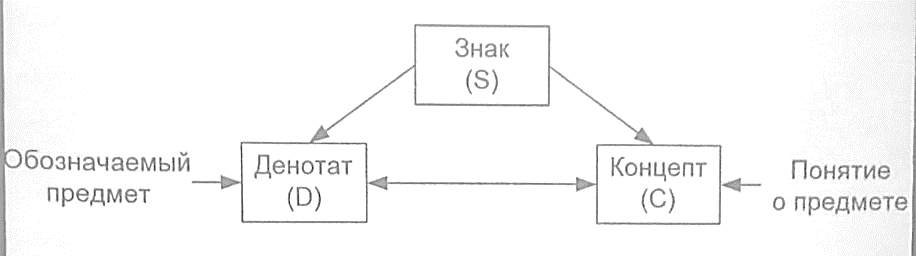
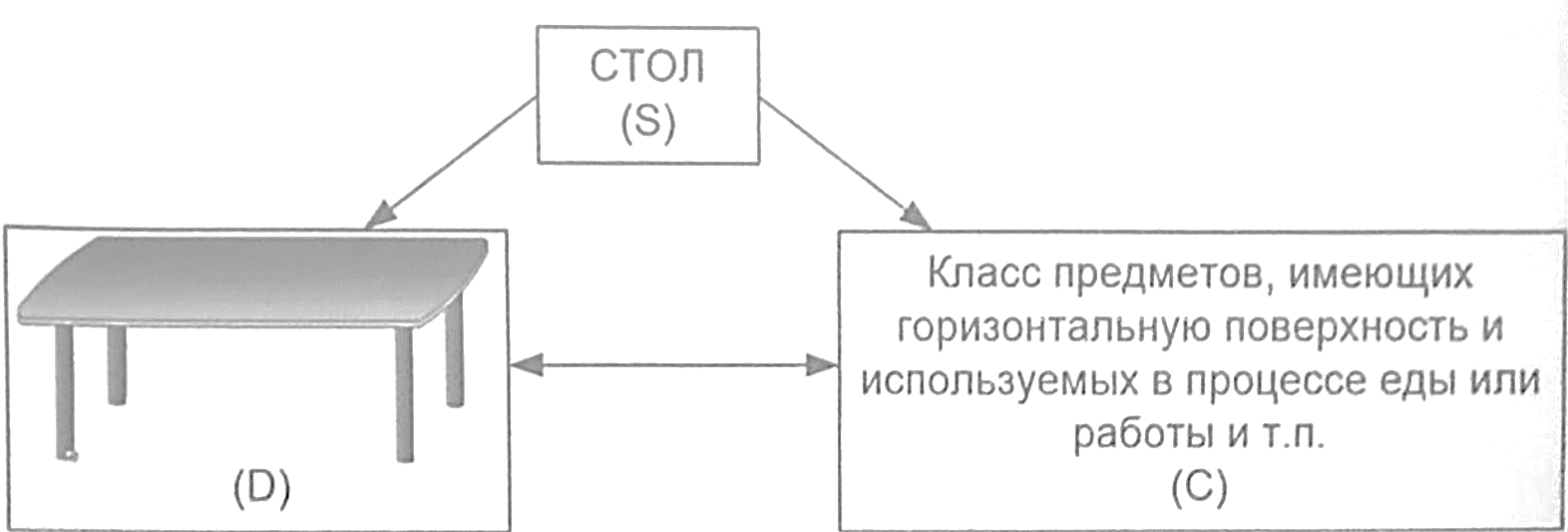
Юрий Борисович Крапивин

**Естественные Языковые Искусственные Интеллектуальные Системы**

**Центральные научные проблемы:**   
- проблема моделирования процесса понимания смысла текстов (переход от текста к формализованному представлению его смысла);  
- проблема синтеза текста и речи (переход от формального представления смысле к текстам на естественном языке).   
Автоматическая обработка текста (АОТ) и обработка естественного языка (ЕЯ) опирается на результаты общей лингвистики: Фонология, Морфология, Синтаксиса, семантика и прагматика, Лексикография.   
АОТ и ЕЯ связаны с научными направлениями, как: Математика, Computer Science, Психология, Социология.  
А это, в свою очередь, означает междисциплинарный характер проводимых исследований, связанных с решением следующих прикладных задач:  
1) Автоматизация инженерии знаний: автоматический анализ текста и выделении из него знаний.  
Основные классические типы знаний:   
- Объекты – конкретные элементы доменов предметной области.   
Именные группы: “позитрон”, “ударопрочность корпуса часов Луч”, “расход топлива”.  
- Факты – отношение между объектами предметной области, характеризующие данные объекты, их свойства и связи между ними.  
Субъект, акция, объект: “солнце испускает поток заряженных частиц”  
- Правила – отношения между фактами, выражающие закономерности предметной области.   
2) Информационный поиск (ИП) – автоматизация поиска в неструктурированных текстовых документах.  
Задачи ИП:   
- Классический поиск по закону Мура, индексирование документов.  
- Поиск документов похожих на исходный.  
- Тематический поиск по тональности и стилистической окраске.  
- Кластеризация и классификация документов.  
Классическое решение IR:  
- представление запроса и документа как векторов слов  
- вычисление степени подобия между векторами.   
Степень подобия векторов – косинус угла между векторами.  
3) Машинный перевод – автоматический перевод из одного языка другой.  
Подходы:   
- статистический – реализуется на базе параллельных корпусов текста.   
- лингвистический – модель перевода строится на основе языковых правил.  
- нейросетевой – реализуется на базе различных архитектур нейронных сетей.  
4) Автоматическое распознавание языка текстового документа.  
Особенности нейросетевого подхода:  
- долгий этап обучения, для которого нужны большие мощности.  
- быстрый лёгкий процесс практической работы.  
5) Автоматическая классификация текстового документа – автоматическая категоризация, кластеризация, генерация таксономий (тематических категорий). Применяются методы основанные на ML.  
6) Автоматическое реферирование – создание краткого содержания.   
Подходы:   
- Извлечение предложений с помощью оценочной функции и статистических показателей текста, например, частота встречаемости слов.  
- Извлечение содержания. Генерация реферата с созданием нового текста обобщающего первичный.   
7) Выделение по именованным сущностям – задача распознавания, извлечения и классификации по заданным категориям, например, отнесение документов по отношению к определенным людям и т.д.  
8) Анализ тональности текста ­– автоматический анализ мнений или эмоциональной окраски текста.  
9) Задача автоматического синтеза текста.   
Применения:   
- вопросно-ответные  
- диалоговые системы  
- системы машинного перевода  
- информационный поиск для целей поисковой оптимизации  
В основе всех этих задач лежит автоматический анализ текста:  
1) Лексический анализ предполагает:   
- сегментация текста  
- определение начальной формы слова и её парадигмы  
2) Лексико-грамматический анализ ­– автоопределение лексико-грамматического класса каждого слова.  
3) Синтаксический анализ – для каждого предложения строится синтаксическое дерево.  
4) Семантический анализ – распознавание концептов и семантических отношений.   
Лингвистический процессор – специальный информационно программный модуль, выполняющий лингвистическую обработку текстов.   
ЛП включает:   
1) Лингвистическую базу знаний   
2) набор программных алгоритмов обработки текста  
Структура ЛБЗ:  
1) Классификаторы лексико-грамматических свойств, синтаксических и семантических свойств ЕЯ  
2) Базовый словарь – все возможные словоформы  
3) Базовый корпус текстов – подобранные тексты для каждого слова, для каждого слова в контексте один лексико-грамматический код  
3) Лингвистические правила анализа текста на различных уровнях глубины ЕЯ

**Корпус тестов:**  
Корпус тестов – большой, представленный электронном виде, унифицированный, структурированный, размеченный, филологически компетентный массив языковых данных предназначенный для решения конкретных лингвистических задач. В понятие также входит система управления корпусом текстов и поисковую систему.  
Цели создания:   
1) Достаточно большой спектр данных гарантирует типичность и полноту всего спектра языковых явлений.  
2) Данные разного типа находятся в их естественной контекстной форме, что создаёт возможность их всестороннего и объективного изучения  
3) Созданный корпус может переиспользоваться   
- Поиск по слову позволяет построить конкорданс – ссылки на связанные слова.  
- Корпусы могут использоваться для получения разнообразных справок и статистических данных о языковых и речевых единицах.  
- Корпусы источники в подготовке исторических или современных словарей.  
- Корпус – это уменьшенная модель языка или подъязыка.  
- Задача создателей корпуса собрать не столько большой набор текстов, сколько пропорциональный.  
- Репрезентативность – пропорциональное содержание разных форм и периодов одного стиля. Репрезентативность – важнейшее свойство корпуса.  
Для решения различных лингвистических задач потребовалось дополнительная информация в тексте, так появились размеченные корпуса.   
tag, или метки, бывают:  
- внешние  
- экстралингвистические (автор, место издания, жанр)  
- структурные (глава абзац, предложение, словоформа)  
- собственно лингвистических, описывающих характеристики элементов текста   
Разметка:   
1) Морфологическая: часть речи, признаки части речи  
2) Синтаксическая: синтаксические связи между предложениями  
3) Семантическая: тематические категории   
4) Анафорическая: фиксирует связи, например, местоимения  
5) Просодическая: ударение, интонация  
Семантическая разметка учитывает:  
1) значения слов  
2) разрешение омонимии и синонимии   
3) категоризация слов (разряды)  
4) тематические классы  
5) признаки каузативности   
6) Оценка   
7) Деривационные характеристики   
8) и т. д.  
Три группы помет:   
1) разряд   
2) лексико-семантические характеристики  
3) деривационные характеристики  
Технология процесса построения:   
1) Определение перечня источников  
2) Оцифровка текстов  
3) Предобработка текста (корректирование, подготовка описания)  
4) Конвертирование и графематический анализ  
5) Разметка текста  
6) Корректировка результатов автоматической разметки  
7) Конвертирование данных в тип информационно-поисковой системы  
8) Обеспечение доступа к корпусу  
Для морфологического и синтаксического анализа существуют программные средства – тэггеры и парсеры.  
Построение разметки:   
1) выделить из БТК “минимальное допустимое” по объёму подмножество Т1, текстов для аннотирования  
2) из корпуса текстов Т выделить его определенную часть Т2 и проанализировать её вручную, положить Т3 = Т1 – Т2   
3) выделить из T3 его определенную часть Т4 и на основе полученных данных из Т2 получить возможные варианты разметки для каждого предложения из Т4.  
4) для каждого варианта аннотирования из текста Т4 вручную убрать лишние варианты  
5) положить Т3 = Т3 – Т4\*, где Т4\* однозначное проаннотированное множество.  
6) повторять шаги 3-5 до тех пор, пока множество текстов на шаге 5 Т3 не окажется пустым.  
Точность решение задачи определяется как:  
Pi = mi / ni \* 100  
а полнота:  
Ri = mi / ki \* 100  
n – получено объектов  
m – правильных объектов  
k – исходных объектов  
Лингвистическая и экстралингвистическая разметка базируются на достаточно широко распространённых и принятых принципах описания текстов и языковых единиц.

**Основные составляющие семиотики:  
Синтактика** – изучает структурные аспекты, сочетаний знаков данной системы, правила их образования и преобразования безотносительно к их значениям и функциям.  
**Семантика** рассматривает знаковые системы как средства выражения смысла, занимается установлением зависимости между структурой знакосочетаний и их выразительными возможностями, изучает интерпретации знаков, знакосочетаний и совокупностей знакосочетаний, образующих осмысленные тексты.  
**Прагматика** изучает отношение воспринимающего знаковую систему (интерпретатора или адресата) к самой знаковой системе. Прагматика изучает восприятие осмысленных выражений знаковой системы в соответствии с разрешающими способностями воспринимающего. Можно сказать, что прагматика определяет, каким образом воспринимающая система на основании информации, не заложенной в воспринимаемом выражении, выбирает из множества интерпретаций наиболее подходящую для данного случая.   
Знак – центральное понятие семиотики.   
- знаками обозначаются предметы, объекты реального мира, процессы, явления.  
- Для обозначения и исследования отношений между знаками рассматриваются **знаковые системы**.   
**Программа с точки зрения семиотики:**- способ написания программы рассматривается как знак, на том или ином языке, есть её **синтаксис**  
- **семантика** программы есть ее смысловое содержание: какую именно задачу призвана решать данная программа и какое содержание заключено в каждой ее строке.   
- **прагматика** реализуется процессом ее выполнения на компьютере, что обеспечивается конструкцией компьютера.  
**Знак** - это материальный, чувственно воспринимаемый предмет, выступающий в познании и общении людей в качестве представителя некоторого предмета или предметов и используемый для приобретения, хранения, преобразования и передачи сообщений (информации, знаний) или компонентов сообщений какого-либо рода. Предметом обозначения могут быть вещи, свойства, объекты, отношения объектов в реальном мире, действия и процессы, психологические явления и абстрактные объекты науки, символы других знаковых систем.  
**Денотат** - это отражение предмета внешнего мира. Выбор денотата определяется конкретной знаковой ситуацией, т.е. зависит от контекста.  
**Концепт** - это понятие о предмете, т.е. информация, которую знак несет о возможных денотатах, об их положение в системе реалий, об их месте в универсуме. Это не вся информация о денотате, а только та ее часть, которая отражается данным знаком.

Иногда в состав знака включается также **коннотат**, т.е. комплекс чувственно-оценочных, в том числе и эстетических оттенков значения.  
Будем называть знаковой системой четверку (S, U, K, /), где  
S - множество знаков,  
U - универсум, т. е. множество денотатов,  
К - система знаний, т. е. множество понятий, в которых описываются концепты и их взаимоотношения,  
I - интерпретации, соотносящие знаку его денотат или концепт.  
EЯ как знаковая система  
ЛЕКСЕМА (от греч. lexis - слово, выражение), слово, рассматриваемое как единица словарного состава языка в совокупности всех его конкретных грамматических форм и выражающих их флексий (окончаний), а также всех возможных значений (смысловых вариантов); абстрактная двусторонняя единица словаря.  
В одну лексему:   
- объединяются разные парадигматические формы (словоформы) одного слова («словарь, словарём, словарю» и т. п.).   
- включаются разные смысловые варианты слова, зависящие от контекста, в котором оно употребляется («соль» в смысле названия вещества и в значении того, что придает остроту или интерес какому-либо высказыванию, мысли и т. п.).  
СЛОВОФОРМА термин, обозначающий конкретное слово в конкретной грамматической форме. Словоформа - это слово, рассматриваемое как представитель определенной лексемы и определенной грамматической формы, например, лесом (= ЛЕС: твор. п. ед. ч.), идет (= ИДТИ: изъяв. накл., наст. вр., 3 л., ед. ч.) и M.П.  
Понятие - это копия предмета реальной действительности в нашем мышлении. Понятие в слове всегда одно, а значений может быть несколько. Понятие «Зеленый» обладает значениями: Зеленый карандаш (цветовая характеристика), Зеленый плод (степень вызревания, можно сравнить: спелый плод); Зеленое лицо (характеристика нездоровья, степень усталости); Зеленый возраст (степень социальной зрелости).  
Разница между понятием и значением в том, что: понятие - это копия, точное обозначение; в значение всегда входит эмоционально-экспрессивная окраска (модальность).  
Например, Солнышко - уменьшительно-ласкательный оттенок. Бабка - пренебрежительный оттенок.  
У понятия оттенков быть не может. Морфемочка, фонемочка - не употребляется!  
I - интерпретация (множество концептов С) - это множество лексических значений слов языка.  
Используются два способа соотнесения знаку сущностей из универсума:   
- экстенсиональный заключается в явном указании этих сущностей в универсуме;  
- интенсиональный состоит в задании этих сущностей в терминах системы знаний с помощью некоей совокупности свойств, выделяющих эти сущности из универсума.  
- экстенсионалом знака называют класс всех его допустимых денотатов.  
- интенсионалом знака называют характеристику концепта, выраженную через общие свойства денотата этого знака и его отношение с другими денотатами.  
- понятия (человек, зарплата, должность, фамилия, имя, отчество, возраст, пол, год рождения и т. д.),  
- отношения (больше, меньше, быть мужчиной, быть женщиной и т. д.) и  
- функции (суммировать, перечислять и т. д.).  
Лингвистика - это наука, изучающая естественный язык, его строение и закономерности использования человеком в процессе общения.  
Специалисты в области лингвистики утверждают, что естественный язык представляет собой некую иерархическую организацию, в которой единицы каждого уровня складываются из единиц более низкого уровня.  
При этом переход от одного уровня к другому осуществляется в соответствии с определенными правилами. Конкретные разделы лингвистики изучают разные уровни этой иерархии и взаимодействия этих уровней между собой.  
- лексикология (и лексикография);  
- фонетика (и фонология);  
- морфология (и морфемика);  
- синтаксис;  
- семантика;  
- прагматика.  
Основной значимой единицей языка является слово. Совокупность всех слов языка составляет его лексику. Раздел науки о языке, изучающий словарный состав языка, называется лексикологией. В лексикологии изучаются самостоятельные слова с точки зрения их лексического значения, употребления и происхождения.  
Лексикография (Aefikóv, lexikon -- «словарь» и үрáфw, grapho - «пишу») -- раздел языкознания, занимающийся вопросами составления словарей и их изучения; наука, изучающая семантическую структуру слова, особенности слов, их толкование.   
- (филол. ) собирание слов какого-либо языка, приведение их в систему и издание в виде лексикона, словаря; совокупность печатных трудов словарного типа, словарная литература.  
- теория и практика составления словарей.  
Фонетика изучает звуковой состав речи.  
Морфология - это раздел лингвистики, который изучает слово как часть речи, а также морфемный состав слова. Морфология - это учение о морфемах языка и их сочетаемости. Части речи были выделены лингвистами с целью классификации лексического состава языка. Было замечено, что разные слова выполняют различные функции в процессе общения, а отдельные группы слов обладают некоторыми общими чертами. Такие группы слов были названы частями речи. В средней школе изучается 10 частей речи. Для более детального анализа некоторые лингвисты выделяют также дополнительные части речи.  
Грамматическое значение – это формальное значение, выступающее как добавочное к лексическому значению слова и выражающее различные отношения: к другим словам в словосочетании или предложении; к типу, совершающему действие или другим лицам; отношение сообщаемого факта к действительности и времени; отношение говорящего к сообщаемому.