ОБ ОДНОМ ПОДХОДЕ К АВТОМАТИЧЕСКОМУ ПОСТРОЕНИЮ ДЕРЕВЬЕВ ЗАВИСИМОСТЕЙ

Волкова И.А., Головин И. Г.
Каф. АЯ факультета ВМК МГУ

Синтаксические анализаторы и корпуса текстов

Синтаксические анализаторы:

Stanza, DeepPavlov, Наташа, UDPipe, SpaCy,…

Основаны на машинном обучении

Датасеты - готовые корпуса размеченных деревьев зависимостей

Для русского языка: СинТагРус, Тайга, Poetry, GSD, PUD,…

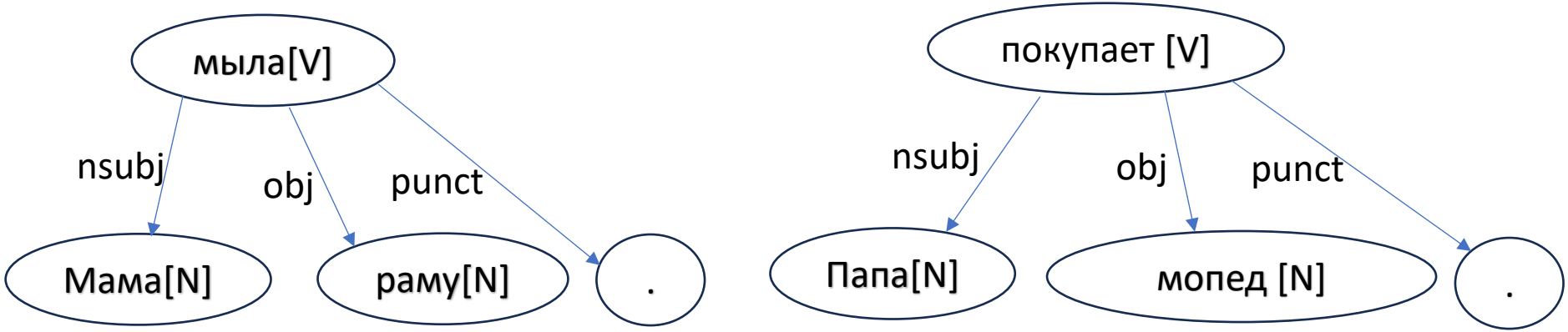
Цель подхода – создание инструмента для поддержки процесса разметки

Идея подхода – подобие деревьев - 1

Идея подхода – пытаться строить деревья, «подобные» деревьям из корпуса

Дерево T1 структурно эквивалентно дереву T2 из корпуса, если существует изоморфизм узлов (слов) деревьев T1 и T2, сохраняющий порядок слов, дуги, синтаксические классы слов и пометки дуг.

Мама мыла раму . ⇔ ***Папа покупает мопед .***



Идея подхода – подобие деревьев - 2

Ярус поддерева T - множество всех дуг, выходящих из вершины T.

Дерево T1 валидно относительно корпуса деревьев К, если для каждого яруса L1 дерева T1 существует ярус L2 дерева T2 из корпуса К, структурно эквивалентный L1.

Например, если корпус содержит предложения:

- ***Сегодня папа покупает мопед .***
- ***Темп экономического роста снижается.***

тогда предложение:

Сейчас физики ищут хиггсовский бозон.

валидно относительно этого корпуса, хотя **сам корпус может и не содержать эквивалентного по структуре предложения.**

Идея подхода – использование РАСП

Как строить?

Используются расширенные сети переходов (РАСП). Предложены Р.А. Вудсом (1973).

РАСП:

- Система именованных конечных автоматов (узлов), дуги которых помечены либо терминальными символами, либо именами узлов.
- Узлы содержат дескрипторы, и в процессе прохождения через дуги или после свертки узла над дескрипторами могут выполняться действия («операторы»).
- Операторы переписывают информацию из одного дескриптора в другой и/или проверяют какие-либо условия, фильтруя варианты разбора.

Генерация РАСП по корпусу деревьев – виды узлов

Для реализации подхода взят корпус СинТагРус – как в разметке НКРЯ (XML-формат), так и в разметке проекта Universal Dependencies (CONLL-U-формат).

Основная идея – реализовать указанное выше «подобие», построив сеть, все узлы которой – ациклические автоматы, распознающие фрагменты предложений с валидными относительно корпуса поддеревьями зависимостей.
Узлы генерируемой сети разбиты на два класса:

- **D(C)** – слово в узле имеет синтаксический класс C . Генерируется **по ярусам** деревьев из корпуса;
- **D(C,R)** - слово в узле имеет синтаксический класс C и связано в каком-либо дереве отношением R с каким-либо главным словом. Генерируется **по зависимым вершинам** деревьев из корпуса

Генерация РАСП по корпусу деревьев – узлы D(C)

Первый этап – генерация путей по ярусам поддеревьев корпуса

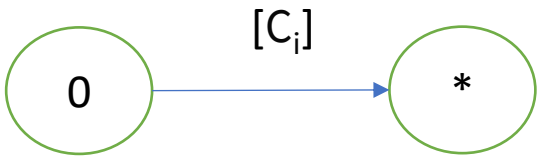
Пусть W(C) – корень какого-то яруса, C – синтаксический класс этого слова
$W_i(C_i)$ – зависимые от W(C) слова, R_i – пометка дуги дерева ($W_i(C_i)$, W(C), и индекс k таков, что слова W_i расположены левее W для i <= k и правее W для i > k
Тогда добавим в узел D(C) новый путь из начального состояния в конечное:

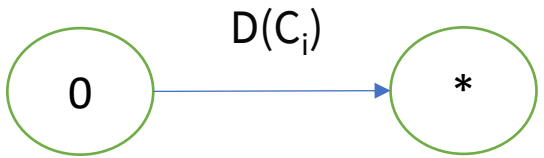
S_i, i = 1…n-1 – новые состояния, добавляемые в узел D(C)

Генерация РАСП по корпусу деревьев – узлы D(C,R)

Пусть W(C) – корень какого-то яруса, C – синтаксический класс этого слова
$W_i(C_i)$ – зависимые от W(C) слова, R_i – пометка дуги дерева ($W_i(C_i)$, W(C).
Тогда для каждого зависимого слова $W_i(C_i)$ добавим в узел $D(C_i, R_i)$ одну дугу сети из начального состояния в конечное.
Пометка этой дуги:

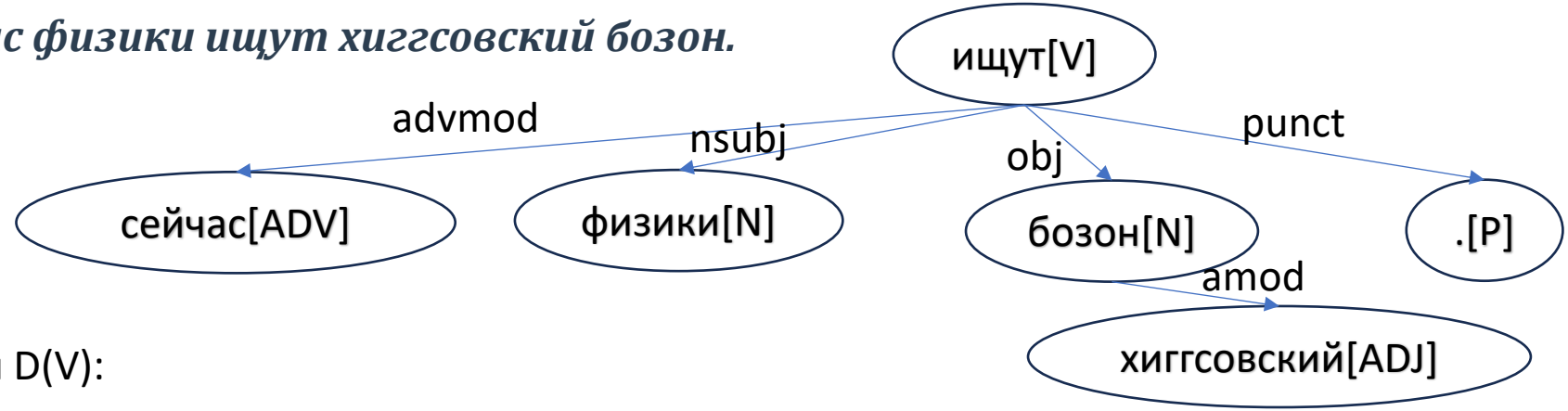
- $[C_i]$, если у $W_i(C_i)$ нет своих зависимостей (терминальная дуга)
- $D(C_i)$, если $W_i(C_i)$ является корнем другого яруса (нетерминальная дуга)



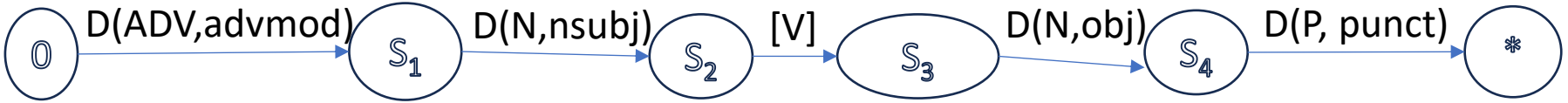


Генерация РАСП по корпусу деревьев – пример(1)

Сейчас физики ищут хиггсовский бозон.



Узел D(V):

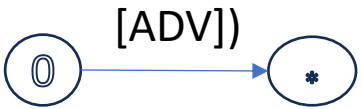


Узел D(N):

Генерация РАСП по корпусу деревьев – пример(2)

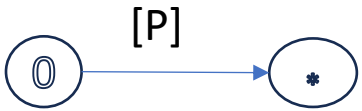
Сейчас физики ищут хиггсовский бозон.

Узел D(ADV, advmod):

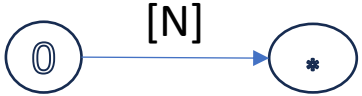


Узел D(ADJ, amod):

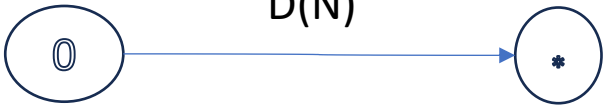
Узел D(P, punct):



Узел D(N, nsubj):



Узел D(N,nobj):



Генерация РАСП по корпусу деревьев – главный узел

Второй этап – минимизация и добавление главного узла

Если предложение в корпусе имеет корень с синтаксическим классом C, то добавляем в главный узел дугу из начального в конечное состояние, помеченную D(C)

Генерация РАСП по корпусу деревьев - операторы

Третий этап – добавление операторов (действий)

- Дуги с пометкой D(C,R) - копирование в дескриптор родительского узла информации о зависимых словах и их связях
- Дуги с пометкой D(C) - копирование в дескриптор родительского узла информации о главном слове узла D(C)
- Дуги с пометкой [C] – сохранение информации о главном слове в родительском узле
- При свертке узла D(C) – @CheckTier – проверка того, что зависимые слова в дескрипторе узла действительно связаны с главным словом нужным отношением

По выводу в такой сети можно построить дерево зависимостей

Использование РАСП

Использование сгенерированной РАСП требует реализации оператора @CheckTier, проверяющего, как связаны главное и зависимые слова в ярусе.
В общем случае оператор @CheckTier решает следующую задачу:

> Даны две словоформы W_1, W_2 и пометка R. Определить, могут ли заданные словоформы образовать дугу (W_1, W_2) с пометкой R в валидном дереве зависимостей.

Варианты решения поставленной задачи:

- Применение классификатора, обученного по корпусу
- Применение векторных представлений и меры семантической близости

(!) Возможно использования подхода для обнаружения несоответствий в существующей разметке корпуса и **без решения проблем проверки связей**.

Использование РАСП – контроль разметки - 1

Проверка корректности разметки:

допускает ли текущий вариант разметки предложения в корпусе,
другие варианты структуры при заданных пометках?

Решение с использованием сгенерированной сети:

разбор предложения с оператором проверки, который пропускает только «правильные» пометки из дерева разметки, отсекая остальные.

Если на входе предложение из подкорпуса, по которому строилась сеть, то результатом должно быть ЕДИНСТВЕННОЕ (и правильное) дерево разбора, а наличие других вариантов свидетельствует о несогласованности разметки.

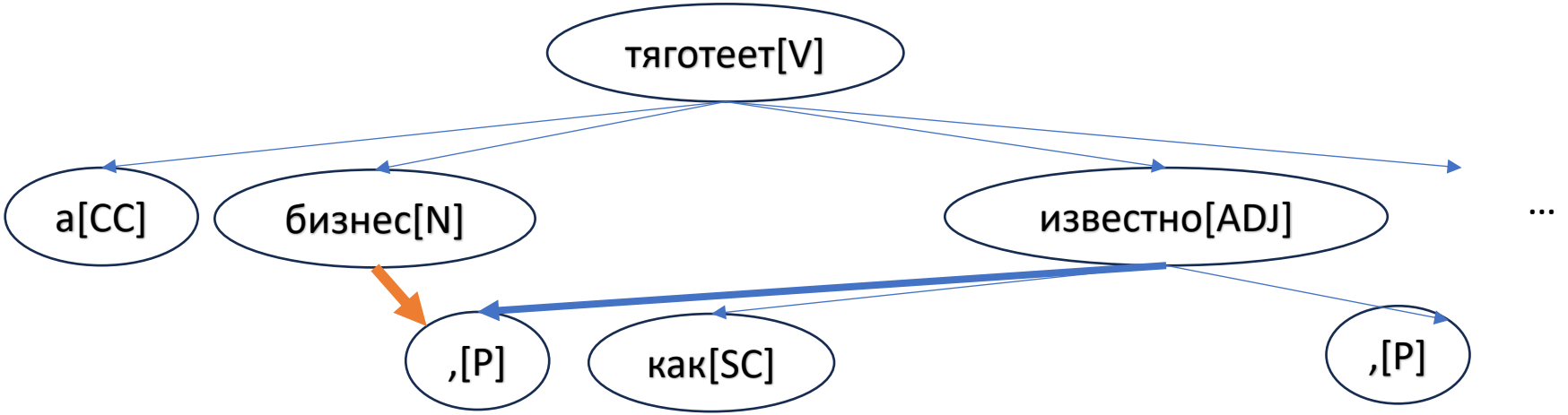
Использование РАСП – контроль разметки - 2

Эксперимент по анализу предложений обучающей выборки корпуса СинТагРус в разметке UD

Общее число предложений	64743
Число предложений с единственным вариантом	55957
Число предложений с количеством вариантов > 1 (только пункт.)	8779 (8220)
Число непрошедших предложений	7
Число предложений с правильным наиболее вероятным вариантом	3482
Общее число вариантов	92257

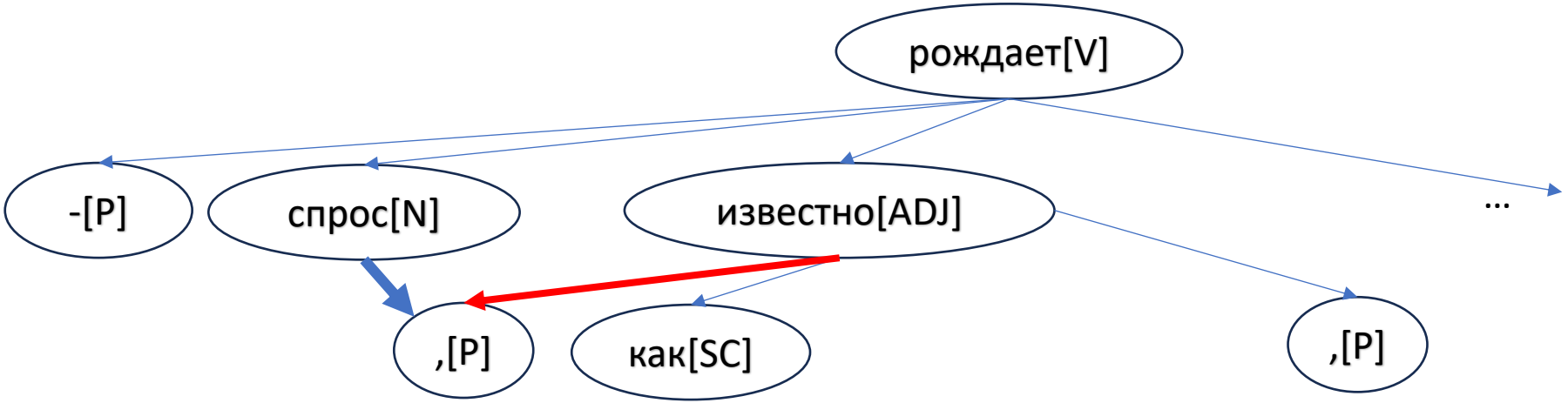
Использование РАСП – контроль разметки - 3

Основная причина проблемы – неоднозначная разметка зависимостей знаков препинания.
Предложение (sent_id = 2009Vysshaya_Shkola_Ekonomiki.xml_10, версия от 11.11.25)
А бизнес, как известно, тяготеет к профессионалам и специалистам, а не к дипломам и регалиям.



Использование РАСП – контроль разметки - 4

Пример разметки из корпуса, противоречащий предыдущему примеру.
Предложение (sent_id = sent_id = 2003Delit_na_vosem.xml_107, версия от 11.11.25).
- Спрос, как известно, рождает предложение.



Достоинства подхода

- Хорошая **интерпретируемость**. Для любого успешно проанализированного предложения можно предъявить примеры разметки, на основе которой построено дерево.
- **Многовариантность**. В процессе анализа входного предложения можно получить все возможные варианты его разбора для последующей оценки.
- Возможность введения **метрики уверенности** в достоверности полученного варианта, например, можно оценивать его вероятность.

Проблемы подхода

- **Не учитываются непроективные** предложения. Так, в текущем варианте UD-разметки корпуса СинТагРус из 78537 предложений в обучающем и валидационном подкорпусах 5298 предложений (7%) имеют непроективные деревья и не могут быть использованы.
- Невозможность выдачи деревьев, **фрагменты** структур которых **не встречаются** в корпусе.
- Порождение **чрезмерно большого числа вариантов для ряда предложений**, особенно со сложной структурой

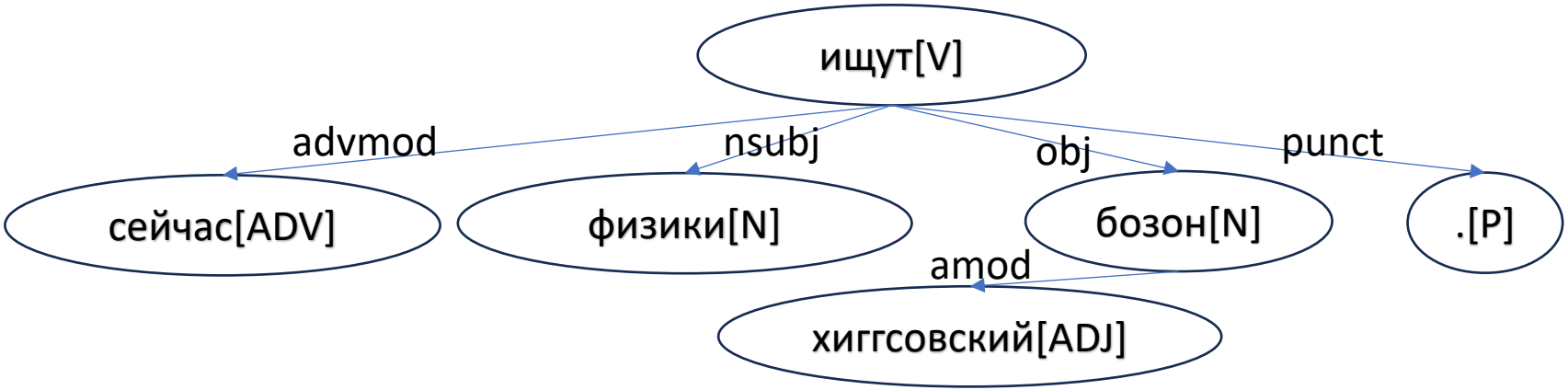
Выводы

- Указанные недостатки подхода не позволяют в настоящее время использовать его для основного синтаксического анализатора в системе автоматической разметки.
- Однако, метод может применяться в автоматизированных системах разметки как вспомогательный инструмент, например, для анализа качества полученных деревьев и для выявления проблем разметки отдельных предложений исходных корпусов.

Спасибо за внимание!

Идея подхода – подобие деревьев - 3

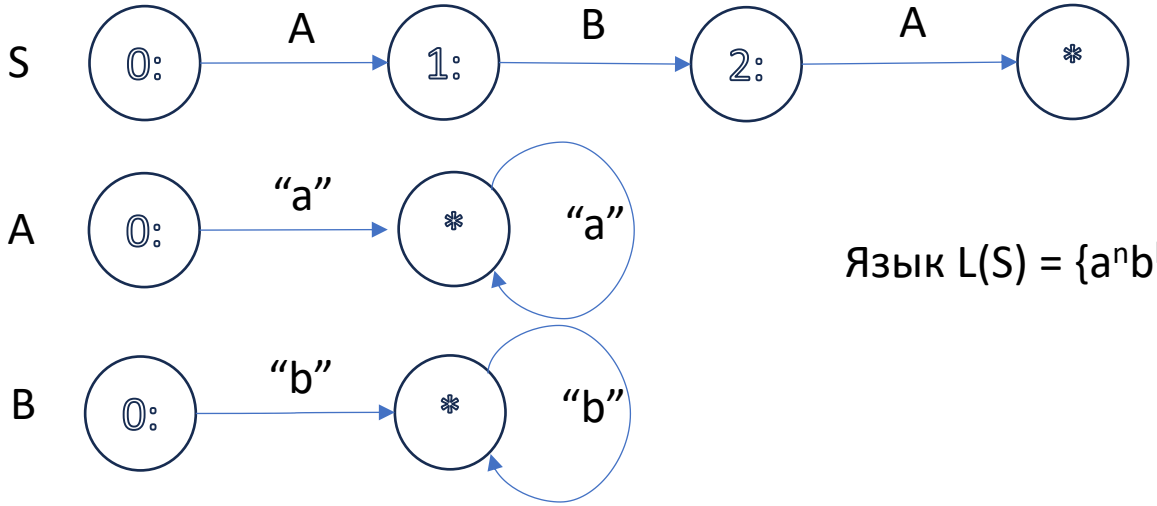
Сейчас физики ищут хиггсовский бозон.



Сегодня[ADV] папа[N] покупает[V] мопед[N] .[P]

Темп экономического[ADJ] роста[N] снижается.

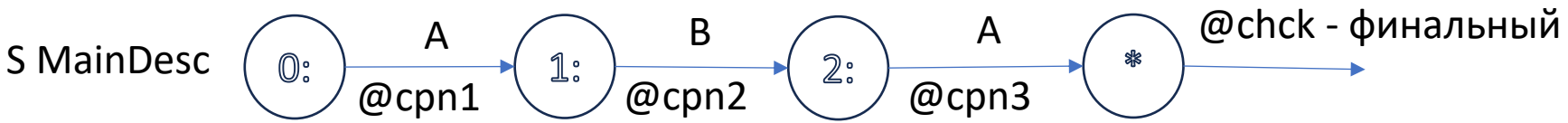
Пример РАСП без операторов

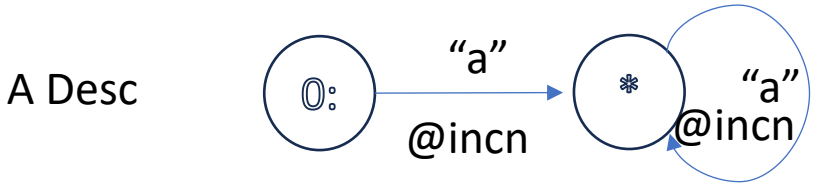


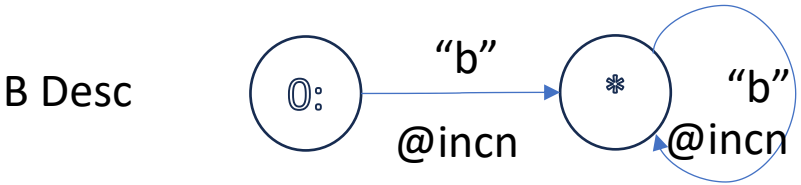
Язык $L(S) = \{a^n b^k a^m \mid n,k,m>=1\}$

Пример РАСП с операторами

DECLARE MainDesc (N1(0), N2(0), N3(0)), Desc (N(0))







@incn – увеличение поля N на 1

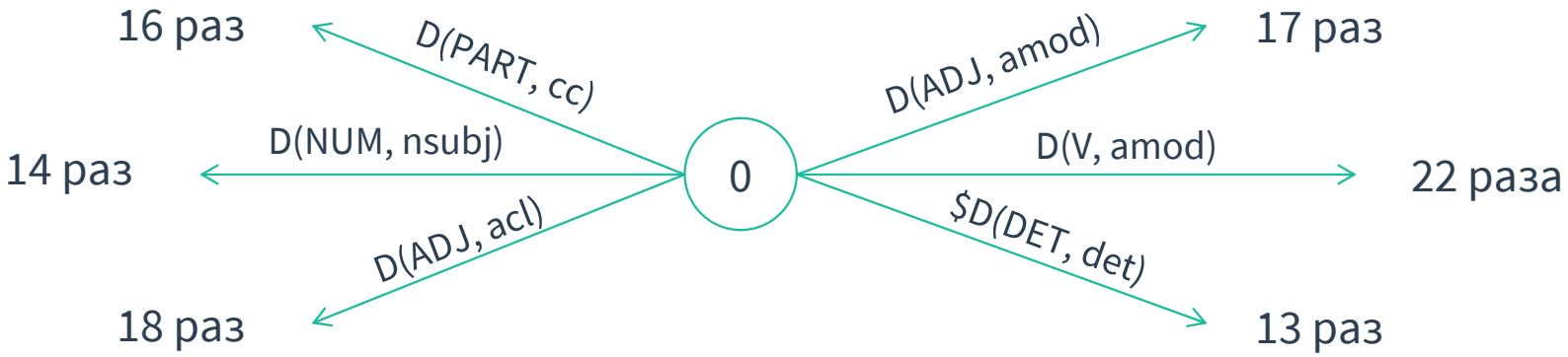
@cpn$_i$ - копирование поля N в поле N$_i$

@chck – проверка полей N$_i$ на равенство

Язык L(S) = {$a^n b^n a^n$ | n,k,m>=1}

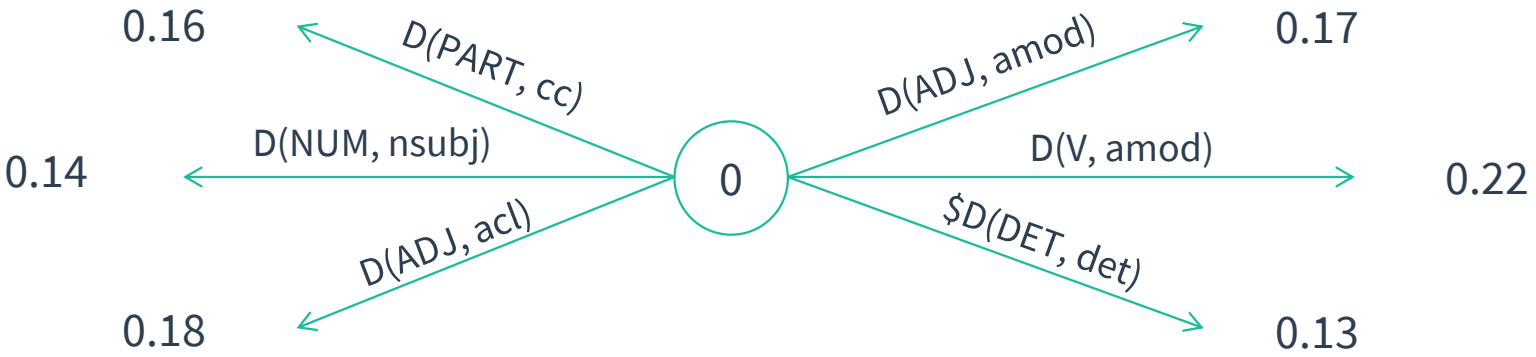
Добавление вероятностей - 1

Каждой дуге приписывается вероятность ее прохождения.
Вероятности рассчитываются по всем деревьям из обучающей части корпуса.



Добавление вероятностей - 2

Каждой дуге приписывается вероятность ее прохождения.
Вероятности рассчитываются по всем деревьям из обучающей части корпуса.



Использование РАСП – контроль разметки - 5

Предложение (sent_id = 2006Vse_vozrastu_pokorny.xml_117, версия от 11.11.25).
С помощью современных приборов у пациента измеряются давление, содержание холестерина, вес , пропорции фигуры , степень облысения , сила пожатия руки , наличие собственных зубов во рту , острота зрения , показатели дыхания , состояние кожи , гибкость суставов , обоняние , двигательные реакции и так далее.

Результат: 1536 вариантов разбора

Предложение (sent_id = sent_id = 2003Somnambula_v_tumane.xml_355, версия от 11.11.25).
Или он приходит в ванную, кладет голову на край умывальника и плачет, плачет, как Денисов , плачет , оплакивает свою бессмысленную жизнь, морскую пустоту, обманчивую красоту лиловых островов, людские пороки, женскую глупость, оплакивает утонувших, погибших, забытых, преданных, ненужных; слезы текут по замызганному рукомойному фаянсу, льются на пол, вот уже поднялось до щиколоток, вот дошли до колена, рябь, круги, ветер, шторм.

Результат: отказ анализатора из-за нехватки памяти