УДК 81 + 81’32 + 81’33
DOI 10.25205/1818-7935-2018-16-2-19-33

Е. В. Люкина

Национальный исследовательский университет
«Высшая школа экономики»
ул. Мясницкая, 20, Москва, 101000, Россия

eliouki@mail.ru

ИСПОЛЬЗОВАНИЕ УНИВЕРСАЛЬНЫХ ЗАВИСИМОСТЕЙ ПРИ ГРАММАТИЧЕСКОМ РАЗБОРЕ МНОГОЯЗЫЧНОГО ТЕКСТА (НА ПРИМЕРЕ БЕЗЛИЧНОГО ПРЕДИКАТИВА)

Статья посвящена инициативе универсальных зависимостей (УЗ), направленной на создание кросс-лингвистически непротиворечивой схемы грамматического разбора предложения. Цель данной инициативы – упрощение кросс-лингвистических исследований, унификация межъязыковой лингвистической типологии, создание основы для автоматизированных многоязычных систем и универсального кросс-языкового парсера текстов.

В первой части статьи рассматриваются основные проблемы, возникающие при грамматическом разборе многоязычного текста, описываются преимущества унификации языковых признаков, даются цели проекта УЗ, приводится история возникновения проекта, на примере трех языков – русского, английского и немецкого – описываются основные принципы построения УЗ в части морфологии и синтаксиса, рассматриваются наиболее характерные варианты синтаксических конструкций.

Во второй части статьи на примере безличного предикатива описывается, как можно вести корпусные исследования с использованием УЗ. Рассматривается методика автоматического выявления безличных предикативов, их частотное распределение в корпусе русского языка УЗ, семантическая категоризация наиболее часто используемых предикативов.

Ключевые слова: универсальные зависимости, кросс-лингвистическое исследование, грамматический разбор, лингвистическая типология, предикатив, корпусное исследование.

Введение

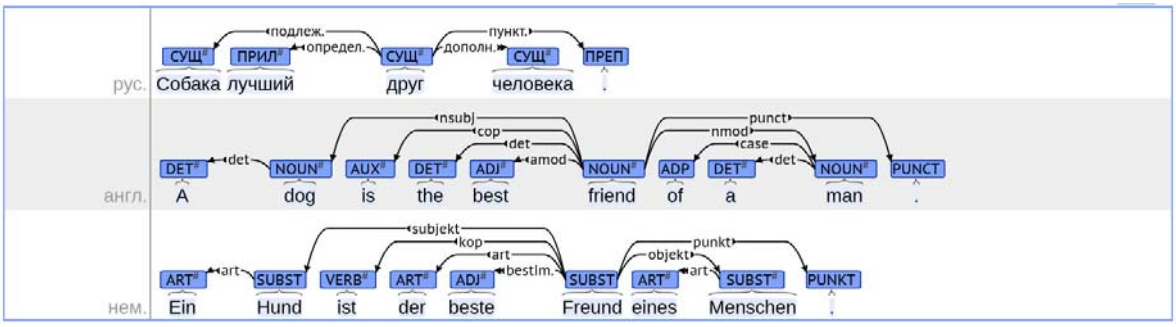
С развитием автоматизированных средств разбора и анализа текстов все большее значение приобретают кросс-лингвистические исследования, позволяющие обнаруживать сходные языковые явления в контексте нескольких языков, выполнять перевод с одного языка на другой, анализировать семантическую структуру многоязычных текстов, проводить сравнительные кросс-лингвистические исследования, формировать лингвистическую типологию на базе нескольких языков.

Однако многоязычным исследованиям в области грамматического разбора текста долгое время препятствовал тот факт, что схемы разбора для различных языков значительно различались. Это сильно затрудняло выполнение сравнительных оценок и кросс-лингвистического анализа экспериментов. Суть проблемы иллюстрирует пример 1, который показывает три параллельных предложения на русском, английском и немецком языках, проаннотированных по правилам русской грамматики [РГ, 1980], стенфордских типизированных зависимостей [Marneffe et al., 2006] и грамматического словаря Дудена [Duden, 1995] соответственно.

Люкина Е. В. Использование универсальных зависимостей при грамматическом разборе многоязычного текста (на примере безличного предикатива) // Вестн. Новосиб. гос. ун-та. Серия: Лингвистика и межкультурная коммуникация. 2018. Т. 16, № 2. С. 19–33.

ISSN 1818-7935
Вестник НГУ. Серия: Лингвистика и межкультурная коммуникация. 2018. Том 16, № 2

Пример 1



Все три предложения имеют сходную грамматическую структуру, но совершенно разные обозначения грамматических признаков и синтаксических связей. Как результат, автоматизированный грамматический разбор таких предложений одним и тем же парсером невозможен, и требуется создание отдельного парсера под каждый анализируемый язык, что существенно затрудняет проведение кросс-лингвистического исследования.

Грамматический разбор предложения, являющийся основой практически любого лингвистического исследования, характеризуется сопоставлением с лексическими единицами предложения морфологических и синтаксических признаков, определяющих роль лексической единицы в предложении, ее зависимость от других членов предложения и пр. Каждый язык имеет свой устоявшийся набор таких признаков.

Современные достижения в области глубокого обучения позволили создать парсеры, основанные на нейросетях, которые обучены на больших объемах текстов целевых языков и являются многоязычными [Straka et al., 2016]. При автоматизированном грамматическом разборе многоязыковых текстов такими парсерами возникает проблема: как сопоставить грамматические языковые признаки разных языков друг с другом и унифицировать общее результирующее представление разбора.

Таким образом, отсутствие унификации языковых признаков создает сложности при решении следующих задач:

- сравнение эмпирических результатов исследований по нескольким языкам;
- выполнение кросс-языкового переноса языковых конструкций;
- оценка кросс-языкового обучения;
- создание и поддержка многоязычных систем;
- проведение сравнительных кросс-языковых исследований;
- проверка лингвистической типологии;
- проведение исследования в направлении построения универсального кросс-языкового парсера текста.

В данной статье рассматривается инициатива «универсальных зависимостей» (universal dependencies, далее УЗ) [Nivre et al., 2016], позволяющая задать общее кросс-языковое представление о грамматических признаках там, где это возможно, и сохранить специфичную для определенного языка конкретику там, где унификация невозможна. Основная цель проекта – создание непротиворечивого и унифицированного по всему множеству поддерживаемых языков набора языковых признаков, а также формирование по каждому поддерживаемому языку размеченного этими признаками корпуса текстов (treebank).

Основными целями проекта являются:

- определение кросс-языковой непротиворечивой схемы грамматического разбора;
- поддержка многоязычных лингвистических исследований;
- дополнение (не замена) специфичных для языка схем;
- аннотации разбора конкретного языка выбираются из универсального набора категорий (язык не обязан поддерживать их все) и являются подмножеством этого набора.

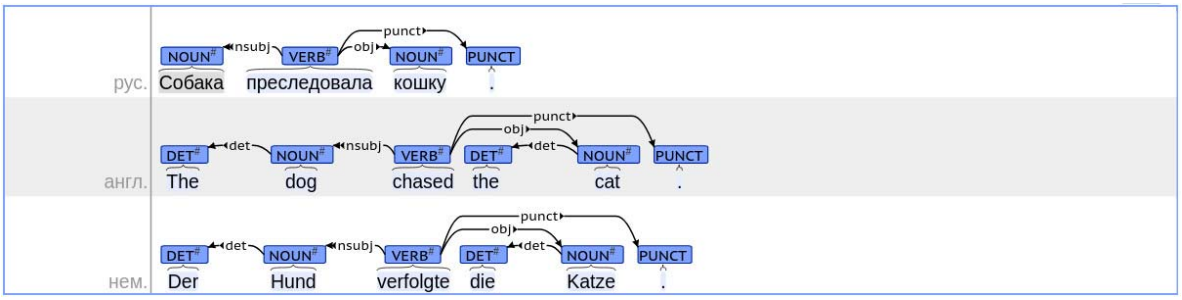
Как было показано в статье [Nivre et al., 2015], проект УЗ может быть рассмотрен как современный подход к универсальной грамматике. Средневековая идея об универсальной грамматике не единожды отвергалась лингвистами во многом по разумным причинам. Одна-

ко она может иметь смысл, если мы будем рассматривать ее как абстракцию: на определенном уровне языки имеют много общего, так почему бы не попытаться найти это общее? В УЗ рабочая гипотеза состоит в том, что языки имеют общие части речи (существительное, глагол, прилагательное и др.), общие грамматические признаки (подлежащее, сказуемое и т. п.). Некоторые языки могут не иметь всех этих признаков, некоторые могут иметь признаки, которые не являются универсальными. Этот подход показал себя как очень успешный, и на текущий момент в УЗ поддерживается более 60 языков.

До УЗ было несколько попыток создать унифицированные кросс-лингвистические схемы разбора (annotation schemes), а УЗ – это объединение нескольких из них. Основой УЗ является объединение подходов универсальных стенфордских зависимостей [Marneffe et al., 2006], универсальной схемы зависимостей Google [McDonald et al., 2013], универсальных тегов частей речи Google [Petrov et al., 2012] и Interset interlingua для микросинтаксических наборов тегов [Zeman, 2008], используемых в корпусе HamleDT [Zeman et al., 2012]. Проект УЗ, таким образом, базируется на де факто стандартах в этой области и призван заменить их все на единую унифицированную схему.

Рассмотрим пример одного и того же предложения на трех языках, разобранный в системе признаков «универсальных зависимостей» (далее УЗ-аннотации). Пример разобран парсером UDPipe, построенным на базе УЗ [Straka et al., 2016].

Пример 2



Как видно из примера, для аналогичных языковых конструкций использованы идентичные аннотации:

- *преследовала* / *chased* / *verfolgte* – сказуемое, глагол (root VERB)
- *собака* / *dog* / *Hund* – подлежащее, существительное (nsubj NOUN)
- *кошку* / *cat* / *Katze* – дополнение, существительное (obj NOUN)
- *The* / *Der* / *Die* – определитель, артикль (art DET)
- знак «точка» – символ пунктуации (PUNCT)

Основные принципы УЗ

Синтаксические аннотации в УЗ базируются на понятии зависимости, которое широко используется в современной обработке естественного языка (Natural Language Processing, NLP) и для аннотирования корпусов текстов, и для разбора текста. Подход УЗ также основан на гипотезе лексикализма, постулирующей, что слова являются базовыми единицами грамматического разбора. Слова имеют морфологические признаки и вступают в синтаксические отношения, которые и призваны отразить аннотации УЗ. На текущий момент подход УЗ не поддерживает слова с пробелами, такие многословные выражения в УЗ выражаются как несколько аннотированных однословных выражений, связанных специальным отношением.

УЗ-аннотации делятся на три основных набора:

- теги частей речи;
- морфологические признаки;
- синтаксические зависимости.

Морфология

Морфологическая спецификация слова в УЗ состоит из трех частей: лемма слова, тег части речи и морфологические признаки, которые определяют лексические и грамматические свойства формы слова. Лемма слова представляет его лексему и, как правило, задается в определенной канонической форме (например, для прилагательного русского языка это им. п., ед. ч., муж. р.).

УЗ имеет фиксированный список тегов частей речи – 17. Язык может не поддерживать все теги, но не может расширить этот список. Вместо этого для уточнения классификации слова используются морфологические признаки.

Теги морфологии задают дополнительные лексические и грамматические признаки слов, которые не отражаются тегами частей речи. Теги морфологии делятся на две группы: теги лексических признаков и теги флективных признаков. Теги флективных признаков задают грамматические признаки слов и, в свою очередь, делятся на именные и глагольные. Деление на группы морфологических признаков является в определенной степени условным, поскольку определенные лексические признаки могут присутствовать у различных частей речи, а определенные флективные признаки в некоторых языках могут являться лексической категорией (например, род является лексическим признаком существительного, но флективным признаком прилагательного или глагола).

Состав тегов морфологии не является закрытым и может пополняться новыми признаками. Конкретный язык также может использовать свои специфичные морфологические признаки.

Лексические признаки считаются свойствами леммы и используются для уточняющей классификации слов. Именные флективные признаки, большей частью, отражают свойства формы слова (как правило, имени), нежели свойства леммы. Глагольные флективные признаки отражают свойства формы глагола.

Пример 3

англ.:	*chased* – Lemma: chase, PartOfSpeech: VERB, Tense=Past (лемма: chase, часть речи: глагол, время: прошедшее)
рус.:	*преследовала* – Lemma: преследовать, PartOfSpeech: VERB, Aspect=Imperf, Gender=Fem, Mood=Ind, Number=Sing, Tense=Past, Voice=Act (лемма: преследовать, часть речи: глагол, вид: несоверш., род: женский, наклонение: изъявит., число: единств., время: прошедшее, залог: активный)
нем.:	*verfolgte* – Lemma: verfolgen, PartOfSpeech: VERB, Number=Sing, Tense=Past, Person: 3 (лемма: преследовать, часть речи: глагол, число: единственное, время: прошедшее, лицо: третье)

Синтаксис

Синтаксис в УЗ определяется с использованием аннотаций синтаксических зависимостей между словами. УЗ определяет большое количество синтаксических зависимостей (более 40 вариантов), используемых для кросс-языкового синтаксического разбора. Не все из определенных зависимостей используются в каждом языке, кроме того каждый язык может определить собственный дополнительный набор зависимостей, если для каких-то его конструкций не найдено соответствующих универсальных зависимостей.

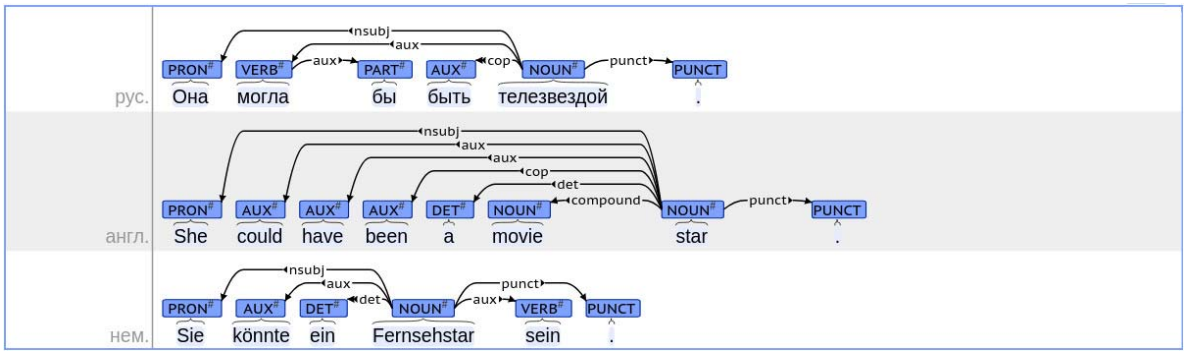
В результате грамматического разбора предложения УЗ совместимым парсером формируется дерево разбора, узлами которого будут слова исходного предложения, а ребрами – синтаксические аннотации, определяющие типизированным образом ту или иную синтаксическую связь между словами. Корнем дерева разбора (маркированного как «root»), как правило,

является сказуемое (глагол или любой другой вид сказуемого). Остальные слова участвуют в формировании промежуточных и листьевых узлов дерева разбора.

В УЗ приняты следующие основные принципы формирования синтаксических аннотаций.

- Главенство знаменательных слов: зависимости определяются преимущественно между знаменательными словами. Служебные слова, связываются со знаменательными словами, которые они специфицируют, отдельными связями. Это обусловлено тем, что повышается вероятность нахождения аналогичных конструкций в разных языках, так как служебные слова в одном языке часто соответствуют флективным признакам в другом или вообще не используются, т. е. являются сильно зависимыми от конкретного языка.

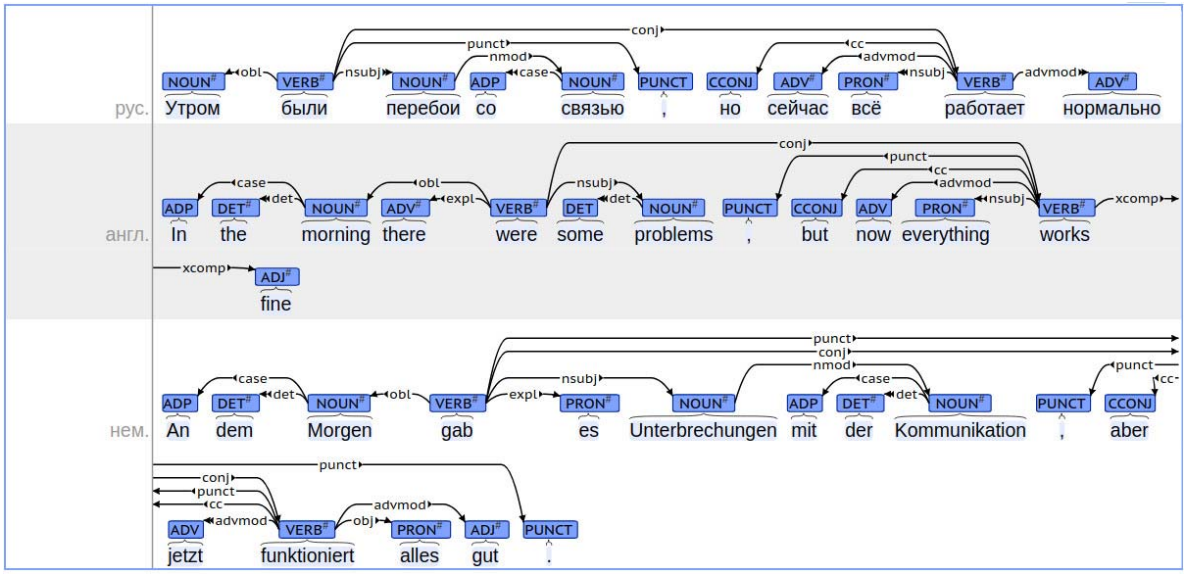
Пример 4



В данном примере (рус.) корневой узел «телезвездой» связан со знаменательным словом «она». Глагол «могла» связан с «телезвездой» связью «aux». Частица «бы» связана с «могла» связью «aux», которая выражает сослагательное наклонение. Глагол «быть» связан с «телезвездой» связью «cop» (copula). Знак пунктуации, обозначающий окончание предложения (запятая, точка, знак вопроса, восклицательный знак и др.), связывается с корнем предложения или словосочетания, которому он принадлежит. Знаки пунктуации, связанные со словами предложения (кавычки, скобки и т. п.), связываются со следующим / предыдущим словом.

- Принцип центральности: в сочинительной связи все знаменательные слова и знаки пунктуации связываются с первым знаменательным словом связи, то же самое применимо к именным словосочетаниям, глагольным словосочетаниям, многословным выражениям, составным именам и пр.

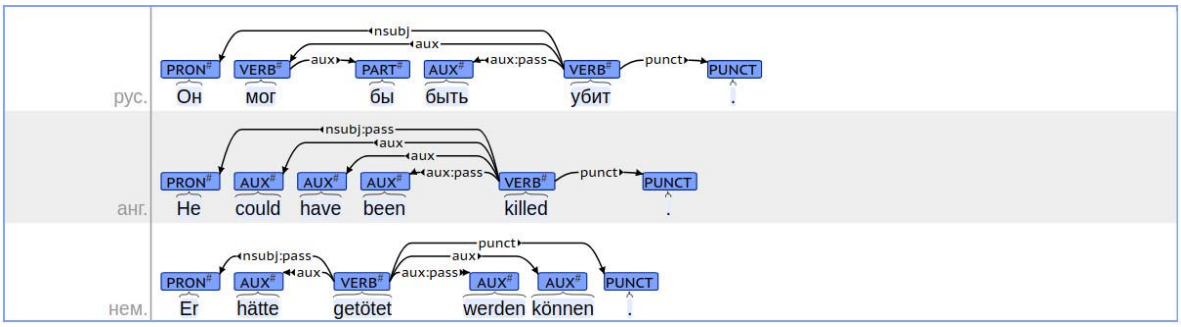
Пример 5



В данном примере (англ.) два сказуемых «were» и «works», формирующие простые предложения, связаны связью «conj» (сочинительная связь) и формируют каждый корневые узлы простых предложений, частью которых они являются.

- Главенство знаменательных слов означает, что служебные слова не образуют самостоятельных связей и в случае употребления нескольких служебных слов, относящихся к одному знаменательному, связываются со знаменательным словом как дочерние узлы. Это также касается вспомогательных глаголов и глаголов-связок.

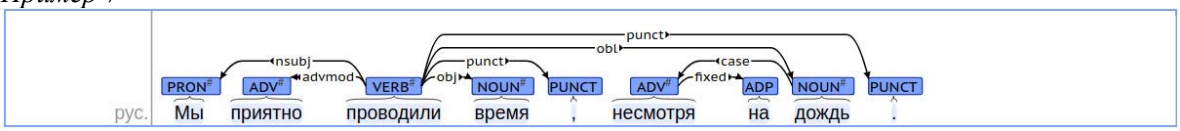
Пример 6



В данном примере (рус.) сказуемое «мог» связано с частицей «бы» связью «aux», определяющей сослагательное наклонение глагола.

- Служебные конструкции, состоящие из нескольких слов (например, составные предлоги, союзы и т. п.), связываются специальной зависимостью «fixed». Первое слово такой конструкции берется за корень дерева, остальные слова связываются с первым. Вся составная конструкция как единое целое связывается с элементом дерева предложения, от которого она зависит.

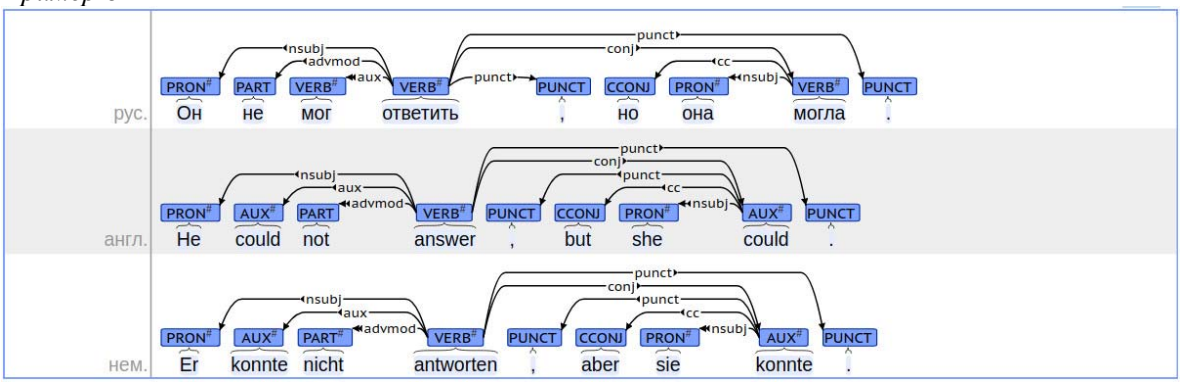
Пример 7



В данном примере (рус.) слово «несмотря» связано с предлогом «на» связью «fixed» и является корневым узлом для составного предлога «несмотря на».

- В случае если знаменательное слово опущено, служебное слово выступает в качестве главного по отношению ко всем словам, зависимым от исходного, т. е. эти слова сформируют зависимости со служебным словом вместо опущенного знаменательного.

Пример 8



В данном примере (англ.) вспомогательный глагол «could» заменил собой опущенное слово «answer».

Варианты синтаксических конструкций УЗ

Для иллюстрации того, как различные синтаксические явления описываются в УЗ, рассмотрим некоторые из синтаксических конструкций, поддерживаемых УЗ (описание полного списка таких конструкций выходит за рамки данной статьи).

Словосочетание

В именном словосочетании основным словом является существительное, с которым связываются остальные слова.

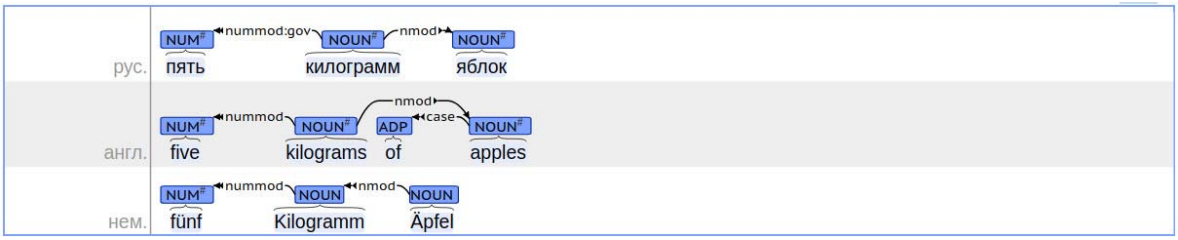
Пример 9

В данном примере (рус.) существительное «рубашка» связано с прилагательным «белая» связью «amod» (определение).

Конструкции с числительными

В УЗ числительное связывается с существительным связью «nummod» (числительное).

Пример 10

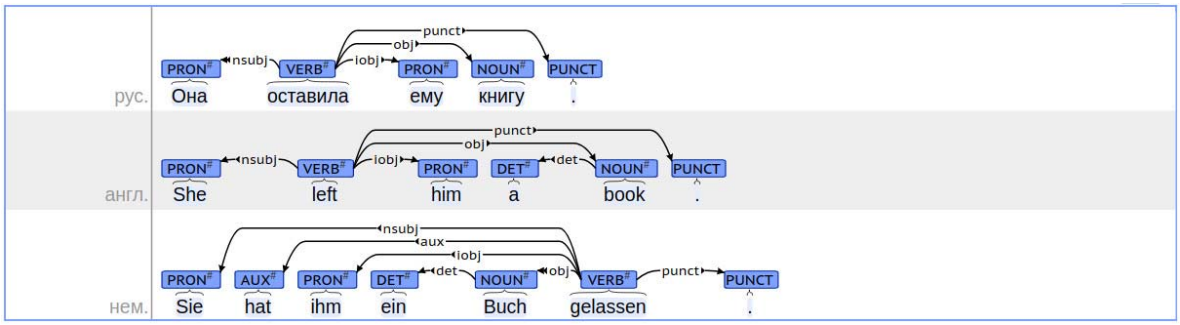


В данном примере (рус.) основное слово – существительное «килограмм» связано с числительным «пять» связью «nummod:gov» (метка «gov» показывает, что числительное управляет существительным).

Простое предложение

В простом предложении корневым узлом разбора является сказуемое (как правило, глагол), с которым связано подлежащее (существительное или местоимение), а также другие слова (например, дополнение, выраженное существительным).

Пример 11

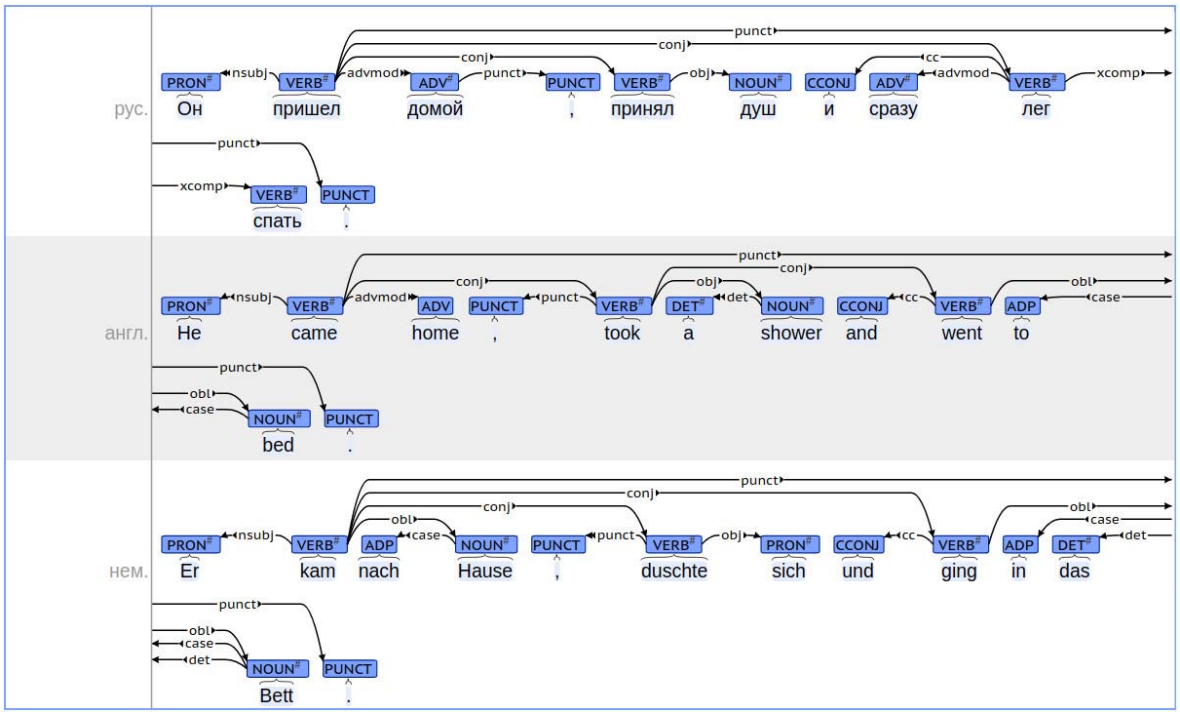


В данном примере (рус.) корневым узлом является сказуемое «оставила», с которым связано подлежащее «она» (связь «nsubj» – подлежащее) и дополнение «ему» (связь «iobj» – дополнение) и «книгу» (связь «obj» – дополнение).

Сочинительная связь

Сочинительная связь в УЗ формируется между корневыми элементами структур, составляющих сочинительную связь (отдельных слов, словосочетаний или простых предложений). Для отдельного слова корневым элементом будет само это слово. Корневым элементом словосочетания будет главное слово (например, существительное для именного словосочетания и глагол для глагольного). В случае простого предложения корневым элементом будет сказуемое. Если в сочинительную связь входят более двух элементов, каждый следующий элемент будет связан с первым связью «conj» (сочинительная связь).

Пример 12

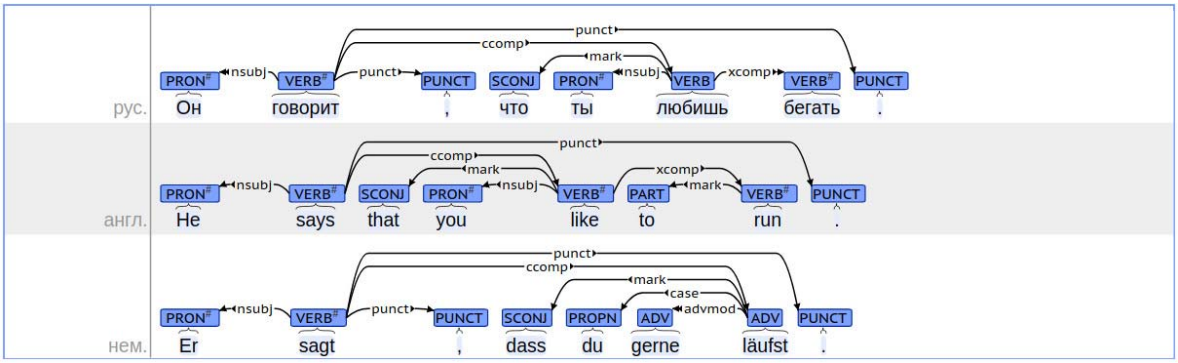


В данном примере (рус.) сочинительная связь формируется между сказуемыми «пришел», «принял», «лег» и выражена связью «conj».

Подчинительная связь

Подчинительная связь между двумя предложениями формируется как связь между сказуемыми этих предложений, выраженная УЗ-аннотациями «csubj» (клаузальное подлежащее), «ccomp» (клаузальное дополнение), «xcomp» (открытое клаузальное дополнение), «advcl» (клаузальное обстоятельство) и «acl» (клаузальное дополнение).

Пример 13

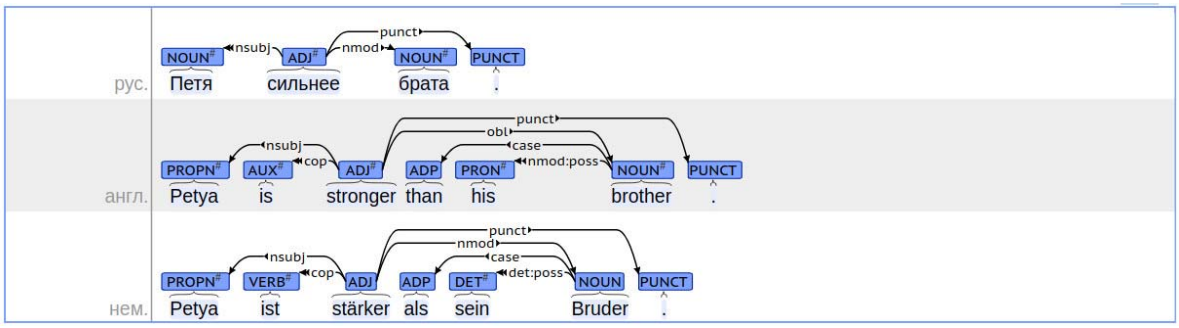


В данном примере (англ.) подчинительная связь формируется между сказуемыми «says» и «like» и выражена связью «ccomp» (клаузальное дополнение).

Сравнение

Конструкции сравнения в различных языках могут достаточно сильно различаться. В простейшем случае сравнение формируется в виде корневого узла, заданного сказуемым, выраженным прилагательным в сравнительной или превосходной степени и связанным с ним подлежащим и дополнением, являющимися аргументами сравнения.

Пример 14



В данном примере (рус.) корневым элементом сравнения является прилагательное «сильнее» (сказуемое), а аргументами сравнения – существительные «Петя» (подлежащее) и «брата» (дополнение).

Безличный предикатив

Предикативом (или безличным предикативом) называется определенный разряд слов, которые в сочетании с нулевой связкой или глаголом «быть» выступают в качестве сказуемого в безличном предложении: «мне интересно», «мне было плохо». Ранее наряду с термином

«предикатив» в указанном значении использовался также термин «категория состояния», сейчас этот термин считается устаревшим и не используется. Термин «предикатив» также имеет более широкое значение, когда предикативом называют именную часть именного сказуемого: «Иван всегда готов помочь другу». В этой статье мы ограничимся рассмотрением только безличных предикативов или слов категории состояния.

На текущий момент среди лингвистов нет единого мнения, считать ли предикатив отдельной частью речи. Одни исследователи считают, что не требуется выделять их как отдельную часть речи, и выделяют среди них два разряда: предикативные существительные (*жаль*, *лень*, *след*) и предикативные наречия (*нельзя*, *холодно*, *жарко*) [РГ, 1980]. Другие выделяют их как отдельную самостоятельную часть речи – слова категории состояния, справедливо считая, что их синтаксические особенности требуют такого выделения [Щерба, 2004] [1]. Третьи считают ее специализированной синтаксической частью речи [Шелякин, 2001].

Морфологически предикативы подразделяются на несколько категорий.

- Слова, мотивированные существительными (*лень*, *грех*, *время*, *пора*, *охота*, *неохота*, *резон*, *досуг*, *недосуг*, *резон*, *толк*). Это ограниченная категория предикативов, которая в настоящее время не пополняется новыми словами.
- Слова, мотивированные краткими прилагательными среднего рода: *грустно*, *легко*, *тяжело*, *трудно*, *странно*, *приятно*, *удивительно*, *дурно*, *плохо*, *невозможно.*
- Слова, мотивированные краткими формами причастий среднего рода: *натоптано*, *накурено*, *закрыто.*
- Предикативные наречия [Шведова, 1980]: *должно*, *можно*, *надо*, *нельзя*, *нужно*, устар. *надобно.*

Большинство предикативов (кроме предикативов, мотивированных существительными) образуют формы сравнительной степени: *плохо – хуже*, *трудно – труднее*.

Хотя семантика предикативов, как и большинства слов русского языка, является многозначной, можно выделить следующие их основные категории [Летучий, 2017]:

- обозначающие оценку;
- отражающие физическое состояние или ощущение;
- отражающие эмоциональное состояние или чувства и восприятие;
- предикативы с ментальной семантикой;
- предикативы с модальной семантикой;
- выражающие пространственные отношения;
- выражающие временные отношения;
- обозначающие свойство места и расположение в нем объектов и не имеющие субъекта оценки;
- отражающие состояние природы;
- отражающие состояние окружающей среды.

Методика исследования

Для изучения вопросов употребления предикатива мы использовали корпус русского языка УЗ. Данный корпус построен на базе корпуса русского языка SynTagRus [Дяченко и др., 2015]. Корпус содержит около 1 млн словоупотреблений и более 60 тыс. предложений из текстов различных жанров (фантастика, научно-популярная литература, газеты, журналы, новости и др.). Корпус поставляется в формате CoNLL-U.

Исследование проведено в три этапа. На первом этапе мы реализовали методику выявления предикативов в произвольном русскоязычном тексте. Для этого на языке программирования Java был создан парсер текстов, позволяющий составить частотный словарь использования предикативов. На вход парсера подается текст, размеченный в формате CoNLL-U. В качестве такого текста мы использовали упомянутый ранее корпус русского языка УЗ. Интересно, что, используя любой УЗ совместимый парсер, например UDPipe [Straka et al., 2016], который выдает результаты в УЗ совместимом CoNLL-U формате, можно без изменения пар-

[1] Первое издание вышло в 1974 г. (Москва, АН СССР).

сера вести исследование на произвольном неразмеченном корпусе текстов. Это подтверждает полезность применения УЗ в исследованиях такого рода.

Парсер работает, опираясь только на синтаксис, и не анализирует семантику предложения, поэтому с достаточной точностью может выявлять только предикативы, мотивированные краткими прилагательными и наречиями, оканчивающимися на -о, -ее, каковых большинство в русском языке. Предикативы, мотивированные существительными, таким образом выявить нельзя. Для обхода этого ограничения нами из различных источников был сформирован начальный список предикативов (около 100 единиц), включающих и предикативы, мотивированные существительными. Этот список также поступает на вход парсера.

УЗ не выделяет предикатив в отдельную часть речи, поэтому для выявления предикатива в произвольном русском предложении парсер использует следующий алгоритм.

1. Среди всех предложений корпуса текстов выявляются все безличные предложения. Предложения, имеющие подлежащее, ссылающееся на корневой элемент предложения, который в УЗ всегда является сказуемым, не рассматриваются. Для этого проверяется, что каждое слово предложения не содержит ссылку типа **nsubj** на слово типа **root**.
2. Среди всех словоформ определенного безличного предложения выделяется словоформа типа **root**, обозначающая сказуемое, остальные словоформы не рассматриваются (главенство знаменательных слов в УЗ предполагает, что глагол связки, если есть, будет зависимым от предикатива, а не наоборот). Сказуемые подчиненных предложений, если есть, также не рассматриваются.
3. Из полученного в п. 2 списка словоформ выделяются все словоформы, являющиеся краткими прилагательными единственного числа среднего рода, оканчивающиеся на -о. В УЗ часть речи: adj, признаки: degree=pos|gender=neut|number=sing|variant=short.
4. Из полученного в п. 2 списка словоформ выделяются все словоформы, являющиеся прилагательными в сравнительной степени, оканчивающиеся на -ее. В УЗ часть речи: adj, признаки: degree=cmp.
5. Из полученного в п. 2 списка словоформ выделяются все словоформы, являющиеся наречиями, оканчивающиеся на -о. В УЗ часть речи: adv, признаки: degree=pos.
6. Из полученного в п. 2 списка словоформ выделяются все словоформы, являющиеся наречиями в сравнительной степени, оканчивающиеся на -ее. В УЗ часть речи: adv, признаки: degree=cmp.
7. Полученные в п. 3–6 списки словоформ объединяются в единый результирующий список предикативов.

На втором этапе по всем найденным предикативам была построена частотная характеристика их использования в предложениях, после чего все слова были отсортированы в порядке убывания частоты использования. На третьем этапе для первых 40 наиболее часто употребляемых слов мы выполнили анализ их распределения по семантическим категориям.

Результаты исследования

В итоге на корпусе русского языка УЗ были получены следующие результаты:

- общее количество предложений: 61 889;
- общее количество словоупотреблений: 905 092;
- предварительно отобранных вручную предикативов: 101, из них реально использованных в данном корпусе: 62;
- количество найденных автоматически уникальных предикативов: 186;
- общее количество использованных в корпусе уникальных предикативов: 248;
- количество предложений с отобранными вручную предикативами: 2 505;
- количество предложений с найденными автоматически предикативами (исключая учтенные по отобранным вручную): 723;
- общее количество предложений с предикативами: 3 228.

Как видно из полученных результатов, количество предложений с предикативами составляет около 5 % от общего числа предложений. Это показывает, что предикатив является значимым языковым явлением.

Распределение уникальных предикативов по частям речи (в скобках показана УЗ часть речи), использованных в исследуемом корпусе:

- прилагательное (adj): 169;
- наречие (adv): 72;
- существительное (noun): 4;
- глагол (verb): 2.

Наиболее продуктивными из частей речи, с точки зрения образования предикативов, являются прилагательные и наречия. Причем прилагательные более чем в два раза опережают по словообразованию наречия. Существительные и глаголы вносят несущественный вклад в словообразование предикативов.

Распределение словоупотреблений предикативов по частям речи в количестве предложений на данную форму предикатива в исследуемом корпусе (в скобках показаны УЗ часть речи и УЗ признаки соответствующей словоформы):

- краткое прилагательное ед. ч., ср. р. (adj – degree=pos | gender=neut | number=sing | variant=short): 1 550;
- наречение (adv – degree=pos): 1618;
- наречие в сравнительной степени (adv – degree=cmp): 18;
- существительное (noun): 40;
- причастие (verb – aspect=perf | gender=neut | number=sing | tense=past | variant=short | verbform=part | voice=pass): 2.

Распределение словоупотреблений показывает, что как предикативы, мотивированные краткими прилагательными, так и предикативы, мотивированные простыми наречиями, используются в языке примерно в равной степени и значительно опережают по частоте использования остальные части речи – сравнительные наречия, существительные и причастия.

Распределение словоупотреблений предикативов по уникальным словам в количестве предложений на данное слово предикатива в исследуемом корпусе (первые 40 наиболее часто употребляемых):

можно: 839	надо: 376	нужно: 221	нельзя: 160
необходимо: 105	невозможно: 100	трудно: 87	важно: 83
понятно: 77	известно: 74	интересно: 46	достаточно: 45
хорошо: 40	легко: 33	ясно: 32	возможно: 28
сложно: 28	видно: 25	очевидно: 25	принято: 25
время: 18	жаль: 18	естественно: 16	пора: 16
неудивительно: 15	просто: 15	странно: 15	целесообразно: 15
приятно: 14	удивительно: 14	непросто: 13	нетрудно: 13
любопытно: 11	правильно: 11	непонятно: 10	тяжело: 10
желательно: 9	жалко: 8	нелегко: 8	

Распределение словоупотреблений предикативов по семантическим категориям в количестве предложений на данную семантическую категорию / данное слово предикатива в исследуемом корпусе (среди выбранных общеупотребительных):

- модальные предикативы (1 946): *можно* (839), *надо* (376), *нужно* (221), *нельзя* (160), *необходимо* (105), *важно* (83), *возможно* (28), *невозможно* (100), *принято* (25), *желательно* (9);
- предикативы ментального состояния (395): *понятно* (77), *известно* (74), *интересно* (46), *достаточно* (45), *ясно* (32), *любопытно* (11), *непонятно* (10), *очевидно* (25), *неудивительно* (15), *естественно* (16), *странно* (15), *целесообразно* (15), *удивительно* (14);
- предикативы физического состояния (207): *трудно* (87), *легко* (33), *сложно* (28), *тяжело* (10), *просто* (15), *непросто* (13), *нетрудно* (13), *нелегко* (8);
- предикативы оценки (111): *хорошо* (40), *время* (18), *жаль* (18), *пора* (16), *правильно* (11), *жалко* (8);
- предикативы эмоционального состояния (39): *видно* (25), *приятно* (14);
- в сумме (по первым 40): 2 698;
- на оставшиеся 208 предикативов приходится 520 словоупотреблений.

Таким образом, модальные предикативы составляют более половины найденных словоупотреблений, что показывает очень высокую частотность их употребления. По-видимому, модальность является одним из основных вариантов использования предикатива в речи.

На втором месте идут предикативы ментального состояния, в большинстве случаев отражающие отношение говорящего к предмету разговора и являющиеся, как кажется, вторым основным вариантом использования предикатива в речи.

На предикативы физического состояния, оценки и эмоционального состояния в сумме приходится около 10 % словоупотреблений.

На оставшиеся некатегоризированными 208 предикативов приходится 520 словоупотреблений (около 15 %).

Заключение

Мы рассмотрели проект универсальных зависимостей, позволяющий унифицировать грамматические признаки между различными языками. Унификация выполняется как по частям речи, так и по линии морфологического разбора, а также в части синтаксиса. Это позволяет упростить кросс-лингвистические исследования, унифицирует межъязыковую лингвистическую типологию, дает основу для создания автоматизированных многоязычных систем и универсального кросс-языкового парсера текстов.

В исследовательской части статьи на примере предикатива и корпуса русского языка УЗ показано, как вести корпусные исследования с использованием УЗ. Универсальность описания грамматических признаков УЗ позволяет вести исследования не только на предварительно размеченном языковом корпусе, но и на произвольном корпусе текстов (используя специализированный УЗ совместимый парсер) без изменения методики и алгоритмов исследования.

Создана методика выявления предикативов в произвольном тексте, определено частотное распределение предикативов по различным критериям: частям речи, уникальным словоформам, семантическим категориям.

По результатам исследования выяснено, что предложения с предикативами занимают около 5 % от общего объема предложений. Наиболее продуктивными из частей речи, с точки зрения образования предикативов, являются прилагательные и наречия. Предикативы, мотивированные краткими прилагательными и простыми наречиями, являются наиболее часто употребляемыми в речи. Модальные предикативы и предикативы ментального состояния составляют основную часть словоупотреблений.

Список литературы

Дяченко П. В., *Иомдин Л. Л.*, *Лазурский А. В.*, *Митюшин Л. Г.*, *Подлесская О. Ю.*, *Сизов В. Г.*, *Фролова Т. И.*, *Цинман Л. Л.* Современное состояние глубоко аннотированного корпуса текстов русского языка (СинТагРус) // Национальный корпус русского языка: 10 лет проекту. Труды Института русского языка им. В. В. Виноградова. М., 2015. Вып. 6. С. 272–299.

Летучий А. Б. Предикатив. Материалы для проекта корпусного описания русской грамматики (http://rusgram.ru). На правах рукописи. М., 2017.

РГ – Русская грамматика / Под ред. Н. Ю. Шведовой. М.: Наука, 1980. Т. 1: Фонетика. Фонология. Ударение. Интонация. Словообразование. Морфология; Т. 2: Синтаксис.

Шелякин М. А. Функциональная грамматика русского языка. М.: Русский язык, 2001.

Щерба Л. В. Языковая система и речевая деятельность. М.: Едиториал УРСС, 2004.

Duden. Mannheim; Leipzig; Wien; Zurich, 1995. Bd. 4: Grammatik der deutschen Gegenwartssprache.

Marneffe M. et al. Universal Stanford Dependencies: A cross-linguistic typology, 2006.

McDonald R., *Nivre J.*, *Quirmbach-Brundage Y.*, *Goldberg Y.*, *Das D.*, *Ganchev K.*, *Hall K.*, *Petrov S.*, *Zhang H.*, *Täckström O.*, *Bedini C.*, *Bertomeu Castelló N.*, *Lee J.* Universal dependency annotation for multilingual parsing // ACL, 2013.

Nivre J., *Marneffe M.*, *Ginter F.*, *Goldberg Y.*, *Hajic J.*, *Manning C.*, *McDonald R.*, *Petrov S.*, *Pyysalo S.*, *Silveira N.*, *Tsarfaty R.*, *Zeman D.* Universal Dependencies v1: A Multilingual Treebank Collection // Proceedings of the Tenth International Conference on Language Resources and Evaluation. European Language Resources Association (ELRA). Paris, France, 2016.

Nivre J. Towards a Universal Grammar for Natural Language Processing // CICLing 2015: Proceedings of Computational Linguistics and Intelligent Text Processing. Cairo, Egypt, 2015. Vol. 9041. P. 3–16.

Petrov S., *Das D.*, *McDonald R.* A universal part-of-speech tagset // LREC, 2012.

Straka M., *Hajič J.*, *Straková J.* UDPipe: Trainable Pipeline for Processing CoNLL-U Files Performing Tokenization, Morphological Analysis, POS Tagging and Parsing // Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016). Portorož, Slovenia, 2016.

Zeman D., *Mareček D.*, *Popel M.*, *Ramasamy L.*, *Štěpánek J.*, *Žabokrtský Z.*, *Hajič J.* HamleDT: To parse or not to parse? // LREC, 2012. P. 2735–2741.

Zeman D. Reusable tagset conversion using tagset drivers // LREC, 2008. P. 213–218.

Материал поступил в редколлегию 16.01.2018

Elena V. Lyukina

National Research University Higher School of Economics
20 Myasnitskaya Str., Moscow, 101000, Russian Federation

eliouki@mail.ru

USE OF UNIVERSAL DEPENDENCIES FOR GRAMMATICAL ANALYSIS OF THE MULTILINGUAL TEXT (ON THE EXAMPLE OF PREDICATIVE)

The paper combines the initiative of universal dependencies (UD), with the aim to promote developing cross-linguistically consistent annotation scheme of grammatical analysis. The purpose of this initiative is to simplify cross-language research methods, develop a unified interlanguage linguistic typology, and thus lay the foundations for the use of automated multilingual systems and a universal cross-language text parser.

The paper points out the main problems of grammatical analysis pertinent to a multilingual text, advantages of using common language features, and purposes of the universal dependencies project. The major principles of universal dependencies are discussed based on the example of morphology and syntax features of Russian, English and German.

In the second part we illustrate how to conduct a corpus research on predicatives using UD. The article proposes a technique of automatic identification of predicatives, examines their frequency distribution in the Russian UD corpus, and the semantic categorization of the most often used predicatives.

Keywords: universal dependencies, cross-linguistic analysis, grammatical analysis, linguistic typology, predicative, corpus-based research.

References

Dyachenko P. V., Iomdin L. L., Lazurskij A. V., Mitjushin L. G., Podlesskaja O. Yu., Sizov V. G., Frolova T. I., Cinman L. L. Sovremennoe sostoyanie gluboko annotirovannogo korpusa tekstov russkogo yazyka (SinTagRus) [Modern state of deeply annotated corpus of Russian language]. *Nacional'nyj korpus russkogo yazyka: 10 let proektu* [*National corpus of Russian language: 10 years of project*]. Trudy Instituta russkogo yazyka im. V. V. Vinogradova. Moscow, 2015, iss. 6, p. 272–299. (in Russ.)

DUDEN. Bd. 4: Grammatik der deutschen Gegenwartssprache. Mannheim Leipzig Wien Zurich, 1995.

Letuchij A. B. Predikativ. Materialy dlya proekta korpusnogo opisaniya russkoj grammatiki (http://rusgram.ru) [Predicative. Materials for project of corpus description of Russian grammar]. Na pravah rukopisi. Moscow, 2017. (in Russ.)

Marneffe M. et al. Universal Stanford Dependencies: A cross-linguistic typology, 2006.

McDonald R., Nivre J., Quirmbach-Brundage Y., Goldberg Y., Das D., Ganchev K., Hall K., Petrov S., Zhang H., Täckström O., Bedini C., Bertomeu Castelló N., Lee J. Universal dependency annotation for multilingual parsing. In ACL, 2013.

Nivre J., Marneffe M., Ginter F., Goldberg Y., Hajic J., Manning C., McDonald R., Petrov S., Pyysalo S., Silveira N., Tsarfaty R., Zeman D. Universal Dependencies v1: A Multilingual Treebank Collection. In Proceedings of the Tenth International Conference on Language Resources and Evaluation. European Language Resources Association (ELRA). Paris, France, 2016.

Nivre J. Towards a Universal Grammar for Natural Language Processing. In CICLing 2015: Proceedings of Computational Linguistics and Intelligent Text Processing, vol. 9041 of LNCS, pages 3–16. Cairo, Egypt, 2015.

Petrov S., Das D., McDonald R. A universal part-of-speech tagset. In LREC, 2012.

Shelyakin M. A. Funkcional'naja grammatika russkogo jazyka. [Functional grammar of Russian language]. Moscow, Russkij yazyk, 2001. (in Russ.)

Shcherba L. V. Yazykovaya sistema i rechevaya deyatel'nost' [Language system and speech activity]. Moscow, Editorial URSS, 2004. (in Russ.)

Shvedova N. Yu. (ed.). Russkaya grammatika [Russian grammar]. Moscow, Nauka [Science], 1980. Vol. I. Phonetics. Phonology. Accent. Intonation. Word formation. Morphology; Vol. II. Syntax (in Russ.)

Straka M., Hajič J., Straková J. UDPipe: Trainable Pipeline for Processing CoNLL-U Files Performing Tokenization, Morphological Analysis, POS Tagging and Parsing. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016). Portorož, Slovenia, 2016.

Zeman D., Mareček D., Popel M., Ramasamy L., Štěpánek J., Žabokrtský Z., Hajič J. HamleDT: To parse or not to parse? In LREC, 2012, p. 2735–2741.

Zeman D. Reusable tagset conversion using tagset drivers. In LREC, 2008, p. 213–218.

For citation:

Lyukina Elena V. Use of Universal Dependencies for Grammatical Analysis of the Multilingual Text (On the Example of Predicative). *Vestnik NSU. Series: Linguistics and Intercultural Communication*, 2018, vol. 16, no. 2, p. 19–33. (in Russ.)

DOI 10.25205/1818-7935-2018-16-2-19-33