КРИТИКА И БИБЛИОГРАФИЯ / BIBLIOGRAPHY. REVIEWS

ОБЗОРЫ / OVERVIEWS

ПРЕДСКАЗАНИЯ, БОЛЬШИЕ ДАННЫЕ И НОВЫЕ ИЗМЕРИТЕЛИ: о возможностях технологий компьютерной лингвистики в теоретических лингвистических исследованиях

© 2016 г. Анастасия Александровна Бонч-Осмоловская

Национальный исследовательский университет «Высшая школа экономики»,
Москва, 101000, Россия
abonch@gmail.com

Статья посвящена обзору работ последних лет, в которых теоретическая исследовательская задача решается с помощью методов или инструментов, используемых в компьютерной лингвистике. В обзоре проводится подробный анализ того, как именно с помощью применения того или иного инструмента или метода можно получить новые знания о природе языка. В частности, выделяются два основных направления, развитие которых в рамках теоретических исследований представляется чрезвычайно перспективным. Это, с одной стороны, применение алгоритмов машинного обучения как предсказательной модели для описания многофакторных языковых явлений, с другой стороны, использование возможностей, открывающихся для типологических исследований и межъязыковых сравнений благодаря созданию множества «глубоко аннотированных» корпусов для разных языков, т. е. корпусов со сложной разметкой, например, синтаксической или референциальной. Уже сейчас объем имеющихся различных данных позволяет делать определенные выводы о свойствах тех или иных универсалий, которые были описаны раньше в теоретических типологических работах.

Ключевые слова: дативная альтернация, категория определенности, компьютерная лингвистика, машинное обучение, синтаксический корпус, теория языка, типология, референциальный выбор

PREDICTORS, BIG DATA AND NEW MEASURING: The impact of computational linguistics on linguistic theory

Anastasia A. Bonch-Osmolovskaya

National Research University «Higher School of Economics». Moscow, 101000, Russian Federation
abonch@gmail.com

The papers, observed in the overview, employ the methods of computational linguistics to enhance theoretical framework. The overview aims to demonstrate a detailed analysis of the benefits that a theoretical linguistic study could gain with the help of methods and instruments of computational approach. In particular, two domains seem to be very perspective. First of all, the use of machine learning technique as a prediction instrument for analysis of multifactorial linguistic phenomena. Secondly, there are completely new opportunities for typological studies due to big data of deeply annotated corpora, created for purposes of computational linguistics for different languages.

Keywords: computational linguistics, dative alternation, definiteness, language theory tree-bank, machine learning, natural language processing, referential choice, theoretical linguistics, typology

1. Введение

Настоящая статья представляет собой обзор нескольких работ последнего времени, выполненных в рамках задач теоретической лингвистики с использованием инструментов компьютерной лингвистики. Этот обзор продолжает основную идею работы [Толдова, Ляшевская 2014], посвященной наблюдаемым тенденциям к сближению между современной компьютерной лингвистикой и теорией языка. Однако, если в [Там же] рассматривалось в том числе и то, как достижения лингвистической науки могут быть использованы для получения более качественных результатов языкового моделирования, задачей этой работы будет показать пути сближения в обратной перспективе: как алгоритмы, методы и подходы, получившие широкое распространение в компьютерной лингвистике, могут быть применены для фундаментальных задач собственно науки о языке, для понимания того, каким образом устроен язык формально и содержательно.

Надо сказать, что за таким развитием лежит не только характерная в целом для современной науки тенденция к междисциплинарному сближению, но и отрефлексированное понимание проблем современной теоретической и компьютерной лингвистики, артикулированное «с обеих сторон» в целом ряде публикаций, появившихся в течение последних лет в центральных научных изданиях. Так, в журнале «Computational linguistics» в последнем номере 2009-го года была опубликована заметка в формате «last words» — заключительных слов — израильского компьютерного лингвиста Шули Уинтнера [Wintner 2009] о том, как связана автоматическая обработка естественного языка с лингвистикой. Уинтнер начинает заметку с обсуждения предложения открыть отдельную тематическую группу по лингвистике на конференции ACL — в настоящее время самой крупной конференции по компьютерной лингвистике. Уинтнер пишет, что несмотря на то, что сначала эта идея показалась ему удивительной (представим, например, что на конгрессе по педиатрии открывается тематическая группа, посвященная медицине в целом), позднее он признал правильность и своевременность этого предложения. Автор формулирует свою позицию как призыв к компьютерной лингвистике «возвратиться к истокам» и быть частью науки о языке. Основная проблема, по мнению автора, возникла как результат решительного поворота компьютерной лингвистики к статистическим методам, общим для искусственного интеллекта в целом, и она, по мнению автора, состоит в том, что современная компьютерная лингвистика по большей части занимается решением инженерных статистических задач, существует совершенно отдельно от лингвистической науки, не использует достижения лингвистической теории и ничего не добавляет к знанию о языке. Следует отметить, что часть вины в таком положении вещей Уинтнер возлагает на доминирующие лингвистические теории, в первую очередь генеративизм: «Теория стала столь непонятной, барочной “вещью в себе”, непроницаемой для тех, кто находится снаружи»[1] [Wintner 2009: 642]. В то же время автор считает совершенно неправильным такое положение вещей, когда компьютерная лингвистика превращается в раздел прикладной статистики, а естественные языки — в последовательность символов, ничем не отличающихся, например, от последовательностей ДНК. Основная идея Уинтнера состоит в том, чтобы существенно расширить базис приложений компьютерной лингвистики, используя ресурсы, методы и результаты как можно более разных лингвистических областей и теорий. Автор указывает и на то, что инженерные возможности компьютерной лингвистики должны быть использованы для лингвистических исследований. В качестве примера приводится работа [Daume III, Campbell 2007], в которой компьютерные методы применяются к данным Всемирного атласа лингвистических структур (WALS, см. подробней об атласе [Dryer, Haspelmath 2013]), а построенная в результате модель позволяет от лингвистических признаков языков перейти к лингвистическим универсалиям.

Безусловно, за пять лет, прошедших со времени публикации «памфлета» Уинтнера, диапазон прикладных задач, находящихся в спектре внимания компьютерной лингвистики,

[1] Перевод мой. — *А. Бонч-Осмоловская.*

расширился радикально, причем прежде всего в сторону все большего вовлечения в нее теоретической лингвистической базы. Разнообразие обсуждаемых вопросов и их решений в полной мере отражено в упомянутом выше обзоре [Толдова, Ляшевская 2014]. При этом нельзя не признать, что в основном это движение направлено на включение лингвистической понятийной базы в решение прикладных задач, например, на более широкое использование синтаксических и семантических характеристик для машинного обучения. Исследований, в которых компьютерные инструменты являются именно инструментом познания, а не самоцелью, по-прежнему немного. Однако, как будет показано ниже, интерес к ним постепенно растет, и в этом немаловажную роль сыграло встречное движение лингвистического сообщества, главным мотивом которого стал поиск объективных оценок надежности лингвистических данных, см., например, [Gibson et al. 2012; Gibson, Fedorenko 2013]. Именно этому типу исследований и посвящен настоящий обзор.

В обзоре обсуждается пять исследований. Они различаются по своим задачам. В работах [Bresnan, Ford 2010] и [Кибрик и др. 2010] компьютерные методы используются для проверки гипотез, которые до этого были сформулированы авторами на теоретическом уровне и за которыми стоит большая история лингвистических исследований и дискуссий. Другие три работы — [Bhatia et al. 2014; Futrell et al. 2014; Toldova et al. 2014] — напротив, представляют собой экспериментальные исследования, цель которых не подтвердить, а выдвинуть гипотезу, обнаружить ранее остававшуюся незамеченной взаимосвязь разных лингвистических факторов. Однако представляется, что сила работ обоих типов состоит в числе прочего в новой методологической базе. Именно поэтому в обзоре они сгруппированы по тому, как именно в них используются компьютерные лингвистические алгоритмы и ресурсы. В части 2 речь пойдет об использовании машинного обучения для оценки значимости разных лингвистических факторов, влияющих на выбор синтаксической структуры. В части 3 будут рассмотрены работы, в которых тестовые корпуса, используемые для проверки качества работы алгоритмов автоматической обработки текстов, применяются как массивы данных, оказавшихся доступными для лингвистического анализа.

2. Сила предсказания

В настоящем разделе речь пойдет о применении алгоритмов машинного обучения для моделирования многофакторных лингвистических явлений. Сама идея использовать машинное обучение связана с тем, что, с одной стороны, накоплено большое количество наблюдений, теоретических построений и даже статистических данных о факторах разной природы. Из их взаимодействия складывается выбор конкретной лингвистической формы: дативной конструкции с двойным объектом или же дативной предложной конструкции, определенной или неопределенной именной группы, некоторого референциального статуса имени. С другой стороны, так и не получилось придумать четкие формальные правила, которые бы учитывали все факторы, а любой конфликт значений факторов приводили к однозначному выбору лингвистической структуры. Метод машинного обучения дает возможность вывести такие правила из данных, максимизируя вероятностные показатели в зависимости от конкретного сочетания факторов. Обучение состоит в обработке большого количества наблюдений отдельных проявлений рассматриваемого феномена — в нашем случае это предложения в корпусе. Каждое такое наблюдение должно сопровождаться обучающей разметкой, отражающей поверхностное проявление тех самых факторов, которые мы считаем важными для анализа интересующего нас явления. Множество наблюдений дает распределение значений факторов, на основании которых строится математическая модель. Главная цель этой модели — подсчет вероятности того или иного лингвистического выбора (например, одного или другого типа дативной конструкции). Как будет показано ниже, сторонники использования обучающих моделей рассматривают их не просто как инструмент анализа, а как модель когнитивных функций человека. Иными словами, ключевая гипотеза состоит

в том, что лингвистический выбор, который делает человек, строится на тех же основаниях сопряжения разных факторов и оценки результата их взаимодействия.

Ниже будут рассмотрены три задачи, которые поставили перед собой три разных группы исследователей. Это задача предсказания выбора конструкции в дативной альтернации, в течение нескольких лет решаемая группой исследователей в Стэнфорде под руководством Джоан Бреснан, а также задача соотнесения семантических параметров определенности и неопределенности с морфосинтаксическими и лексическими характеристиками именной группы, выполненная исследователями из института Карнеги Мелон и университета Питтсбурга, и задача моделирования референциального выбора, выполняемая группой ученых из МГУ им. М. В. Ломоносова под руководством Андрея Кибрика.

Работы Дж. Бреснан и ее коллег, выполненные в последние несколько лет [Bresnan 2007; Bresnan et al. 2007; Bresnan, Nikitina 2009; Bresnan, Ford 2010], посвящены применению квантитативных методов для описания явления, известного как dative alternation, дативное варьирование. В настоящем обзоре я буду рассматривать результаты, представленные в работе [Bresnan, Ford 2010], хотя описываемый подход был предложен еще в работе 2007 г. и в дальнейшем модель уточнялась, но не менялась существенно.

Дативное варьирование является чрезвычайно популярной темой для синтаксического и семантического анализа, библиография исследований этого вопроса насчитывает множество работ начиная еще с 70-х годов прошлого века ([Green 1971; 1974] и далее [Pinker 1989], подробную библиографию современных исследований см. в [Bresnan, Ford 2010]). Собственно языковой проблемой является объяснение дистрибуции двух типов оформления аргументов у трехместных глаголов, первый из них предполагает подъем реципиентной именной группы в позицию прямого дополнения (1), а второй — синтаксическое выражение реципиента с помощью косвенного дополнения с предлогом *to* (2).

(1) *John gave Mary a book.*
(2) *John gave a book to Mary.*
‘Джон дал Мэри книгу’.

Подход Бреснан и ее коллег отличается от прочих квантитативных подходов тем, что авторы не просто используют статистические методы для определения частотности одной или другой конструкции в определенном контексте, но строят **предсказательную модель**. В этом смысле метод Бреснан оказывается близок к классификаторам, используемым в компьютерной лингвистике для машинного обучения. Сама идея лингвистической осмысленности понятия «предсказания» — **prediction** — требует обоснования: в [Bresnan, Ford 2010] этой задаче посвящен целый раздел. Авторы пишут о том, что в последние годы появляется все больше доказательств того, что предсказательность является важным параметром для восприятия и понимания языка. Это подтверждается нейролингвистическими исследованиями, в которых преактивация вызванных потенциалов мозга находится в корреляции с предполагаемым появлением исследуемых словоформ в определенном контексте. Другие исследования показывают, что слушатели используют механизмы порождения языка, которые помогают им строить предсказания, облегчающие понимание. Наконец, предсказательные модели объясняют многие эффекты частотности в овладении языком, его использовании и в исторических изменениях языка.

Бреснан и Форд утверждают, что и для синтаксических структур высокого уровня абстрактности можно построить работающую предсказательную модель. В целом довольно много известно про те параметры, которые так или иначе связаны с использованием каждой из дативных структур: доступность референтов, «сложность» (длина) лексического выражения, обозначающего референта, использование местоимений, одушевленность референта и так далее. Кроме того, оказалось, что большое значение имеет повтор той структуры, которая только что встретилась. Тем не менее, основная проблема, считают авторы, состоит в том, что большинство этих признаков оказываются взаимосвязанными или, другими словами, имеют взаимную корреляцию: личные местоимения короткие, определенные

и обычно одушевленные. Референты одушевленных имен, как правило, бывают определенными и часто являются «данным» в дискурсе, и они же часто могут обозначаться с помощью определений. Непонятно и то, каким образом, по каким правилам разрешаются конфликтные сочетания признаков. Каким образом можно понять, как именно характеристики влияют на выбор конструкции дативной альтернации? И далее, можно ли построить модель, которая, опираясь на известные лингвистические признаки, предсказывала бы выбор конструкции. Бреснан и коллеги [Bresnan 2007; Bresnan et al. 2007; Bresnan, Nikitina 2009; Bresnan, Ford 2010] предлагают в качестве такой модели **логистическую регрессию** — фактически это классификатор, который, получая на вход дистрибуции набора признаков, считает для каждого конкретного случая шансы получить одно из значений бинарного ответа: 1 или 0. В данном случае классификатор должен поделить все конструкции на два класса: объектно-дативный (VP NP NP) и предложно-дативный (VP NP PP). Хочется отдельно подчеркнуть, что задача модели в данном случае не инженерная, несмотря на то, что, как показано в работе, точность (в данном случае аккуратность — accuracy) предсказаний равняется 94 %. Этот классификатор дает возможность понять, каким образом устроено взаимодействие множества факторов, предсказательная сила классификатора в данном случае мыслится изоморфной человеческой способности угадывать выбор между двумя конструкциями. Кроме того, классификатор предоставляет возможности для дальнейших экспериментов: проверки значимости разных признаков, сравнения разных корпусов, экспериментального сравнения с результатами носителей разных разновидностей английского.

В чем суть математической модели, которую применяют к своим данным Бреснан и ее коллеги? Рассмотрим вначале то, как организованы данные, которые подаются на вход модели. Авторы используют базу данных, состоящую из 2 386 вхождений с дативными конструкциями, которые были получены из трехмиллионного англоязычного корпуса телефонных разговоров. Далее данные были размечены по множеству параметров («предикторов»), начиная от собственно глагольной леммы, наличия местоимения, числа и определенности темы и реципиента и кончая весьма нетривиальными предикторами типа параллелизма синтаксической структуры (дативная конструкция повторяет структуру, заданную в предшествующем предложении). Затем были получены числовые коэффициенты для всех предикторов. Числовой коэффициент в данном случае выводится из значения распределения этого параметра в выборке, например из значения распределения местоименного выражения реципиента, определенности или неопределенности темы и так далее. Фактически, каждый рассматриваемый случай употребления дативной конструкции представлен в виде **суммы числовых коэффициентов** тех предикторов, которые реализуются в данной конкретной конструкции, см. рисунок 1:

Model B: Response modeled as depending on

fixed effects: semantic class + accessibility of recipient + accessibility of theme + pronominality of recipient + pronominality of theme + definiteness of recipient + definiteness of theme + animacy of recipient + person of recipient + number of recipient + number of theme + concreteness of theme + structural parallelism in dialogue +length difference (log scale) −1

random effect: verb sense

Рис. 1. Модель предикторов дативной альтернации [Bresnan et al. 2007]

При этом коэффициенты предикторов различаются своим знаком. Знак задается таким образом, чтобы соотнести лингвистические основания предсказания признака с выбором соответствующей конструкции. Предиктор с положительным знаком предсказывает выбор 1 (VP NP PP), а с отрицательным знаком выбор 0 (VP NP NP). Так, как видно из рисунка 1, предиктор типа «pronominality of recipient» со значением местоимения имеет отрицательное значение, потому что он предсказывает конструкцию с двойным объектом, в то время как,

например, предиктор «number of theme» со значением множественного числа, напротив, предсказывает предложную дативную конструкцию и имеет в модели положительное значение. Интересным образом вычисляется предиктор длины именной группы. Из исследований дативной альтернации известно, что наиболее короткая именная группа становится ближе к глаголу — если короче реципиент, то к глаголу стремится реципиент, если короче тема, то тема. Для вычисления этого параметра авторы используют разность логарифма длин реципиента и темы. Таким образом, положительное значение предиктор принимает, если именная группа реципиента длиннее именной группы темы (выбор 1, предложная дативная конструкция), а отрицательный, если длиннее именная группа темы (выбор 0, конструкция с двойным объектом).

Еще один важный коэффициент модели, обозначаемый в формуле модели через переменную *û,* используется для того, чтобы отразить лексическое значение глагола — дело в том, что в зависимости от семантики глаголы в большей или меньшей степени склонны предпочитать одну или другую конструкцию. Числовым коэффициентом в данном случае является оценка, усредняющая влияние признаков относительного нормального распределения конструкций с тем или иным семантическим типом глагола. Пример работы модели можно увидеть на рисунке 2.

$$\text{Probability}\{\text{Response} = \text{V NP PP}\,|\mathbf{X}, \mathbf{u_i}\} = \frac{1}{1+e^{-(\mathbf{X}\boldsymbol{\beta}+\mathbf{u_i})}}, \text{ where}$$

$$\begin{aligned}\mathbf{X}\hat{\boldsymbol{\beta}} = {} & 1.1583 \\ & -3.3718\{\text{pronominality of recipient} = \text{pronoun}\} \\ & +4.2391\{\text{pronominality of theme} = \text{pronoun}\} \\ & +0.5412\{\text{definiteness of recipient} = \text{indefinite}\} \\ & -1.5075\{\text{definiteness of theme} = \text{indefinite}\} \\ & +1.7397\{\text{animacy of recipient} = \text{inanimate}\} \\ & +0.4592\{\text{number of theme} = \text{plural}\} \\ & +0.5516\{\text{previous} = \text{prepositional}\} \\ & -0.2237\{\text{previous} = \text{none}\} \\ & +1.1819 \cdot [\log(\text{length(recipient)}) - \log(\text{length(theme)})]\end{aligned}$$

$$\text{and } \hat{\mathbf{u}}_\mathbf{i} \sim N(0, 2.5246)$$

Рис. 2. Пример реализации предсказательной модели дативной альтернации [Bresnan, Ford 2010]

Построив модель, авторы тестируют ее работу методом кросс-валидации (cross-validation), также очень часто применяющимся в компьютерной лингвистике в случае обучения на ограниченных данных. Суть метода состоит в том, что база 100 раз делится случайным образом на две части — часть большего объема для обучения (обучающий корпус), из этой части базы будут выводиться все коэффициенты, используемые в модели, и часть меньшего объема, тестовый корпус. Роль последней части состоит в том, чтобы проверить, насколько точны предсказания модели. Предложения этой части базы не участвовали в подсчете параметров, модель будет выводить предсказания на основе коэффициентов обучающего корпуса, а далее результаты будут сопоставляться с реальными данными базы. Суть проверки состоит в том, чтобы обработать 100 тестовых выборок, а затем посчитать среднюю классификационную аккуратность (отношение правильных ответов к общему числу ответов). Результат предсказательной силы классификатора в работе [Bresnan, Ford 2010] равен 0,945, что, безусловно, является очень высоким показателем.

В чем состоит научный потенциал классификатора, описанного в работах Дж. Бреснан и ее соавторов? Во-первых, такая модель помогает оценить значимость вклада различных признаков, выделяемых ранее интуитивно: с помощью специального алгоритма разные предикторы последовательно изымаются из модели и оценивается то, насколько сильно меняется результат. Важным теоретическим следствием является опровержение так называемых

«редукционистских» моделей, в которых противопоставление дативных конструкций сводится к противопоставлению двух ключевых параметров. Во-вторых, модель может быть использована для дальнейших исследований. Так, в [Bresnan, Ford 2010] описывается серия экспериментов, связанных с пониманием и порождением дативных конструкций, которые проводились в двух группах носителей английского: в группе американцев и в группе австралийцев. В результате обучения модели на данных экспериментов удалось показать, что носители австралийского варианта английского обладают несколько иными представлениями о значимости предикторов, чем американцы. Так, параметр длинной именной группы реципиента оказывается более значимым для австралийцев при выборе предложной конструкции, чем для американцев. Например, в контексте (3) вероятность выбора предложной конструкции (ii) *to my kids* равна 0,3309, а вероятность выбора двойной объектной конструкции (i) — 0, 6691:

(3) *Speaker A:* *I wish they had just one central place, you know, where you can just dump all the recycling.*
Because really I am not really looking for the money portion of it, you know.
Speaker B: *Well I used to. It used to be a good days work.*
(i) Instead of giving my kids an allowance,
(ii) Instead of giving an allowance to my kids,
I just told them they could go around the neighborhood and collect things to be recycled and then I would drive them over and they would get some money.

‘Говорящий А: Я бы хотел, чтобы было одно основное место, куда можно сваливать все перерабатываемые отходы. Поскольку я, правда, тут думаю не о деньгах.
Говорящий Б: Я и о деньгах думал. Это всегда было хорошим заработком.
(i/ii) Вместо того, чтобы давать моим детям на карманные расходы, я просто говорил им, что они могут пройтись по округе и собрать вещи, подходящие для переработки, а потом я их отвезу и они получат немного денег’.

Если в примере (3) используется распространенная именная группа, например *my kids and their cousin who is staying with us*, то модель предсказывает бóльшую вероятность выбора предложной модели для австралийцев, чем для американцев.

Возвращаясь к использованию алгоритмов компьютерной лингвистики в теоретических исследованиях, отметим, в чем состоит роль классификатора в теоретических исследованиях. При построении инженерных моделей классификатор обучается на ряде параметров, чтобы достичь определенного качества (performance) в получении правильных ответов на моделируемый вопрос, в данном случае бинарный 1/0. Конкретные числовые параметры предикторов неважны, если система работает хорошо, сокращение предикторов интересно лишь для оптимизации инженерной задачи. При использовании классификатора в задачах теоретической лингвистики именно вклад (вес) предиктора оказывается наиболее ценным результатом. Если модель работает хорошо, то ее переобучение на других выборках и корпусах открывает возможности для измерения языковой вариативности в наиболее сложно уловимых областях, например, в области синтаксических предпочтений. Таким образом, ключевой потенциал классификатора при решении теоретических задач состоит, во-первых, в возможности **оценить вес** различных параметров при оценке сложного мультифакторного лингвистического явления, а во-вторых, в возможности использовать **числовые измерения** для нелексической языковой вариативности.

Еще одним примером использования машинного обучения для получения моделирования сложного языкового явления является работа [Bhatia et al. 2014], представленная на конференции Coling 2014 в Рейкьявике. В работе рассматривается проблема коммуникативной функции грамматической категории определенности. Сложность и многофакторность этого явления понятна каждому носителю русского языка — языка без артиклей, — который хоть раз пытался порождать тексты на английском — языке, в котором определенность и неопределенность маркируется грамматически. Надо сказать, что сама постановка задачи

связывается авторами с задачами машинного перевода с безартиклевых языков на языки с артиклями. Авторы приводят внушительную библиографию решения этой задачи для улучшения качества машинного перевода, однако в работе подчеркивается теоретическая ценность избранного авторами подхода. Так же, как и в работе [Bresnan, Ford 2010], авторы строят классификатор, основанный на множестве признаков, которыми может задаваться определенность или неопределенность именной группы. Однако в отличие от задачи объяснения дативной альтернации, в которой семантические параметры используются для предсказания синтаксической структуры, в данном случае, напротив, авторы используют синтаксические предикторы для предсказания семантических категорий. Сами семантические категории кодируются весьма сложным образом, в виде иерархической схемы, в которой собственно объектом категоризации являются термы иерархии (leaf nodes), а их «родители» представляют собой лишь атрибуты, которые они наследуют, см. рисунок 3.

- **NONANAPHORA** [$-A,-B$] **999**
  - **UNIQUE** [$+U$] **287**
    - ∗ **UNIQUE_HEARER_OLD** [$+F,-G,+S$] **251**
      - · UNIQUE_PHYSICAL_COPRESENCE [$+R$] 13
      - · UNIQUE_LARGER_SITUATION [$+R$] 237
      - · UNIQUE_PREDICATIVE_IDENTITY [$+P$] 1
    - ∗ UNIQUE_HEARER_NEW [$-F$] 36
  - **NONUNIQUE** [$-U$] **581**
    - ∗ **NONUNIQUE_HEARER_OLD** [$+F$] **169**
      - · NONUNIQUE_PHYSICAL_COPRESENCE [$-G,+R,+S$] 39
      - · NONUNIQUE_LARGER_SITUATION [$-G,+R,+S$] 117
      - · NONUNIQUE_PREDICATIVE_IDENTITY [$+P$] 13
    - ∗ NONUNIQUE_HEARER_NEW_SPEC [$-F,-G,+R,+S$] 231
    - ∗ NONUNIQUE_NONSPEC [$-G,-S$] 181
  - **GENERIC** [$+G,-R$] **131**
    - ∗ GENERIC_KIND_LEVEL 0
    - ∗ GENERIC_INDIVIDUAL_LEVEL 131
- **ANAPHORA** [$+A$] **1574**
  - **BASIC_ANAPHORA** [$-B,+F$] **795**
    - ∗ SAME_HEAD 556
    - ∗ DIFFERENT_HEAD 329
  - **EXTENDED_ANAPHORA** [$+B$] **779**
    - ∗ BRIDGING_NOMINAL [$-G,+R,+S$] 43
    - ∗ BRIDGING_EVENT [$+R,+S$] 10
    - ∗ BRIDGING_RESTRICTIVE_MODIFIER [$-G,+S$] 614
    - ∗ BRIDGING_SUBTYPE_INSTANCE [$-G$] 0
    - ∗ BRIDGING_OTHER_CONTEXT [$+F$] 112
- **MISCELLANEOUS** [$-R$] **732**
  - PLEONASTIC [$-B,-P$] 53
  - QUANTIFIED 248
  - PREDICATIVE_EQUATIVE_ROLE [$-B,+P$] 58
  - PART_OF_NONCOMPOSITIONAL_MWE 100
  - MEASURE_NONREFERENTIAL 125
  - OTHER_NONREFERENTIAL 148

	+	−	0		+	−	0		+	−	0		+	−	0
Anaphoric	1574	999	732	Generic	131	1476	1698	Predicative	72	53	3180	Specific	1305	181	1819
Bridging	779	1905	621	Familiar	1327	267	1711	Referential	690	863	1752	Unique	287	581	2437

: CFD (Communicative Functions of Definiteness) annotation scheme, with frequencies in the corpus. Internal (non-leaf) labels are in bold; these are not annotated or predicted. +/− values are shown for ternary attributes Anaphoric, Bridging, Familiar, Generic, Predicative, Referential, Specific, and Unique; these are inherited from supercategories, but otherwise default to 0. Thus, for example, the full attribute specification for UNIQUE_PHYSICAL_COPRESENCE is $[-A,-B,+F,-G,0P,+R,+S,+U]$. Counts for these attributes are shown in the table at bottom.

Рис. 3. Семантические признаки категории определенности, использованные при построении модели [Bhatia et al. 2014]

Такой подход был выбран как компромисс между множеством классификаций семантических признаков, иерархическая таблица используется как способ представления семантических классов, однако, как пишут авторы, для использования классов в рамках модели осуществлялась декомпозиция иерархии. Каждая категория, обозначенная жирным шрифтом, принимает одно из значений +, − или 0. Категории, обозначенные нормальным шрифтом, наследуют признаки более высоких категорий, добавляя к ним свои собственные. Таким образом, полная спецификация так называемого атрибута (категории, которую будет предсказывать модель) состоит из закодированной восьмибуквенной последовательности, в которой каждая буква имеет положительное, отрицательное или нулевое значение. Так, например, атрибут, называемый Unique_Physical_Copresence (уникальный референт, связанный с физическим присутствием) получает отрицательные характеристики по параметрам анафоричности (A), бриджинга[2] (B), генеричности (G), нулевую характеристику по параметру

[2] Под бриджингом в терминологии компьютерной лингвистики понимается неанафорическая референциальная связанность лексем, такая как, например, в предложении *Петя вошел в комнату*

предикативности (P) и положительные характеристики по параметру известности (F), референциальности (R), специфичности (S) и уникальности (U), см. легенду на рисунке 3.

В соответствии с принятой схемой аннотации был размечен корпус из 868 предложений, включающий в себя выступления политиков, материалы СМИ и художественную прозу, см. рисунок 4 со схемой разметки отрывка из «Красной Шапочки».

Once upon a time there was a dear little girl who was loved by everyone who looked at her, but most of all by her grandmother, and there was nothing that she would not have given to the child.

Once she (SAME_HEAD) gave her (DIFFERENT_HEAD) a little riding hood of red velvet (OTHER_NONREFERENTIAL), which suited her (SAME_HEAD) so well that (NONUNIQUE_HEARER_NEW_SPEC)

she (SAME_HEAD) would never wear anything else (QUANTIFIED); so she (SAME_HEAD) was always called 'Little Red Riding Hood.' (UNIQUE_HEARER_NEW)

Рис. 4. Пример разметки корпуса для обучения модели в [Bhatia et al. 2014]

Далее, для каждого из такого восьмизначного атрибута авторы работы строят классификатор, который предсказывает, относится ли именная группа к классу рассматриваемого атрибута или нет. Предсказывающими параметрами (предикторами), или, как их называют авторы работы, **перцептами** (percepts), в данном случае служат синтаксические характеристики именной группы, получаемые с помощью автоматического синтаксического анализа корпуса в формализме грамматики зависимостей, а также автоматического разрешения кореференции. Перцепты представляют собой данные трех типов. Во-первых, были использованы сами словоформы, их морфологические характеристики и леммы этих словоформ. Кроме того, использовались и другие словоформы, связанные с рассматриваемым именем, — его вершина, зависимые и главное слово (governor), ближайший глагол, с которым имя связано отношениями зависимости, а также вспомогательный глагол и отрицательная частица, если они были. Во-вторых, использовались структурные перцепты: длина зависимости от имени до основного узла и до ближайшего глагола, количество зависимых членов и количество зависимостных отношений. В-третьих, использовались позиционные свойства: длина именной группы, положение именной группы в предложении (первая или вторая половина), положение глагола, от которого зависит именная группа (слева или справа). Наконец, учитывался еще и контекст — для каждой рассматриваемой именной группы рассматривались также другие именные группы: непосредственный «родитель», непосредственный «ребенок», ближайшая предшествующая и последующая именные группы. Для всех контекстуальных именных групп также извлекаются основные перцепты. Входными данными для обучения являются именные группы корпуса, размеченные по атрибутам и связанные с набором перцептов. В результате применения лог-линейной модели предсказывается вероятность семантических характеристик (восьми значений атрибута), связанных с определенностью именной группы на основании набора лексических, синтаксических и контекстуальных характеристик (перцептов). Как и в работах [Bresnan et al. 2007] и [Bresnan, Ford 2010], особое место в исследовании занимает оценка того, насколько качественно работает модель, — ведь именно благодаря ей можно понять, насколько надежны будут теоретические выводы о значимости тех или иных лингвистических параметров для коммуникативной определенности. Авторы работы [Bhatia et al. 2014] используют несколько мер для оценки своей модели. Для сравнения они используют «золотой стандарт», размеченный корпус, который не участвовал в обучении. Предсказания, сформированные для этого корпуса, сравниваются с заранее приписанной разметкой «эталонной» категорией. Кроме общей оценки, авторам важна также и индивидуальная оценка качества распознавания каждого атрибута — так, можно понять, какие именно семантические категории можно надежно предсказывать. Еще одна оценка — soft match, так называемое мягкое совпадение, — использовалось для того, чтобы учитывать

и закрыл окно. Мы понимаем, что окно находится в той самой комнате, в которую вошел Петя. Эта связь носит название бриджинга.

те атрибуты, которые были предсказаны не полностью, т. е. не все из восьми семантических параметров предсказанного атрибута совпали с эталонным. Результаты оценки отдельных атрибутов (leaf labels) представлены в таблице 1. В первой числовой колонке таблицы отражено количество вхождений каждого атрибута в тренировочный корпус — тот корпус, откуда были взяты данные для обучения модели. Во второй и третьей колонках таблицы представлены основные метрики оценки в процентном выражении: Precision, точность, и Recall, полнота. Параметр точности показывает нам меру того, насколько много ошибок было допущено в предсказании атрибута, например стопроцентная точность говорит о том, что все предсказания атрибутов данной категории были сделаны верно. Параметр полноты определяет то, насколько много было допущено пропусков. Так, уровень в 78 % в первой строке таблицы говорит о том, что 22 % случаев вхождения атрибута не были опознаны. Мера F — гармоническое среднее — показывает усредненное значение двух показателей. Из таблицы 1 видно, что для некоторых атрибутов явно не хватило данных в обучаемой выборке, а некоторые, например «bridging other context», распознаются очень плохо, так как совершенно не связаны с теми лингвистическими параметрами, которые подаются модели на вход. Действительно, связь определенности имени с внешним контекстом за пределами рассматриваемого предложения не определяется теми перцептами, которые использовались для построения модели. Авторы признают, что для того, чтобы строить предсказания таких атрибутов, нужно использовать более сложные семантические и дискурсивные характеристики. Кроме того, авторы замечают, что ряд проблем возникает из ошибок автоматической разметки. Так категории Same_head и Different_head распознаются очень плохо, несмотря на хороший объем обучающего материала. Причина неудачи видится авторам в плохой работе парсера, устанавливающего кореферентные связи, который был использован при предварительной обработке корпуса.

Таблица 1

Результаты работы модели [Bhatia et al. 2014]

Leaf label	No. of Instances	Prec.	Recall	F_1
Pleonastic	44	100	78	88
Bridging_Restrictive_Modifier	552	58	84	68
Quantified	213	57	57	57
Unique_Larger_Situation	97	52	58	55
Same_Head	452	41	41	41
Measure_Nonreferential	98	88	26	40
Nonunique_Hearer_New_Spec	190	36	46	40
Other_Nonreferential	134	39	36	37
Different_Head	271	32	33	32
Nonunique_Larger_Situation	97	29	25	27
Part_of_Noncompositional_MWE	88	20	17	18
Bridging_Nominal	33	33	10	15
Generic_Individual_Level	113	14	11	13
Nonunique_Nonspec	173	9	25	13
Bridging_Other_Context	96	33	6	11
Bridging_Event	9	—	0	—
Nonunique_Physical_Copresence	36	0	0	—
Nonunique_Predicative_Identity	10	—	0	—
Predicative_Nonidentity	57	0	0	—
Unique_Hearer_New	26	—	0	—

Преимущество модели состоит в том, что она дает возможность увидеть неожиданные связи между морфосинтаксическими свойствами именной группы и семантическими признаками, составляющими категорию определенности и неопределенности. Так, выясняется, что, кроме вполне ожидаемых грамматических признаков специфичности, таких как артикли, посессивные местоимения, имена собственные, местоимения второго лица, важным

фактором оказывается зависимость имени от предлога *from*. С другой стороны, именные группы, зависящие от прилагательных в сравнительной степени, почти всегда имеют характеристику неспецифичности. Выдвигая еще несколько неожиданных гипотез, авторы признают, что не могут точно сказать, насколько они действительно отражают общую тенденцию или же связаны с отклонениями в тренировочном корпусе. Для того, чтобы это проверить, нужны дополнительные, более обширные данные.

Наконец, интересным результатом работы предсказательной модели является эксперимент по сокращению обучающих характеристик (перцептов). Оказывается, например, что устранение из моделей артиклей не снижает существенным образом результатов предсказаний. Авторы делают вывод о том, что категория коммуникативной определенности в английском языке на самом деле гораздо богаче ее грамматического выражения.

Рассматриваемая работа является, по сути, первым шагом к анализу весьма сложной семантической категории определенности с помощью методов машинного обучения. Авторы считают своим важным достижением то, что им удалось показать многофакторность этой категории. В то же время заметим, что, в отличие от работ Дж. Бреснан и ее коллег, рассматриваемая работа все-таки ближе к инженерной модели машинного обучения, нежели к лингвистическому исследованию. Тем не менее, она дает представление о том, каким образом может быть продолжено изучение связанности семантических параметров определенности и морфосинтаксических характеристик имени.

Работа [Кибрик и др. 2010] посвящена применению машинного обучения для моделирования референциального выбора. Авторы указывают, что предложенная модель продолжает модели, которые были сделаны раньше [Kibrik 1996; 1999; Grüning, Kibrik 2003; 2005], однако ее отличие в том, что она построена на данных существенно большего объема. Корпус, положенный в основу исследования, имеет сложную разметку, включающую в себя разметку риторической структуры в рамках формализма, предложенного в работе [Mann, Thompson 1988], разметку референциальных выражений, используемых в анафорической функции, и их свойств, а также разметку именных или предложных групп, выполняющих функцию антецедента, и их характеристик. Между маркабулами — аннотируемыми элементами — определяются отношения кореферентности. Каждое такое отношение соединяет два аннотируемых элемента — анафор и его антецедент. Задачей предсказательной модели являлось определение формального выражения анафора: полной именной группы или же местоимения. Обучение модели происходило с помощью наборов распределений признаков, извлеченных из корпуса. Признаки включали в себя, во-первых, общие свойства референта — первое или не первое упоминание в дискурсе, одушевленность и протагонизм, во-вторых, грамматические и позиционные признаки антецедента и анафора: входит ли антецедент в состав прямой речи, тип синтаксической фразы, грамматическая роль, также отдельно для антецедента указывалась референциальная форма (например, определенность). Наконец, последним типом признаков были свойства кореферентной связи: расстояние в словах, расстояние в маркабулах между анафором и антецедентом, линейное расстояние в клаузах, риторическое расстояние в элементарных дискурсивных единицах по Манну и Томпсон. Для машинного обучения были выбраны логические алгоритмы классификации, позволяющие интерпретировать получаемые классы признаков, и логистическая регрессия. Как уже говорилось выше, преимущество логистической регрессии состоит в том, что она позволяет увидеть значимость отдельных факторов. Оценка качества предсказания модели показывает уровень в 85 %. Авторы сравнивают результаты работы модели с анализом выбора, совершаемого носителями, и приходят к выводу, что существует класс случаев, в котором выбор между местоимением и именной группой равновероятен. В этом случае предсказательный потенциал системы упирается в потолок. По аналогии с инженерными компьютерными лингвистическими моделями можно сказать, что точность оценки вероятности того или иного лингвистического явления (в нашем случае референциального выбора) не может быть выше уровня согласия носителей языка в том, к какому классу это явление относится (так называемый параметр interannotator agreement). В нашем случае это значит,

что при определенных условиях говорящий имеет более одной референциальной опции. Основной вывод авторов работы согласуется с выводом, сделанным в исследованиях, рассмотренных выше: оценка вероятности использования местоимения является аналогом когнитивного по своей природе коэффициента активации — «интегрального показателя, представляющего собой равнодействующую всех одновременно действующих факторов активации» [Кибрик и др. 2010: 179].

Итак, мы рассмотрели три работы, в которых были использованы алгоритмы машинного обучения для анализа многофакторных языковых явлений. Основные тенденции этого подхода, которые, очевидно, получат в ближайшем будущем широкое развитие, состоят в том, что классификатор, обученный на статистике распределения лингвистических признаков, используется для **предсказания**. В рассмотренных нами работах предсказание строилось от семантических признаков к ограниченному выбору формального выражения. Интуитивно такой тип предсказания наиболее близок к выбору лингвистического выражения, который делает говорящий. Для того чтобы поддержать это содержательное сходство, авторы использовали результаты дополнительных экспериментальных данных, полученных при работе с носителями языка. Такая параллель чрезвычайно важна, поскольку дает возможность рассматривать исследование не как инженерную конструкцию, пусть и отличающуюся высоким качеством, но как экспериментальное исследование, позволяющее получить новые знания о когнитивных процессах. Собственно, на сегодняшний момент мы можем видеть три области, которые могут быть исследованы в полной мере только с помощью предсказывающих технологий. Это, во-первых, **взвешивание признаков** при их взаимодействии и связь этих весов с результирующим предсказанием. Логистическая модель на сегодняшний день не считается самой эффективной с точки зрения достижения оптимального результата, однако это модель, дающая исследователю доступ в «черный ящик». Во-вторых, это **оценка значимости признаков** для предсказания результирующего выбора: появилась возможность измерить, насколько ухудшится предсказательная сила модели, если устранить тот или иной признак. Представляется, что именно у этого экспериментального метода очень большое будущее в области типологических исследований. Нельзя исключить, что модели, обученные на одинаковых мультиязыковых корпусах, будут показывать разные результаты при устранении того или иного признака. Иными словами, давно наблюдаемый факт того, что разные языки по-разному определяют важность или же несущественность тех или иных семантических параметров, получит не только объективное подтверждение, но и вполне конкретные измерители. Именно с **измерителями** связано, на мой взгляд, третье направление развития предсказательного подхода. В работе [Bresnan, Ford 2010] было показано, как одна и та же модель, натренированная на ответах носителей разных вариантов языка, давала разные результаты. Что принципиально важно, модель позволяет чрезвычайно точно определить, чем и насколько эти результаты отличаются. Иными словами, открывается возможность точного описания и содержательного сравнения всего многообразия языковой вариативности — форм, стилей, диалектов и идиолектов, причем именно в областях синтаксиса, семантики и прагматики, которые при традиционном подходе чрезвычайно трудно уловить и перевести в исчисляемые параметры.

3. Большие данные и ранжирование языков

Три работы, рассмотренные выше в части 2, различаются источниками данных, которые используются в исследовании. В работе [Bresnan, Ford 2010] речь идет фактически о базе данных, в которой каждый вход (предложение) снабжен определенным дескриптором. Модель, предложенная в [Кибрик и др. 2010], использует аннотированный вручную корпус с разметкой весьма сложной структуры. Наконец, в наиболее инженерно ориентированной работе [Bhatia et al. 2014] для построения модели используются автоматически размеченные корпуса. Именно источникам данных и возможности применения для получения лингвистических выводов автоматически размеченных корпусов будет посвящен настоящий раздел.

Развитие статистического направления в компьютерной лингвистике было бы невозможно без постоянного производства массивов данных — обучающих корпусов. Как правило, тексты для обучающих корпусов берутся из уже имеющихся корпусов, так например, самый известный корпус синтаксической разметки Пенсильванского университета Penn TreeBank использует для разметки не менее известный Brown corpus и корпус телефонных разговоров Switchboard. Однако очень часто в качестве материала разметки берется архив новостных сообщений или газетных статей, например, корпус Wall Street Journal, который был использован и в рассмотренной выше работе [Кибрик и др. 2010]. В определенный момент возникает практика выкладывания обучающих корпусов в открытый доступ, в частности благодаря распространению так называемых форматов Shared Tasks на конференциях, посвященных технологиям автоматической обработки языка. Shared Tasks представляют собой соревнования, к участию в которых приглашаются команды, разработавшие технологии по тематике объявленного задания. Задания могут быть самые разнообразные, и, надо сказать, с каждым годом они связываются со все более сложными лингвистическими явлениями и классификациями. Так, если первыми соревнованиями еще в 90-х годах были соревнования по автоматическому морфологическому анализу (см. например, краткий обзор таких соревнований в [Ляшевская и др. 2010]), далее весьма популярны были соревнования по автоматическому разрешению лексической многозначности (см., например, историю форумов Senseval и впоследствии Semeval в [Agirre, Edmonds 2007]), лингвистические темы в соревнованиях последних лет связаны с семантикой, грамматикой и синтаксисом. Так, в 2006—2009 гг. на конференции CoNLL (Natural language learning) проходили соревнования по автоматическому распознаванию синтаксических, а затем и семантических зависимостей, причем особенностью этих соревнований с 2007 г. стал мультиязычный обучающий корпус [Nilsson et al. 2007; Surdeanu et al. 2008; Hajič et al. 2009]. В результате участники, используя обучающие корпуса соревнований, должны были выдать размеченные тестовые корпуса на нескольких десятках языков, а эксперты имели возможность оценить в том числе и то, какие языки в среднем лучше поддаются автоматическому анализу в рамках предложенной задачи, а какие хуже, что само по себе является небезынтересным лингвистическим фактом. Соревнования форума CoNLL в 2010 г. были связаны с автоматическим определением эпистемической модальности [Farkas et al. 2010], в 2011 и 2012 гг. — с автоматическим моделированием кореферентных связей [Pradhan et al. 2011; Pradhan et al. 2012], в 2013 г. Shared Task был посвящен автоматическому исправлению грамматических ошибок [Ng et al. 2013]. Тематика соревнований форума Semeval в последние годы тоже исключительно разнообразна: объектом автоматического моделирования стали такие неэлементарные лингвистические явления, как семантические роли, метонимия, временная референция, пространственные отношения и др. Причем для большинства этих явлений соревнования проводились на мультиязычных данных. Наконец, нельзя не отметить, что с 2010 г. соревнования парсеров проводятся и в России на материале русского языка как часть программы международной конференции по компьютерной лингвистике «Диалог». Так, были проведены соревнования по морфологическому, синтаксическому анализу, по задачам разрешения референциальной неоднозначности, по лексической близости [Ляшевская и др. 2010; Толдова и др. 2012; Toldova et al. 2014]. Принципиально важно для темы настоящего обзора, что основным продуктом всех подобных мероприятий являются корпуса с разметкой, соответствующей тематике соревнования, находящиеся в открытом доступе. Эти корпуса используются в дальнейшем для обучения автоматических систем, но, как представляется, их потенциал в теоретических лингвистических исследованиях остается пока чрезвычайно мало оцененным. Фактически на сегодняшний день уже имеется огромный мультиязычный массив данных, связанных с самыми разными лингвистическими объектами и явлениями. Корпуса с разметкой могут быть использованы и непосредственно как источник статистических данных о тех или иных явлениях, и как тренажер для обучения классификатора, которым исследователь впоследствии мог бы обработать тот корпус текстов, который его интересует. Ниже я чуть подробней рассмотрю две работы, в которых

теоретические импликации базируются на корпусных данных, «оставшихся» от обучения автоматических модулей.

Работа [Futrell et al. 2014] была представлена на конференции AMLaP (“Architecture and Mechanisms for Language Processing”) в 2014 г. в Эдинбурге. Авторы поставили перед собой задачу, используя материал мультиязыковых корпусов соревнования CoNLL 2007, проверить взаимосвязь между некоторыми синтаксическими характеристиками разных языков, а именно свободным порядком слов и длиной зависимостей. Принципиально отличает эту работу от работ по квантитативным типологическим исследованиям то, что использование корпусов дает возможность использовать не бинарные меры (свободный / несвободный порядок слов), а градуальные. Используя корпуса с размеченными деревьями зависимостей, можно подсчитать в числовом выражении длины зависимостей и уровни свободы порядка слов (в данном случае энтропию отклонения от нейтрального порядка слов). Такие подсчеты дают возможность представлять полученные данные в виде шкал и считать более сложные корреляции.

Авторы работы отталкиваются от двух гипотез, которые были сформулированы ранее в теоретических исследованиях. Во-первых, считается, что имеется корреляция между свободным порядком слов и наличием в языке падежей. Вопрос, который ставится в данном случае, звучит так: связана ли большая свобода порядка слов с большим количеством падежей? Во-вторых, исследуется длина зависимостной цепочки. Опять же опираясь на предшествующие теоретические работы, исследователи проверяют гипотезу о том, что по причинам, связанным с обработкой высказываний, развитие языков характеризуется постепенным уменьшением линейной длины зависимостной цепочки (длины от вершины к ее зависимому). Таким образом, второй задачей является определить среднюю длину зависимости. И наконец, итоговая цель исследования — посмотреть, насколько велика связь между свободой порядка слов в языке и увеличенной длиной зависимости. В таблице 2 представлен список из 34 языков, на данные которых опиралось исследование. Как мы видим, корпуса имеют неравномерный размер, однако в основном они превышают объем в сто тысяч словоформ. Поскольку исследуются явления, которые есть в каждом предложении, то такой объем представляется достаточным.

Таблица 2

Языковые корпуса, использованные в работе [Futrell et al. 2014]

Language	# Tokens	Source	Family / Region
English	470 367	HamleDT	IE / West Germanic
Dutch	214 389	HamleDT	IE / West Germanic
German	929 454	HamleDT	IE / West Germanic
Swedish	208 554	HamleDT	IE / North Germanic
Danish	105 750	HamleDT	IE / North Germanic
Spanish	493 794	HamleDT	IE / Romance
Catalan	458 241	HamleDT	IE / Romance
Portuguese	221 904	HamleDT	IE / Romance
French	412 933	UDT	IE / Romance
Italian	79 654	UDT	IE / Romance
Romanian	40 192	HamleDT	IE / Romance
Latin	56 616	HamleDT	IE / Classical
Ancient Greek	330 255	HamleDT	IE / Classical
Modern Greek	73 125	HamleDT	IE / Greek
Czech	1 591 651	HamleDT	IE / West Slavic

Продолжение табл. см. на с. 114

Продолжение табл. со с. 113

Language	# Tokens	Source	Family / Region
Slovak	958 706	HamleDT	IE / West Slavic
Slovenian	38 552	HamleDT	IE / South Slavic
Bulgarian	209 372	HamleDT	IE / South Slavic
Russian	532 360	HamleDT	IE / East Slavic
Persian	202 027	HamleDT	IE / Iranian
Hindi	307 783	HamleDT	IE / Indic
Bengali	8 381	HamleDT	IE / Indic
Basque	162 818	HamleDT	Isolate
Finnish	62 883	HamleDT	Finno-Ugric / Finnic
Estonian	10 806	HamleDT	Finno-Ugric / Finnic
Hungarian	145 567	HamleDT	Finno-Ugric / Ugric
Turkish	70 677	HamleDT	Turkic
Hebrew	162 500	HamleDT	West Semitic
Arabic	284 970	HamleDT	West Semitic
Tamil	10 181	HamleDT	Dravidian
Telugu	7 172	HamleDT	Dravidian
Indonesian	127 516	UDT	Austronesian
Japanese	174 925	UDT	East Asian / Isolate
Korean	76 029	HamleDT	East Asian / Isolate

Авторы используют формулу условной энтропии для того, чтобы подсчитать вариативность порядка слов, обусловленного типом синтаксических отношений. Фактически это вычисление можно интерпретировать как уровень неуверенности того, каким является порядок внутри зависимостного отношения. Первая, чрезвычайно интересная, диаграмма отражает энтропию ветвления: измеряется то, насколько легко предсказать позицию вершины, т. е., по сути, насколько позиция вершины следует правилам нейтрального порядка слов, см. рисунок 5 (с. 115).

Мы видим, что верхние и нижние строки списка выглядят совершенно ожидаемо: японский и корейский максимально предсказуемы с точки зрения положения вершины, а латынь и древнегреческий максимально свободны в смысле порядка слов. Наиболее информативна и интересна средняя часть списка, которая может быть получена только с помощью применения квантитативных методов к синтаксически размеченным корпусам.

На второй диаграмме отражена степень энтропии для порядка подлежащего и прямого дополнения, см. рисунок 6 (с. 115). Заметим, что для составления такой диаграммы не требуются сведения о том, каков порядок этих синтаксических элементов в языке. Источником данных здесь является стабильность и предсказуемость для любого порядка.

На диаграмму также нанесено деление языков на группы по количеству падежей (для этого использовалась информация из грамматик). Делается вывод о том, что языки со свободным порядком слов обязательно имеют богатую падежную систему, но в то же время многие языки, имеющие падежи, демонстрируют низкую вариативность порядка слов.

Результатом исследования является график, показывающий слабую позитивную корреляцию между длиной зависимости и свободным порядком слов (из графика исключены латынь и древнегреческий). Для сравнения данных авторы работы используют технику минимизации длины зависимости, предложенную в [Gildea, Temperley 2007], — алгоритм линеаризации зависимостей, суть которого заключается в том, что по определенным правилам сложные зависимостные деревья представляются как сумма небольших деревьев с минимальными расстояниями от главного слова к зависимому. Алгоритм линеаризации

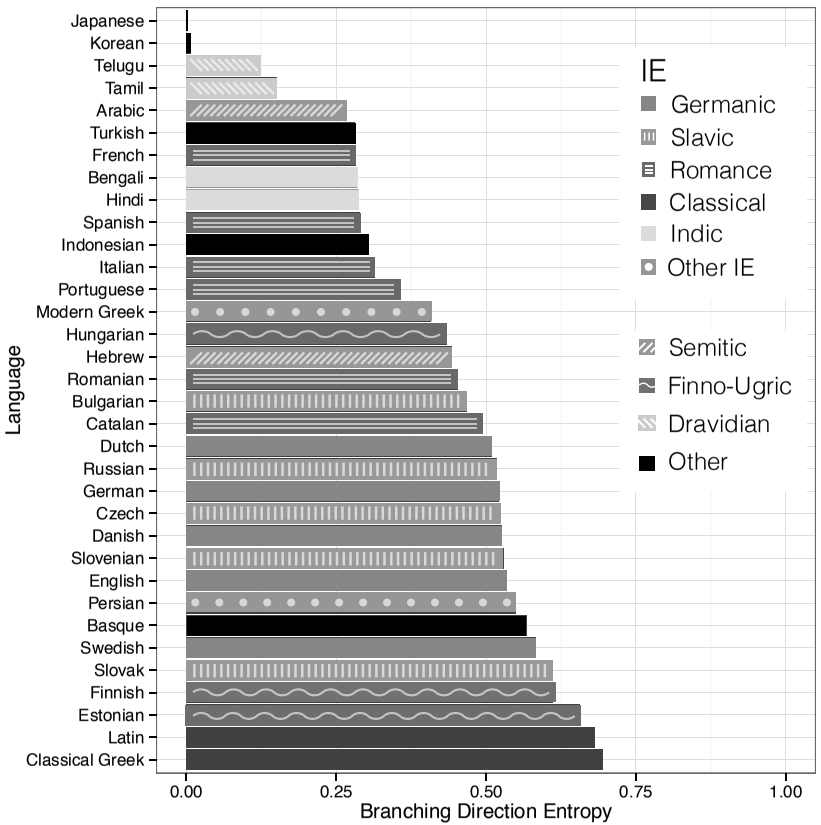

Рис. 5. Шкала языков по уровню варьирования направления ветвления [Futrell et al. 2014]

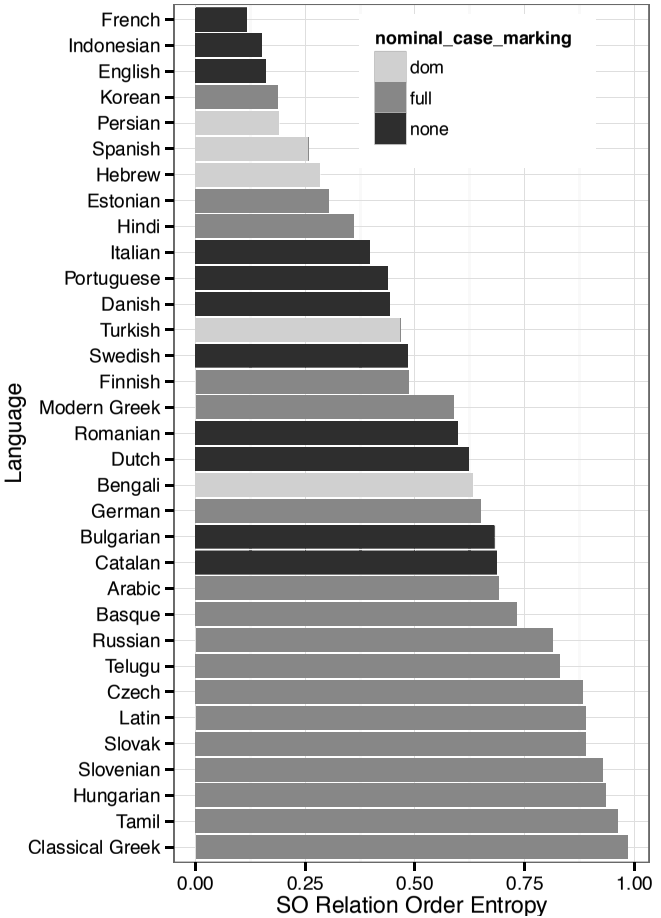

Рис. 6. Шкала языков по уровню варьирования порядка слов для подлежащего и прямого дополнения [Futrell et al. 2014]

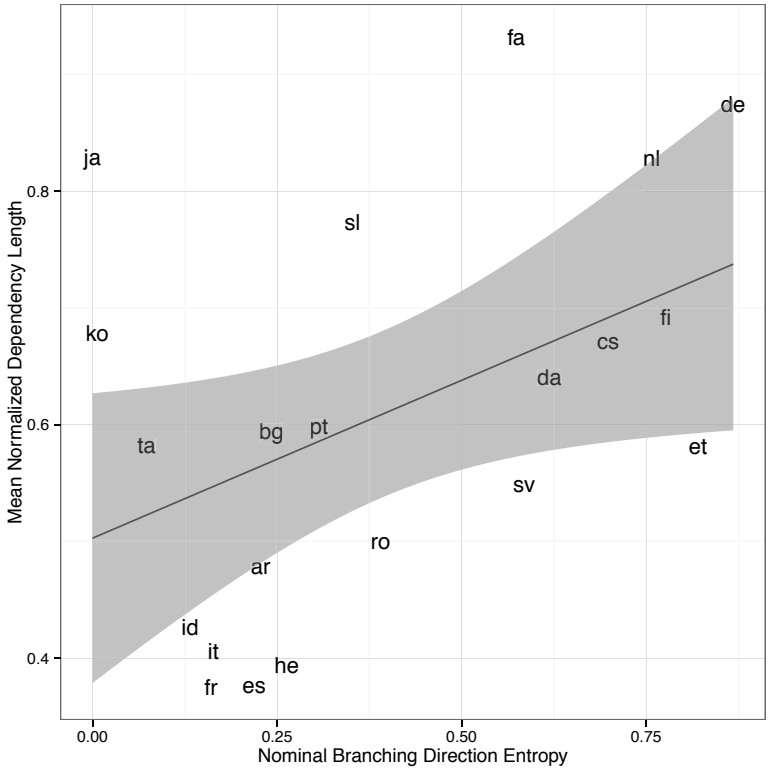

Рис. 7. График соотношения энтропии направления ветвления и средней нормализованной длины зависимости [Futrell et al. 2014]

позволяет представить данные каждого языка в трех графиках: средней длины зависимости, минимальной длины линеаризованной зависимости и случайной длины линеаризованной зависимости. Такое сравнение задает своего рода шкалу, распределение по которой различает языки. Эта шкала сопоставляется со шкалой варьирования порядка слов.

На графике (см. рисунок 7) оси абсцисс соответствует возрастание показателя уровня энтропии, отражающего уровень варьирования порядка слов, а ось ординат отражает шкалу трех длин зависимостей, полученных для каждого языка, которая нормирована до диапазона между 0 и 1. Авторы замечают, что слабая положительная корреляция не совсем соответствует концепции о том, что длина зависимости должна со временем стремиться к уменьшению: у языков со свободным порядком слов имеется склонность к увеличению длины зависимости.

Предварительные наблюдения, сделанные в работе [Futrell et al. 2014], безусловно ценны методологической новизной. Авторам удалось показать новый метод, дающий возможность построения языковых шкал типологических признаков, позволяющий также оценить взаимосвязь этих признаков между собой. Особо интересным в предложенном методе является то, что он дает континуальный способ представления типологических данных, не противопоставляя жестко языки по наличию того или иного признака, но лишь ранжируя их. Можно предположить, что при таком подходе менее остро стоит проблема адекватности языковой выборки для типологического исследования. При традиционном делении языков на две или несколько групп по наличию некоторого параметра (например, жесткого или свободного порядка слов) добавление новых языков в выборку меняет отношение между объемами групп в целом. В случае ранжирования на плавно возрастающей кривой каждый из языков является в некотором смысле отдельной группой. Добавление нового языка меняет изгиб кривой только в области ближайших соседей, не затрагивая другие части списка.

Еще один интересный опыт использования эталонного тестового корпуса в академическом исследовательском проекте представлен в работе [Nedoluzhko et al. 2015], посвященной сравнительному исследованию кореферентных цепочек в трех языках — английском, чешском и русском. Авторы используют материалы Пражского чешско-английского параллельного корпуса (Prague Czech-English Treebank (PCEDT)), имеющего три уровня разметки, в том числе и разметку кореферентных связей, а также корпус текстов с разметкой анафоры, созданный для проведения соревнования по автоматическому разрешению кореферентной неоднозначности в 2014 г. [Toldova et al. 2014]. Для исследования было взято примерно одинаковое количество новостных текстов на каждом из трех языков. Авторы сравнили частотность длин и структурных элементов кореферентных цепочек в текстах и показали, что за различиями в выборе оформления кореферентных связей стоят различия в предпочтениях тех или иных синтаксических структур в рассматриваемых языках. Оказалось, что три языка незначительно различаются по параметру длины цепочки: доли коротких и длинных цепочек примерно одинаковы в английском, чешском и русском. В то же время сравнение структурного состава цепочек обнаружило интересные особенности. Так, например, сравнительный анализ представленности разных типов маркабул — размеченных элементов кореферентных цепочек — показал, что в английском языке анафорические местоимения в позиции подлежащего составляют 8,6 % от общего числа маркабул, в русском языке — 3,8 %, а в чешском только 0,1 % от всех маркабул. И русский, и чешский языки позволяют subject pro-drop — опущение местоименного подлежащего, однако для чешского языка такая стратегия является основной, а в русском эта синтаксическая возможность оказывается менее частотной. Как замечают авторы, можно сказать, что русский менее pro-drop-язык, чем чешский. Развивая этот подход, можно было бы, расширив языковую выборку, построить ранжирование языков по степени свободы опущения местоимения по аналогии с рассмотренной выше иерархией языков по варьированию порядка слов.

Другой интересный вывод из результатов сравнения кореферентных цепочек связан с использованием релятивного местоимения (для русского языка это *который*). Выясняется, что ранжирование языков по этому параметру устроено следующим образом. Наиболее часто релятивная анафора употребляется в чешском языке (8,5 %), следующим за ним идет английский язык (6,9 %), в русском языке кореферентные цепочки с *который* встречаются лишь в 4,1 % случаев. Иначе говоря, мы видим, что два родственных языка ведут себя поразному. Авторы предполагают, что такое расхождение связано с различиями структур клауз в рассматриваемых языках. Русский язык предпочитает релятивным оборотам нефинитные формы глагола, в то время как в чешском чаще встречаются относительные предложения. Эта гипотеза подтверждается при сопоставлении частотностей цепочек, включающих в себя инфинитивные конструкции с кореферентным, но не восстановимым на поверхностном уровне синтаксическим нулем (PRO) — см. таблицу 3, в которой иллюстрируется соотношение между финитными и нефинитными клаузами в английском, русском и чешском языках.

Таблица 3

Сравнение частотных характеристик финитных и нефинитных конструкций в трех языках [Nedoluzsko et al. 2015]

	Czech	**English**	**Russian**
number of finite clauses	1166	1005	663
number of nonfinite clauses	97	200	379

Как видно из таблицы 3, в этой мини-иерархии чешский является наиболее «финитным» языком, а русский наиболее «нефинитным». При этом авторы указывают на некоторую смещенность данных, поскольку имеют дело с параллельным корпусом, в котором тексты переводились с английского на чешский: возможно, структура предложения языка оригинала могла влиять на структуру переведенного предложения. Иначе говоря, при сравнении

самостоятельных текстов можно ожидать от чешского еще большей частотности финитных клауз.

Рассматриваемое исследование является пока только пилотным, однако в нем сделан очень важный шаг к выработке новых инструментов и методов лингвистических измерений. Распространение корпусов с глубоким аннотированием, включающим в себя не только морфологическую, но и синтаксическую, семантическую и анафорическую разметку, позволяет продвинуть методы корпусной лингвистики от лексического и морфологического уровня к изучению явлений более абстрактных уровней. Появление достаточного количества мультиязычных размеченных корпусов дает возможность использовать новые методы для сравнения языков: кроме деления языков на классы по наличию или отсутствию определенного грамматического признака (ср. принцип языковых карт и описаний в атласе WALS [Dryer, Haspelmath 2013]) появляется возможность определить степень выраженности этого признака с помощью числовой величины и упорядочить языковую выборку относительно этой величины.

4. Заключение

В обзоре ставилась задача показать на примере нескольких работ новые методы и подходы к исследованию теоретических вопросов лингвистики. Эти методы успешно применяются для решения задач компьютерной лингвистики, более того, они постоянно развиваются и совершенствуются. Одним из преимуществ их распространения является их доступность. Уже сейчас существует достаточно много пакетов и сред, позволяющих использовать методы машинного обучения, не владея специальным языком программирования[3]. С каждым годом в открытый доступ выкладывается все больше корпусов на разных языках, снабженных лингвистической разметкой, а также инструменты (парсеры), с помощью которых исследователь может разметить собственный корпус текстов. Так, например, в рассмотренной выше работе [Bhatia et al. 2014] экспериментальное исследование строилось на разметке, полученной в результате автоматического синтаксического и анафорического анализа собственного корпуса. Использование ресурсов и алгоритмов компьютерной лингвистики дает возможность получить количественные измерители «скрытых» явлений, таких, которые не могут быть получены с помощью прямого запроса к корпусу. Представляется, что принципиально новым свойством этих измерителей является их относительность: их суть состоит в том, что они количественно оценивают (предсказывают) роль отдельных факторов в сложных многофакторных процессах, обуславливающих то или иное поверхностное формальное выражение.

СПИСОК ЛИТЕРАТУРЫ / REFERENCES

Кибрик и др. 2010 — Кибрик А. А., Добров Г. Б., Залманов Д. А., Линник А. С., Лукашевич Н. В. Референциальный выбор как многофакторный вероятностный процесс // Компьютерная лингвистика и интеллектуальные технологии. По материалам ежегодной Международной конференции «Диалог». М.: Изд-во РГГУ, 2010. Т. 9 (16). С. 173—179. [Kibrik A. A., Dobrov G. B., Zalmanov D. A., Linnik A. S., Lukashevich N. V. Referential choice as multifactorial probabilistic process. *Komp'yuternaya lingvistika i intellektual'nye tekhnologii. Po materialam ezhegodnoi Mezhdunarodnoi konferentsii «Dialog»*. Moscow: Russian State Univ. for the Humanities, 2010. Vol. 9 (16). Pp. 173—179.]

Ляшевская и др. 2010 — Ляшевская О. Н., Астафьева И., Бонч-Осмоловская А., Гарейшина А., Гришина Ю., Дьячков В., Ионов М., Королева А., Кудринский М., Литягина А., Лучина Е., Сидорова Е., Толдова С., Савчук С., Коваль С. Оценка методов автоматического анализа текста: морфологические парсеры русского языка // Компьютерная лингвистика и интеллектуальные технологии. По материалам ежегодной Международной конференции «Диалог». М.: Изд-во РГГУ, 2010. Т. 9 (16). С. 318—326. [Lyashevskaya O. N., Astaf'eva I., Bonch-Osmolovskaya A., Gareishina A.,

[3] См., например, среду weka (http://www.cs.waikato.ac.nz/ml/weka/) или rapidminer (https://rapidminer.com/). Обе среды приспособлены для проведения экспериментов, имеют графические интерфейсы и очень подробные описания.

Grishina Yu., D'yachkov V., Ionov M., Koroleva A., Kudrinskii M., Lityagina A., Luchina E., Sidorova E., Toldova S., Savchuk S., Koval' S. Assessment of methods of text automatic analysis: Morphological parcers of the Russian language. *Komp'yuternaya lingvistika i intellektual'nye tekhnologii. Po materialam ezhegodnoi Mezhdunarodnoi konferentsii «Dialog»*. Moscow: Russian State Univ. for the Humanities, 2010. Vol. 9 (16). Pp. 318—326.]

Толдова и др. 2012 — Толдова С. Ю., Соколова Е. Г., Астафьева И., Гарейшина А., Королева А. Н., Привознов Д., Сидорова Е., Тупикина Л., Ляшевская О. Н. Оценка методов автоматического анализа текста 2011—2012: синтаксические парсеры русского языка // Вестник компьютерных и информационных технологий. 2012. № 8. С. 14—19. [Toldova S. Yu., Sokolova E. G., Astaf'eva I., Gareishina A., Koroleva A. N., Privoznov D., Sidorova E., Tupikina L., Lyashevskaya O. N. Assessment of methods of text automatic analysis: Syntactic parcers of the Russian language. *Vestnik komp'yuternykh i informatsionnykh tekhnologii*. 2012. No. 8. Pp. 14—19.]

Толдова, Ляшевская 2014 — Толдова С. Ю., Ляшевская О. Н. Современные проблемы и тенденции компьютерной лингвистики // Вопросы языкознания. 2014. №. 1. С. 120—145. [Toldova S. Yu., Lyashevskaya O. N. Contemporary issues and trends in computational linguistics. *Voprosy jazykoznanija*. 2014. No. 1. Pp. 120—145.]

Agirre, Edmonds 2007 — Agirre E., Edmonds P. G. (eds). *Word sense disambiguation: Algorithms and applications* (*Text, speech and language technology*. Vol. 33). Dordrecht: Springer Science & Business Media, 2007.

Bhatia et al. 2014 — Bhatia A., Lin Ch., Schneider N., Tsvetkov Yu., Talib Al-Raisi F., Roostapour L., Bender J., Kumar A., Levin L., Simons M., Dyer Ch. Automatic classification of communicative functions of definiteness. *Proceedings of the 25th International conference on computational linguistics.* Dublin City University and Association for Computational Linguistics, 2014. Pp. 1059—1070.

Bresnan 2007 — Bresnan J. Is knowledge of syntax probabilistic? Experiments with the English dative alternation. *Linguistics in search of its evidential base, series: Studies in generative grammar.* Featherston S., Sternefeld W. (eds). Berlin: Mouton de Gruyter, 2007. Pp. 75—96.

Bresnan et al. 2007 — Bresnan J., Cueni A., Nikitina T., Baayen R. H. Predicting the dative alternation. *Cognitive foundations of interpretation*. Boume G., Krämer I., Zwarts J. (eds). Amsterdam: Royal Netherlands Academy of Science, 2007. Pp. 69—94.

Bresnan, Ford 2010 — Bresnan J., Ford M. Predicting syntax: Processing dative constructions in American and Australian varieties of English. *Language*. 2010. Vol. 86. No. 1. Pp. 168—213.

Bresnan, Nikitina 2009 — Bresnan J., Nikitina T. The gradience of the dative alternation. *Reality exploration and discovery*. *Pattern interaction in language and life*. Uyechi L., Wee L.-H. (eds). Stanford: Center for the Study of Language and Information, 2009. Pp. 161—184.

Daume III, Campbell 2007 — Daume III, H. Campbell L. A Bayesian model for discovering typological implications. *Proceedings of the 45th annual meeting of the Association for Computational Linguistics (ACL)*. Prague, June 23—30, 2007. Association for Computational Linguistics, 2007. Pp. 65—72.

Dryer, Haspelmath 2013 — Dryer M. S., Haspelmath M. (eds). *The world atlas of language structures online*. Leipzig: Max Planck Institute for Evolutionary Anthropology, 2013. Available at: http://wals.info, Accessed on 2015-04-10.

Farkas et al. 2010 — Farkas R., Vincze V., Móra G., Csirik J., Szarvas G. The CoNLL-2010 shared task: Learning to detect hedges and their scope in natural language text. *Proceedings of the Fourteenth conference on computational natural language learning: Shared task*. Association for Computational Linguistics, 2010. Pp. 1—12.

Futrell et al. 2014 — Futrell, Mahowald K., Gibson E. *CLIQS: Crosslinguistic investigations in quantitative syntax*. Poster presented at AMLaP 2014.

Gibson, Fedorenko 2013 — Gibson E., Fedorenko E. The need for quantitative methods in syntax and semantics research. *Language and Cognitive Processes*. 2013. Vol. 28. № 1—2. Pp. 88—124.

Gibson et al. 2012 — Gibson E., Piantadosi S. T., Fedorenko E. Quantitative methods in syntax / semantics research: A response to Sprouse and Almeida. *Language and cognitive processes.* 2012. Vol. 28. No. 3. Pp. 229—240.

Gildea, Temperley 2007 — Gildea D., Temperley D. Optimizing grammars for minimum dependency length. *Proceedings of the 45th annual meeting of the Association for Computational Linguistics (ACL)*. Prague, June 23—30, 2007. Association for Computational Linguistics, 2007. Pp. 184—191.

Green 1971 — Green G. Some implications of an interaction among constraints. *Papers from the Seventh regional meeting, Chicago Linguistic Society.* Chicago: Chicago Linguistic Society, The University of Chicago, 1971. Pp. 85—100.

Green 1974 — Green G. *Semantics and syntactic regularity*. Bloomington: Indiana University Press, 1974.

Grüning, Kibrik 2003 — Grüning A., Kibrik A. A. A neural network approach to referential choice. *Komp'yuternaya lingvistika i intellektual'nye tekhnologii. Trudy mezhdunarodnoi konferentsii «Dialog-2003»*. Moscow: Nauka, 2003. Pp. 260—266. [Компьютерная лингвистика и интеллектуальные технологии. Труды международной конференции «Диалог-2003». М.: Наука, 2003. С. 260—266.]

Grüning, Kibrik 2005 — Grüning A., Kibrik A. A. Modeling referential choice in discourse: A cognitive calculative approach and a Neural Networks approach. *Anaphora processing: Linguistic, cognitive and computational modelling*. Branco A., McEnery T., Mitkov R. (eds). Amsterdam: John Benjamins, 2005. Pp. 163—198.

Hajič et al. 2009 — Hajič J., Ciaramita M., Johansson R., Kawahara D., Martí M. A., Màrquez L., Meyers A., Nivre J., Padó S., Štěpánek J., Straňák P., Surdeanu M., Xue N., Zhang Y. The CoNLL-2009 shared task: Syntactic and semantic dependencies in multiple languages. *Proceedings of the Thirteenth conference on computational natural language learning: Shared task*. Association for Computational Linguistics, 2009. Pp. 1—18.

Kibrik 1996 — Kibrik A. A. Anaphora in Russian narrative discourse: A cognitive calculative account. *Studies in anaphora*. Fox B. (ed.). Amsterdam: John Benjamins, 1996. Pp. 255—304.

Kibrik 1999 — Kibrik A. A. Reference and working memory: Cognitive inferences from discourse observation. *Discourse studies in cognitive linguistics*. Van Hoek K., Kibrik A. A., Noordman L. (eds). Amsterdam: John Benjamins, 1999. Pp. 29—52.

Mann, Thompson 1988 — Mann W. C., Thompson S. A. Rhetorical structure theory: Toward a functional theory of text organization. *Text*. 1988. Vol. 8 (3). Pp. 243—281.

Nedoluzhko et al. 2015 — Nedoluzhko A., Toldova S., Novák M. Coreference chains in Czech, English and Russian: Preliminary findings. *Komp'yuternaya lingvistika i intellektual'nye tekhnologii. Po materialam ezhegodnoi Mezhdunarodnoi konferentsii «Dialog»*. Moscow: Russian State Univ. for the Humanities, 2015. Vol. 14 (21). Pp. 474—486. [Компьютерная лингвистика и интеллектуальные технологии. По материалам ежегодной Международной конференции «Диалог» (Москва, 27—30 мая 2015 г.). М.: Изд-во РГГУ, 2015. Т. 14 (21). С. 474—486.]

Ng et al. 2013 — Ng H. T., Wu S. M., Wu Y., Hadiwinoto C., Tetreault J. The CoNLL 2013 shared task on grammatical error correction. *Proceedings of the Seventeenth conference on computational natural language learning*. Association for Computational Linguistics, 2013. Pp. 1—24.

Nilsson et al. 2007 — Nilsson J., Riedel S., Yuret D. The CoNLL 2007 shared task on dependency parsing. *Proceedings of the CoNLL shared task session of EMNLP-CoNLL*. Association for Computational Linguistics, 2007. Pp. 915—932.

Pinker 1989 — Pinker S. *Learnability and cognition: The acquisition of argument structure*. Cambridge: MIT Press, 1989.

Pradhan et al. 2011 — Pradhan S., Ramshaw L., Marcus M., Palmer M., Weischedel R., Xue N. CoNLL 2011 shared task: Modeling unrestricted coreference in ontonotes. *Proceedings of the Fifteenth conference on computational natural language learning: shared task*. Association for Computational Linguistics, 2011. Pp. 1—27.

Pradhan et al. 2012 — Pradhan S., Moschitti A., Xue N., Uryupina O., Zhang Y. CoNLL-2012 shared task: Modeling multilingual unrestricted coreference in OntoNotes. *Joint conference on EMNLP and CoNLL-shared task*. Association for Computational Linguistics, 2012. Pp. 1—40.

Surdeanu et al. 2008 — Surdeanu M., Johansson R., Meyers A., Màrquez L., Nivre J. The CoNLL-2008 shared task on joint parsing of syntactic and semantic dependencies. *Proceedings of the Twelfth conference on computational natural language learning*. Association for Computational Linguistics, 2008. Pp. 159—177.

Toldova et al. 2014 — Toldova S., Roytberg A., Ladygina A., Vasilyeva M., Azerkovich I., Kurzukov M., Sim G., Gorshkov D., Ivanova A., Nedoluzhko A., Grishina Y. RU-EVAL-2014: Evaluating anaphora and coreference resolution for Russian. *Komp'yuternaya lingvistika i intellektual'nye tekhnologii. Po materialam ezhegodnoi Mezhdunarodnoi konferentsii «Dialog»*. Moscow: Russian State Univ. for the Humanities, 2014. Vol. 13 (20). Pp. 77—90. [Компьютерная лингвистика и интеллектуальные технологии. По материалам ежегодной Международной конференции «Диалог». М.: Изд-во РГГУ, 2014. Т. 13 (20). С. 77—90.]

Wintner 2009 — Wintner S. What science underlies natural language engineering? *Computational Linguistics*. 2009. Vol. 35. No. 4. Pp. 641—644.

Статья поступила в редакцию 07.04.2015.