военнослужащих в речи старшины Васкова:

Кирьянова: «*Знаете, товарищ старшина? Есть вопросы, на которые женщина отвечать не обязана.*»

Васков: «*Нету, нету здесь* ***женщин****. Есть* ***бойцы****, есть* ***командиры****.* ***Война*** *идет! И покуда она не кончится, все* ***в среднем роде*** *ходить будем.»*

Интересно, что в русском языке для солдат-женщин нет общеупотребительного существительного женского рода. Например, директор – директриса, учитель – учительница, студент – студентка. В то же время, слово «солдатка» обозначает, согласно словарям, не военнослужащую, а жену солдата.

Таким образом, в данном исследовании рассмотрены языковые средства выражения антитезы концептов «женщина» и «война», что позволяет глубже понять замысел авторов произведения – как повести, так и фильма. Тем не менее, следует отметить, что проведенный анализ языковых средств имеет некоторые недостатки.

В первую очередь, это трудоемкость анализа. Для проведения анализа концептов в нескольких произведениях, каждое произведение потребуется тщательно изучить, затратив на это длительное время.

Во-вторых, проведенный в нашем исследовании анализ не может претендовать на высокую надежность результата, поскольку его результаты могут по-разному интерпретироваться разными людьми.

На основании изложенного выше, представляется целесообразным использование компьютерно-корпусного подхода для анализа концептов. Это позволит получить однозначно трактуемые данные на больших выборках речевых употреблений.

Научный руководитель – д-р филол. наук, проф. Ким И. Е.

АВТОРСКИЙ УКАЗАТЕЛЬ

ОГЛАВЛЕНИЕ

Научное издание

МАТЕРИАЛЫ
55-Й МЕЖДУНАРОДНОЙ НАУЧНОЙ
СТУДЕНЧЕСКОЙ КОНФЕРЕНЦИИ

МНСК–2017

ПРИКЛАДНАЯ ЛИНГВИСТИКА

Материалы конференции публикуются в авторской редакции

Подписано в печать 31.03.2017 г. Формат 60x84/16
Уч.-изд. л. 2,9. Усл. печ. л. 2,7.
Тираж 100 экз. Заказ № 55.

Издательско-полиграфический центр НГУ
630090, г. Новосибирск, ул. Пирогова, 2