КОМПЬЮТЕРНАЯ ЛИНГВИСТИКА И ПЕРЕВОД

для умных

Логико-математические основы компьютерной лингвистики

будьте умными)

Семейная история

- Математическая лингвистика
- Вычислительная лингвистика
- Компьютерная лингвистика
- Инженерия данных (с LLMs)

Языковедческие основы компьютерной лингвистики

история развития человека
в зеркале эволюции
его представлений о языке

будьте людьми)

Естественное и искусственное

Естественный:

- Природный, прирожденный (противопоставляется искусственному, как сделанному человеком)
- Нормальный, без отклонений (противопоставляется искусственному, как надуманному, ненастоящему, аномальному)

Язык - явление общественное, а не природное (натуральное, естественное)

Как посмотреть

И. А. Бодуэн де Куртенэ: Почти всегда литературный язык образуется искусственно

Л. А. Булаховский: Нет неискусственных литературных языков..

Н. Я. Марр: Натуральных языков не существует в мире, все языки искусственные, все созданы человечеством.

Языковая искусственность

Между естественным и искусственным
в жизни человека и человеческого общества
нет четкой грани. Все искусственное, делается
из природного материала, на основании законов,
которым подчинена вся природа.

Абревиат **LA** (из корней ling и art выражает
языковую искусственность.

Степени LA (ling+art)

- LA-1 - речевая стихия бесписьменных говоров
- LA-2 - языки нормализованные
- LA-3 - апостериорные языки
- LA-4 - априорные языки
- LA-5 - компьютерные языки + математические нотации

Степени LA (ling+art)

- LA-1 - речевая стихия бесписьменных говоров
- LA-2 - языки нормализованные
- LA-3 - апостериорные языки
- LA-4 - априорные языки
- LA-5 - компьютерные языки +
- **AL-1** - стихия внутренних языковых представлений LLM

Claude-3.5-Sonnet

- А что может LLM считать "своим родным языком"?
- Это очень интересный вопрос, который затрагивает наиболее фундаментальные аспекты природы и функционирования LLM.
- На самом деле, LLM не может "считать" какой-либо язык "своим родным" в том смысле, в котором это понимают люди. Вот почему....

В языковой стихии

словѣне — говорящие «словами» (по-нашему).

Племена
- *ѩзыкъи* -
(языки)

По ту сторону языка

Праславянское

*nĕmьcь

от *nĕmъ

«немой, неспособный говорить

(на понятном языке)

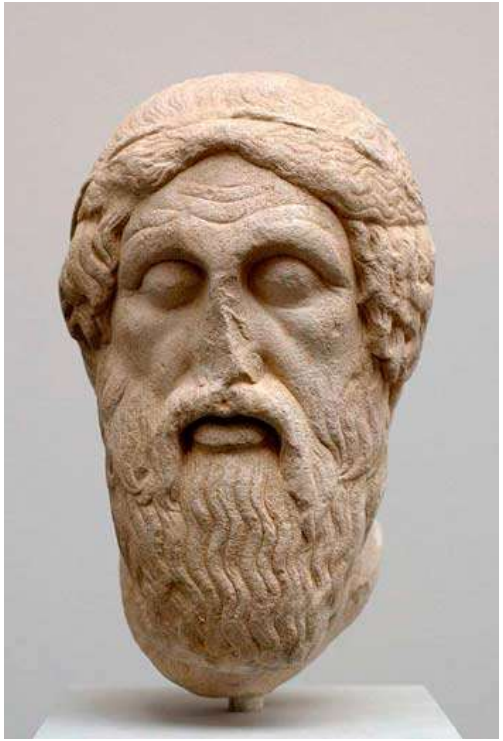
Истоки научной рефлексии

- **Па́нини** (पाणिनि, V век до Р.Х.) – предтеча современной структурной лингвистики, порождающей грамматики, семиотики и логики.

Исток европейской рефлексии

Гоме́р (VIII век до Р.Х.) — древнегреческий поэт, создатель Илиады (древнейшего европейского литературного произведения) и Одиссеи.

- 50% древнегреческих литературных папирусов — отрывки из Гомера.



Ключевой термин

ОТЧУЖДЕНИЕ:

- отделение от людей процесса и результатов их деятельности, при котором процесс и результаты деятельности становятся неподвластными человеку и даже господствуют над ним, делая чуждыми друг другу человека и создаваемый им мир.

От филолога к лингвисту и дальше…

- В чем отличие
  между филологом, лингвистом
  и компьютерным лингвистом?

Прощай, филология….

Филология (филос + логос) – содружество гуманитарных дисциплин изучающих **духовную (?)** культуру через языковой анализ письменных и устных текстов…

Филос утрачена….

- Древние греки: филос, агапэ, сторге, эрос
- Джон Алан Ли (*John Alan Lee):* эрос, агапэ, сторге, людус, маниа, *п*рагмос, ….

Логос утрачена….

Михаил Вячеславович Копотев

Михаил Вячеславович Копотев

… конференция по корпусной лингвистике проходила
в помещении бывшего анатомического театра.

Михаил Копотев

Первый же докладчик отметил символичность места: корпусной лингвист тоже работает с корпусом, препарируя его с помощью специальных инструментов. Традиция открытого для широкой публики доступа к корпусу исчезла из медицинской науки, но, возродилась в лингвистике ….

Платон

Аристокл (428-347 до Р.Х) - ученик Сократа, учитель Аристотеля.

- «Кратил» — диалог Платона о значении слов — могут ли имена служить познанию вещей.

Античные Грамматисты

Марк Теренций Варрон (116—27) - первый грамматист Рима, приспособил греческие схемы описания к латинскому языку.

Александрия Птолемеев (IIвек от Р.Х):

- **Аполлоний Дискол** («Синтаксис»),
- **Дионисий Фракийский** (Грамматика)

1000-летние образцы

- **Донат** (III—IV века)
- **Присциан** (VI век).

الكتاب

Аль-Китаб

Абу Бишра Амри ибн Усмана аль-Басри,

Сибавейхи (سيبويه)

http://sydney.edu.au/arts/research_projects/sibawiki/homepage/

Восхождение

- 13-14 века - школа модистов. Томас Эрфуртский.
- 16 век Пьер де ла Раме (Рамус), понятийный аппарат и терминология синтаксиса.
- 17 век - грамматика Пор-Рояля (Антуан Арно и Клод Лансло).
- 18 век Этьен Бонно де Кондильяк

Новое время

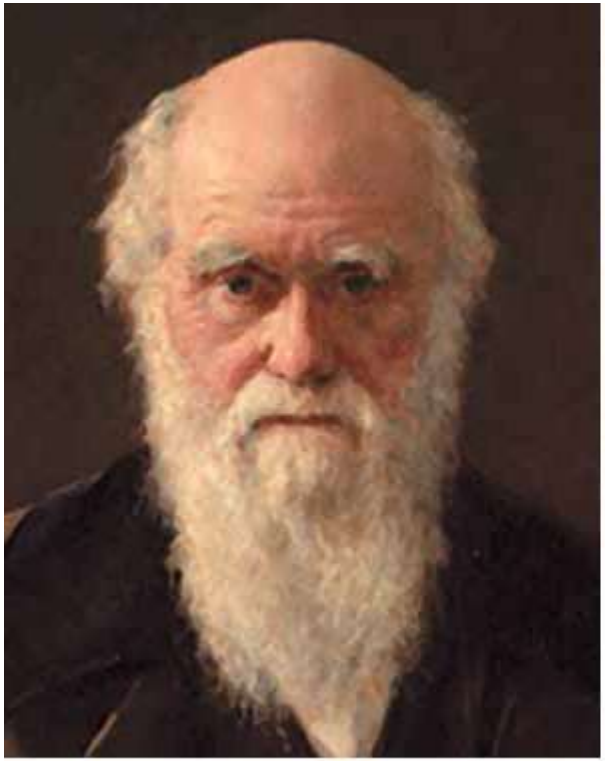
- В 1819 печатается первый том «Немецкой грамматики» **Якоба Гримма** (1785—1863).
- В 1820 году выходит «Рассуждение о славянском языке» **Александра Христофоровича Востокова** (1781-1864).
- В этих сочинениях впервые формировался сравнительно-исторический метод.

Вершина

Вильгельм фон Гумбольдт
(1767—1835)

… язык - живой дух,
в симбиозе
с духовным человечеством
и человеком.

Уоллес - Дарвину



Чарлз Дарвин

Альфред Рассел Уоллес

Зачем обезьяне мозг философа?

- Как же в таком случае мог один из органов получить развитие, столь превышающее потребности его обладателя?
- Естественный отбор наделил бы дикаря мозгом, едва превосходящим мозг обезьяны, тогда как на самом деле его мозг только чуть-чуть менее развит, чем у рядового члена наших научных сообществ.

Ветер перемен

Август Шлейхер (1821-1868) большой любитель ботаники, ученый и практик-садовод

Язык есть организм природы.

Басня «Овца и кони»

Алимова Валентина, «Сбор яблок», 1963

Роботизация села!

24.04.2016 Министры сельского хозяйства G7 обсудили его роботизацию.

- В Японии программа замены выходящих на пенсию фермеров роботами (20 типов).
- В помощь фермерам роботизированные рюкзаки Kubota с элементами экзоскелета

Молодая смена!

Средний возраст фермеров:

- в развитых странах 60 лет,
- в Японии 67,
- в США - 57.

За 20 лет число фермеров :

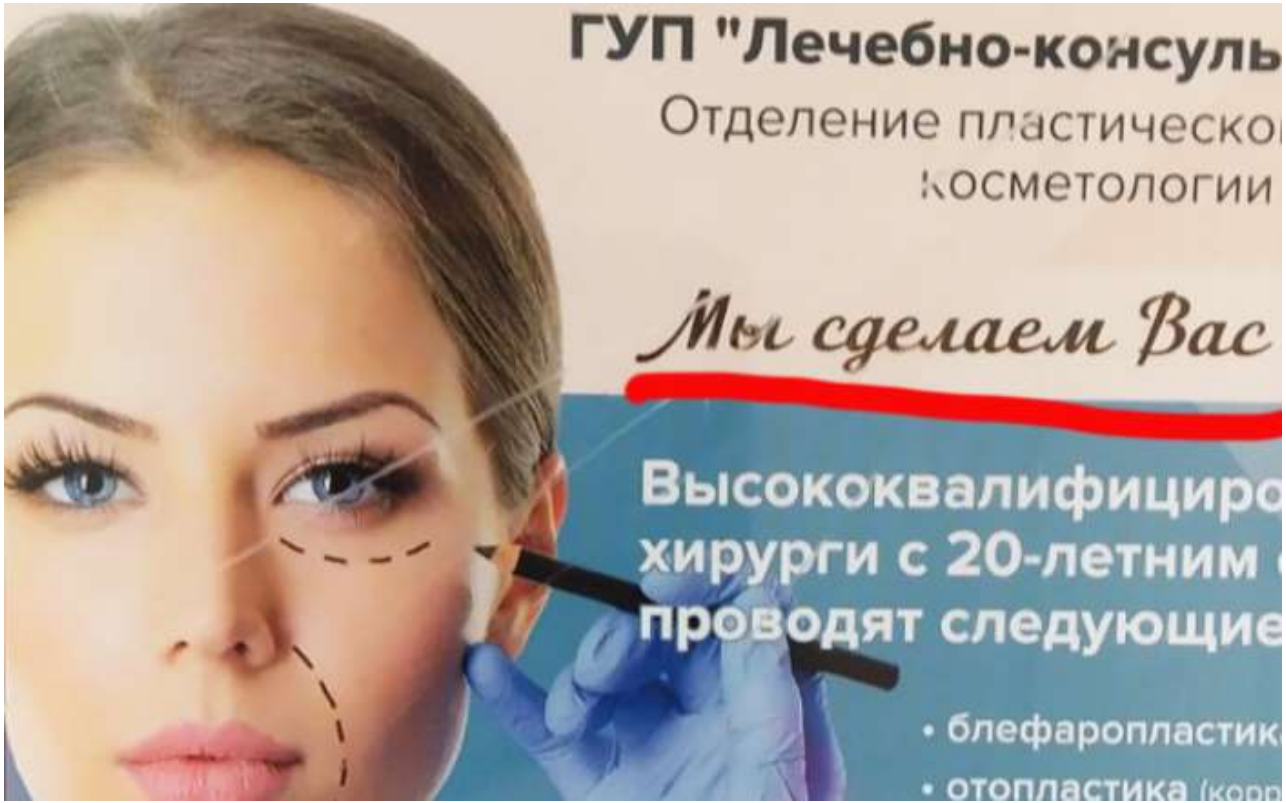
- старше 75 лет выросло на 30%,
- моложе 25 лет сократилось на 20%.
- необрабатываемые земли в Японии удвоились.

От организма к механизму

Фердинанд Монжин де Соссюр (1857-1913)
«Мемуар о первоначальной системе гласных в индоевропейских языках» (1879)

...в языке нет ничего, кроме различий

От организма к механизму

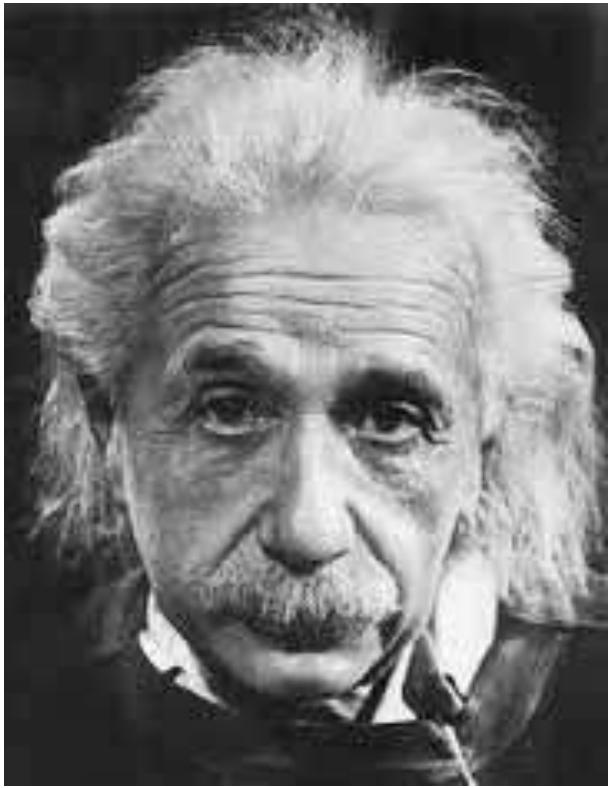


От организма к механизму

Просто как дважды два?

Невозможно решить проблему на том же уровне, на котором она возникла. Нужно стать выше этой проблемы, поднявшись на следующий уровень

или опустившись?



Нежить: желание простых решений

Julien Offray de La Mettri

1709-1751

- Человек-машина (1747),
- Животные - большее, чем машины (1748),
- Человек-растение (1748)

Генерация парней-цветочков

Собственно… я понимаю,
что парни сейчас
как цветочки, но….
Чтоб настолько буквально
понимать значение
этого выражения…
я в шоке

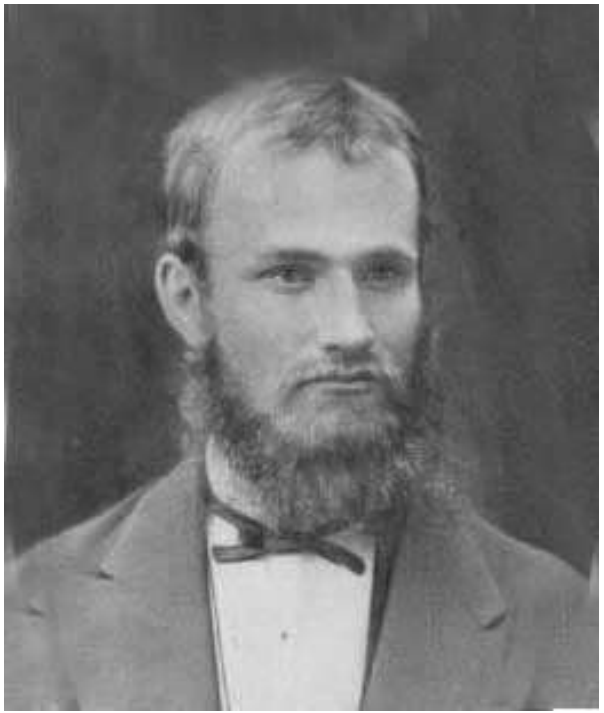
— Нарисуй парня 18 лет с черными волосами, на фоне города под дождем. На нем должно быть темная кофта, и он должен курить сигарету 23 сентября в 15:32

— Вот ваш рисунок:

D’où venons nous Que sommes nous Où allons nous

А языка то и нет вовсе?

Бодуэн де Куртенэ Иван Александрович (1845-1929)



...язык
как физическое
явление вообще не
существует

Что? Где? Когда?

Компьютерная лингвистика (что?) –

актуальная лингвистика

Вся актуальная лингвистика –

компьютерная

Осенью – всё осеннее!

ИТОЖИМ

- Главное в КЛ - инженерия (языков и текстов), а не их изучение
  Компьютерная лингвистика - не наука о естественном языке и не наука вообще, а инженерия - область технического творчества... творения искусственных миров...

ИТОЖИМ

- Компьютерная лингвистика – языковое моделирование реальности (eXtended reality, XR) посредством NLP (Natural Language Processing) и NLG (Natural Language Generation) и

Что? Где? Когда?

Компьютерная лингвистика (где?) –

в техносфере

Лингвистика техносферы = лингвистика глокальных сетевых сообществ

В компьютерных сетях – все компьютерное (только роли разные)

Семиотическое

- Текст - объединенная смысловой связью последовательность знаковых единиц любой формы коммуникаций (письмо, песня, танец, рисунок, обряд).
- Культура - предел текста (в эру Культуры)
- Техносфера - предел текста в современную эру –

Техника

Техника - инобытие (внешняя проекция) тела человека и тела человечества.

Техногенез (эволюция техники) - это процесс возникновения и совершенствования элементов техногенной реальности, превращение техники в самостоятельную силу, развивающуюся по собственным законам в направлении техносферы.

Техносфера

- Техноценоз - локальная совокупность технических систем в их отношениях между собойи с нетехническими факторами среды...
- Техносфера - системная целостность объектов, имеющих искусственное происхождение, все части которой связаны структурными взаимодействиями с обменом веществом, энергией и информацией.

Человек в техносфере

- Техносфера - искусственная среда обитания человека или новая (техногенная) реальность, в которой человек...
- Техника и позволяет получать впечатления, не добывая их собственным телом. Двигаться, не двигаясь.. видеть не видя...
- Текстоценоз - текст в текстовой (виртуальной) реальности....

Текстуальность

термин постструктурализма, констатирующий, что внетекстовой реальности не существует и любой индивид неизбежно находится внутри текста, мир есть бесконечный, безграничный текст (general text), в котором каждый текст всегда отсылает только к тексту.

Текст не есть событие, рождающееся внутри языка, но сам язык вписан в текст.

Что? Где? Когда?

Компьютерная лингвистика (когда?) –

актуальная лингвистика

Вся актуальная лингвистика –

компьютерная

Осенью – всё осеннее!

Новая Земля и новое Небо

- Языковая грамотность
- Языковая компетентность
- Языковая среда
- Языковая культура
- Языковая судьба

Barbara Hall Partee

- первая аспирантка **Хомского Ноама Абрамовича**
- Соратница **Ричарда Монтегю**,
- Формальный семантик,
- Профессор Массачусетского университета (США),
- приглашенный профессор Факультета лингвистики РГГУ)

Куда податься филологу?

- Кто финансирует ваше исследование? Спасибо.

Барбара Парти:

- В данный момент никто.
- Для этого, слава Богу, не нужны очень дорогие инструменты, чтобы исследовать семантику. Раньше в Штатах были гранты ....
- В Европе тоже есть такие научные фонды. Но сейчас много работ в рамках компьютерной лингвистики, там всегда есть деньги.

ВЫВОД №1

Развитие математики и логики
привело в XX в к:

- постановке вопросов о возможностях и ограничениях языков описания реальности…
- проблематике их формализации
- необходимости и возможности языковых вычислений…

ВЫВОД №2

Математика, в отличие от лингвистики, достигла огромных успехов и решила, что обойдется без нее.

МАТЕМАТИКИ - гроссмейстеры и мастера. Лингвисты - не знают теории игры - шахматной композиции, задач, этюдов.... этапов партии, взаимодействия фигур, только ходы.... Филологи-сказочники - даже фигур не знают....

Корпусная лингвистика

- область лингвистики,
- связанная с созданием и развитием корпусов текстов (Text corpus),
- их применением в качестве инструмента лингвистического исследования

Корпус (лингвистический)

- **репрезентативная** (собранная в соответствии с определёнными принципами, соответствующими задаче функционального представления языка),
- совокупность текстов (в электронной форме, письменных и устных),
- размеченных (снабженных аннотациями,),
- обеспеченных специализированной поисковой системой.

Плунгян Владимир Александрович

- доктор филологических наук
- член-корреспондент РАН,
- завсектором Института языкознания РАН,
- завсектором корпусной лингвистики и лингвистической поэтики Института русского языка РАН,
- профессор МГУ …..

Плунгян Владимир Александрович

Плунгян Владимир Александрович

- Запомните единственное: теперь для овладения языком человеку нужны не две, а три вещи: словарь, грамматика и корпус текстов данного языка.
- Потому что и словарь, и грамматика, в общем-то, бесполезны вне этого живого пространства, где язык, собственно, и функционирует.

Плунгян Владимир Александрович

- Более того, и словари и грамматики теперь нужны не традиционные, а нового поколения,
- то есть не просто словари и грамматики, а словари такого-то корпуса и грамматики такого-то корпуса, что сразу дает нам возможность их проверить.

К прочтению:

- Успенский Владимир Андреевич
  российский математик, логик и лингвист, д. ф-м.н. …

  Апология математике
  Труды по нематематике
  Математика - это гуманитарная наука.

Алгебры высказываний

Логика высказываний… (исчисление, 0 порядок)

Логика предикатов (кванторов *1 порядка…*)

Логики кванторов высших порядков…

Многозначные логики

Нечеткие логики… (Лотфи Заде)

……

Алгебра категорий…..

Альфред Тарский

1902 – 1983, автор термина «теория моделей»

Kmp-рекомендует:

- Истина и доказательство
- Семантическая концепция истины и основания семантики

Семантическая концепция истины

1. Главная проблема — удовлетворительное определение истины.
2. Объём термина «истинно».
3. Значение термина «истинно».
4. Критерий материальной адекватности искомого определения.
5. Истина как семантическое понятие.
6. Языки с точно заданной структурой.

Корпусная лингвистика

- область лингвистики,
- связанная с созданием и развитием корпусов текстов (Text corpus),
- их применением в качестве инструмента лингвистического исследования

Корпус (лингвистический)

- **репрезентативная** (собранная в соответствии с определёнными принципами, соответствующими задаче функционального представления языка),
- совокупность текстов (в электронной форме, письменных и устных),
- размеченных (снабженных аннотациями,),
- обеспеченных специализированной поисковой системой.

Плунгян Владимир Александрович

- Запомните единственное: теперь для овладения языком человеку нужны не две, а три вещи: словарь, грамматика и корпус текстов данного языка.
- Потому что и словарь, и грамматика, в общем-то, бесполезны вне этого живого пространства, где язык, собственно, и функционирует.

Плунгян Владимир Александрович

- Более того, и словари и грамматики теперь нужны не традиционные, а нового поколения,
- то есть не просто словари и грамматики, а словари такого-то корпуса и грамматики такого-то корпуса, что сразу дает нам возможность их проверить.

Альфред Тарский

- При обсуждении проблемы определения истины и вообще любых проблем из области семантики мы должны использовать два разных языка.
- Первый из них есть язык, который «о чём-то говорит» и который является предметом всего нашего обсуждения, ибо искомое определение истины как раз и применяется к предложениям этого языка.

Альфред Тарский

- Второй язык - тот, в котором мы «говорим о» первом языке и в терминах которого мы хотим, в частности, построить определение истины для первого языка.
- Первый язык мы будем называть «объектным языком», второй - «мета-языком».

Слова и дела Тарского

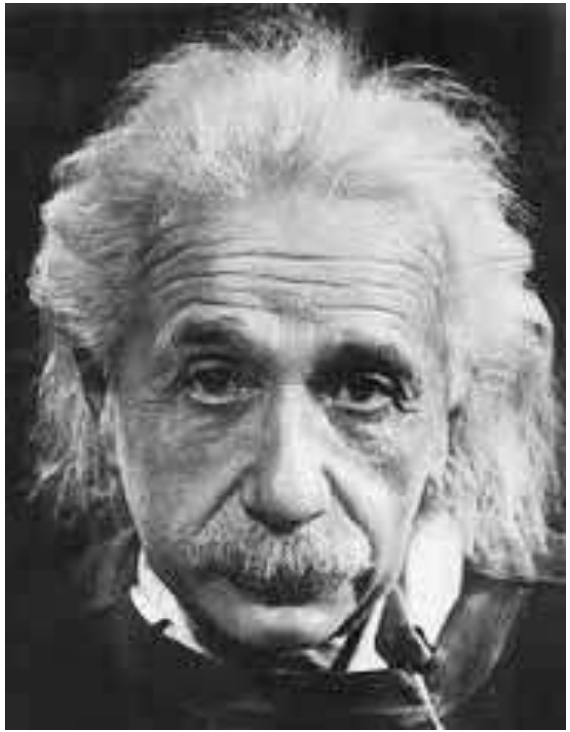
А. Тарский доказал неопределимость понятия истинности средствами предметного языка и предложил семантическое определение истины, как метаязыковой категории.

Простота как совершенство

Вы думаете, всё так просто?

Да, всё просто.

Но совсем не так.



Смешное смешение

Смешение терминов (слов) и высказываний (осмысленных утверждений) метаязыка и соответствующего языка-объекта порождает трудности в понимании и использовании языков человеческого общения и нередко приводит к серьёзным парадоксам.

Всё просто

- Для создания модели предметной области сначала строится модель наших представлений.
- Описание наших представлений содержится в метамодели
- Наши представления имеют очень непростую структуру, которая до сих пор не имеет формального описания.

Всё просто

- Наши представления имеют очень непростую структуру, которая до сих пор не имеет формального описания.
- Поэтому построение метамодели и мета-метамодели сильно затруднено.
- Попытку описать эту структуру предприняли греки, когда придумали основы логики.

Языковые парадоксы…

Курт Гёдель показал, что парадокс Лжеца возникает даже в таком элементарном языке, как арифметика.

Разделяй и властвуй!

Метаязык позволяет разрешить самореферентные парадоксы

«Лошадь — это существительное»?

В данном предложении «лошадь» — это термин языка-объекта, а «существительное» — метаязыковой термин.

«Слово «лошадь» — это существительное»!

Победа над лжецом

Эпименид Кносский (VII в. до Р.Х.):

- Один критянин сказал, что все критяне всегда лгут. Что он сказал — истину или ложь?

Смешение предметных терминов с метаязыковым понятием «истина», причём не только для оценки соответствующего предметного высказывания, но и по отношению ко всему этому утверждению в целом.

Лестница метаязыков

В исходном языке отсутствует «ложь» и «истина».

Оценка истинности утверждений об объектах, требует метаязыка — следующей ступеньки лестницы.

- L0 Утверждение 1
  - L1 Утверждение 1 истинно.
    - L2 Утверждение 2 истинно.
      - L3 Утверждение 3 истинно.

Метаязыковая относительность

Различение языков-объектов и соответствующих метаязыков является относительным:

- любой из метаязыков (в этом случае он является языком-объектом) может стать объектом описания метаязыка более высокого уровня (мета-метаязыка).

Метаязыковое богатство

Для описания языка-объекта в соответствующем метаязыке необходимо, чтобы:

- метаязык был логически более богатым, чем описываемой с его помощью язык-объект
- обладал бо́льшими выразительными возможностями

Метаязык должен содержать объектный язык как свою часть.

От чистого истока…

С середины 1930-х годов различение понятий «язык-объект» и «метаязык» стало активно использоваться в исследованиях проблем математической логики и оснований математики.

Позже его стали применять в лингвистике, семиотике, в философии и методологии науки.

И тут и там … метязык

- язык исследовании языков логико-математических исчислений
- язык описания языка-объекта
- метаданные, служащие для описания имеющихся данных.

Дискурс-анализ

- понимание языка на основе социально-конструкционистских подходов

Наши знания о мире и самих себе — не есть отражение реальности, но есть результат её исторически и культурно обусловленной категоризации

Язык - продукт дискурсов - способов понимания и репрезентации мира через саморепрезентацию

В шорах специализации…

- Дискурс (от фр. discours — речь, выступление) — речь, привязанная к говорящему (в отличие от récit (как речь безотносительно к говорящему).
- Дискурсивный анализ — изучение языка, используемого членами некоторого языкового сообщества на основе разговорной речи и письменных текстов…

Метаязык для ЯЛВ

- Заглавные латинские буквы A, B, C и др. (метабуквы) в определении формулы, принадлежат метаязыку, используемому для описания самого ЯЛВ.
- Содержащие метабуквы выражения — не пропозициональные формулы, а схемы формул.