РЕФЕРЕНЦИАЛЬНЫЙ ВЫБОР

КАК МНОГОФАКТОРНЫЙ ВЕРОЯТНОСТНЫЙ ПРОЦЕСС

А.А.Кибрик, Г.Б.Добров,
Д.А.Залманов, А.С.Линник,
Н.В.Лукашевич

aakibrik@gmail.com

Референциальный выбор в дискурсе

- Когда говорящему нужно упомянуть некоторый конкретный, определенный референт, он делает выбор из нескольких возможностей, в том числе:
  - полной именной группы (ИГ)
    - имя собственное
    - имя нарицательное (с модификаторами) = дескрипция
  - редуцированной ИГ, напр. местоимения 3 лица или нулевого выражения

➢ Как осуществляется этот выбор?

Пример (фрагмент из рассказа Ф. Искандера “Сталин и Вучетич”)

Полная ИГ

антецедент

кореферентность

Сталин мирно беседовал с Вучетичем. "Товарищ Сталин, что такое старость?" - спросил Вучетич, разумеется, имея в виду философский смысл проблемы.

И вдруг лицо Сталина мгновенно исказилось гневом и ненавистью. Он стал страшен. Вучетич помертвел, ∅ не в силах осознать, чем ∅ разгневал Сталина.

Место-имение

нуль

План доклада

- I. Референциальный выбор как многофакторный процесс
- II. Количественный и нейросетевой подходы к референциальному выбору
- III. Корпусное исследование RefRhet: состояние и перспективы

Многофакторный характер референциального выбора

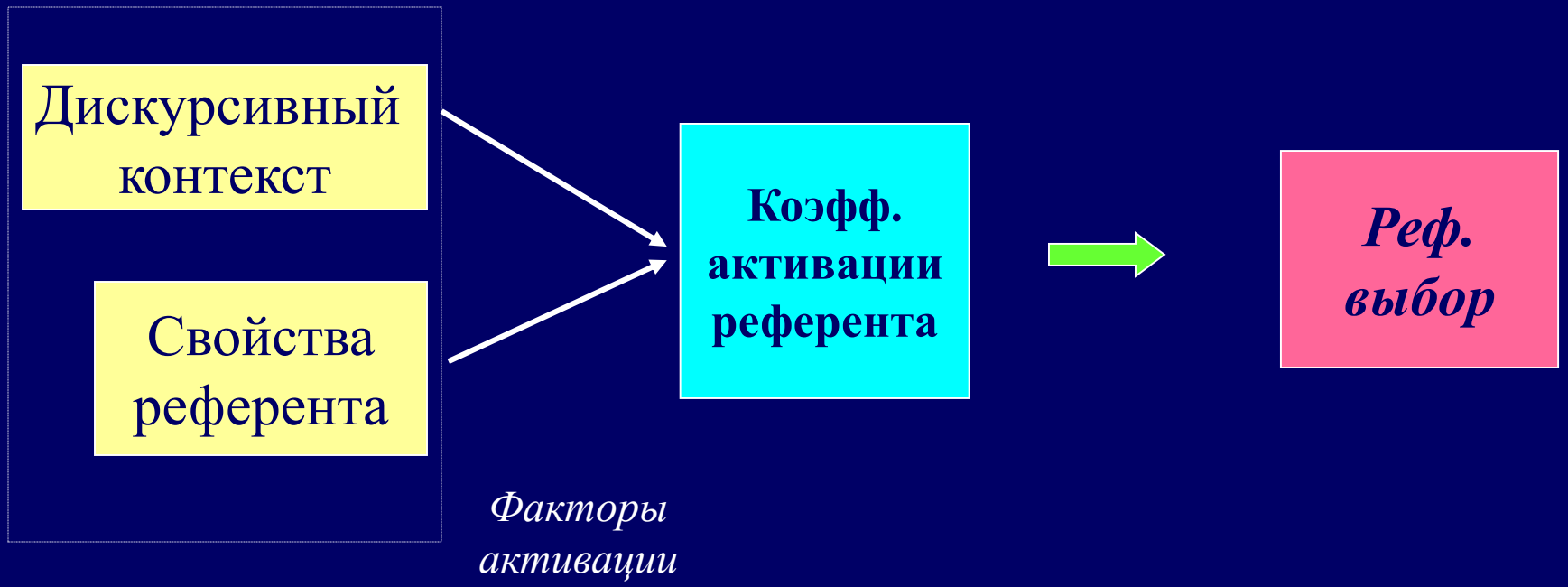
- Существует большое число факторов референциального выбора
  - Расстояние до антецедента
    - По линейной структуре дискурса
    - По иерархической структуре дискурса
    - По глобальной структуре дискурса
  - Роль антецедента
  - Одушевленность референта
  - Протагонизм

  .....
- Ни один из этих факторов в отдельности не может объяснить референциальный выбор

Интеграция факторов

- В каждой точке дискурса все факторы некоторым образом суммируются и порождают интегральную характеристику, которую можно назвать коэффициентом активации референта
- Коэффициент активации предопределяет референциальный выбор
  - Низкий → полная ИГ
  - Средний → полная или редуцированная ИГ
  - Высокий → редуцированная ИГ

Когнитивная многофакторная модель референциального выбора



Количественный подход (Kibrik 1996, 1999)

- Каждый фактор – это переменная, имеющая набор возможных значений
- Каждому из значений переменной соответствует числовой вес
- В каждой точке дискурса для каждого референта могут быть идентифицированы значения всех факторов и, соответственно, все их количественные вклады
- Проблемы исследования:
  - Детерминированная зависимость
  - Не моделируется нелинейное взаимодействие между факторами
  - Веса были подобраны вручную

Нейросетевой подход (Gruening and Kibrik 2005)

- Алгоритм машинного обучения
- Нелинейное взаимодействие факторов
- Автоматическое приписывание весов
- Возможность редуцировать число факторов («обрезка»)
- Проблемы исследования:
  - Малый объем данных
  - Лишь один метод машинного обучения
  - Невысокая скорость обучения
  - Низкая трактуемость результата
  - Исчезновение когнитивной интерпретации

Дальнейшее развитие исследований

- Большой корпус (несколько десятков тысяч реф. выражений)
- Более точные процедуры контроля качества
- Определение оптимального набора факторов, объясняющего референциальный выбор
- Применение большего числа методов машинного обучения
- Построение статистической модели реф. выбора
- Восстановление когнитивной интерпретации

Корпус RefRhet

- Английский язык
- Деловая проза
- Исходный материал - корпус RST Discourse Treebank
  - Аннотирован по иер. структуре
  - 385 газетных статей из Wall Street Journal
- Дополнительный компонент – референциальная разметка
- Корпус RefRhet
  - Около 30 000 референциальных выражений

Пример иерархического графа

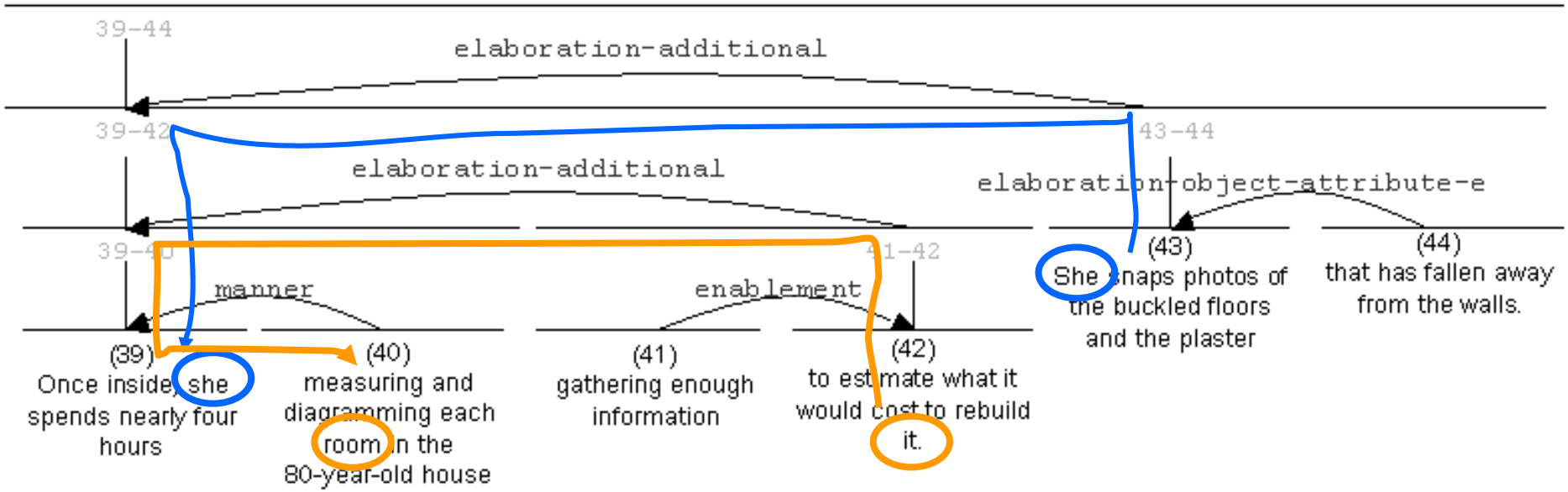
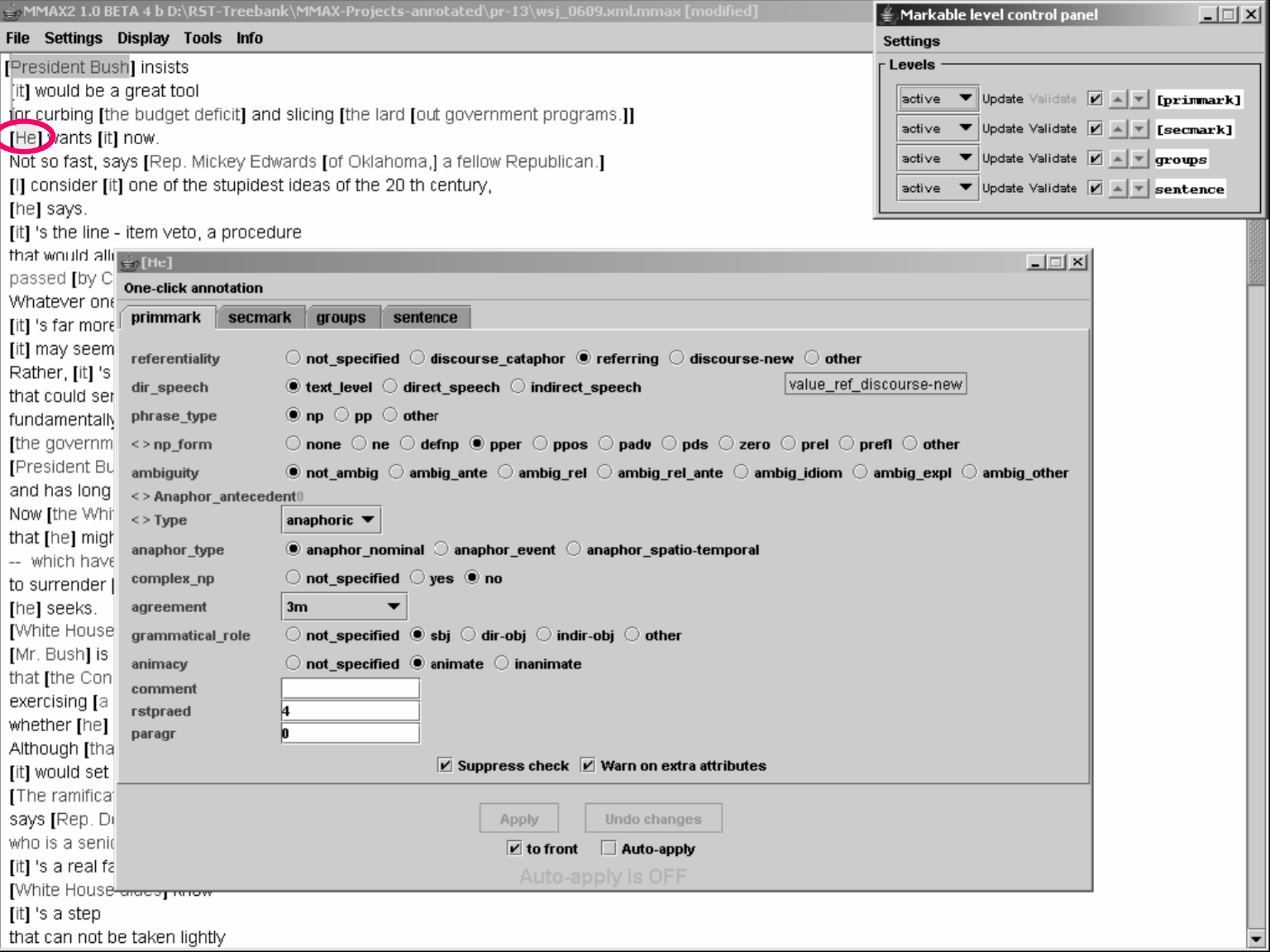


Схема референциальной разметки

- Программа MMAX2
- Krasavina and Chiarcos 2007
- Размечены все аннотируемые выражения (маркабулы – markables)
  - референциальные выражения
  - их антецеденты
- Размечены кореферентные связи
- Размечены признаки реф. выражений и контекста, которые могут быть факторами реф. выбора


MMAX2 1.0 BETA 4 b D:\RST-Treebank\MMAX-Projects-annotated\pr-13\wsj_0609.xml.mmax [modified]
File Settings Display Tools Info
[President Bush] insists
[it] would be a great tool
for curbing [the budget deficit] and slicing [the lard [out government programs.]]
[He] wants [it] now.
Not so fast, says [Rep. Mickey Edwards [of Oklahoma,] a fellow Republican.]
[I] consider [it] one of the stupidest ideas of the 20 th century,
[he] says.
[it] 's the line - item veto, a procedure
that would all
passed [by C
Whatever one
[it] 's far more
[it] may seem
Rather, [it] 's
that could ser
fundamentally
[the governm
[President Bu
and has long
Now [the Whi
that [he] migh
-- which have
to surrender
[he] seeks.
[White House
[Mr. Bush] is
that [the Con
exercising [a
whether [he]
Although [tha
[it] would set
[The ramifica
says [Rep. Du
who is a senic
[it] 's a real fa
[White House aides] know
[it] 's a step
that can not be taken lightly
Markable level control panel
Settings
Levels
active Update Validate [primmark]
active Update Validate [secmark]
active Update Validate groups
active Update Validate sentence
[He]
One-click annotation
primmark secmark groups sentence
referentiality not_specified discourse_cataphor referring discourse-new other
dir_speech text_level direct_speech indirect_speech
value_ref_discourse-new
phrase_type np pp other
< > np_form none ne defnp pper ppos padv pds zero prel prefl other
ambiguity not_ambig ambig_ante ambig_rel ambig_rel_ante ambig_idiom ambig_expl ambig_other
< > Anaphor_antecedent
< > Type anaphoric
anaphor_type anaphor_nominal anaphor_event anaphor_spatio-temporal
complex_np not_specified yes no
agreement 3m
grammatical_role not_specified sbj dir-obj indir-obj other
animacy not_specified animate inanimate
comment
rstpraed 4
paragr 0
Suppress check Warn on extra attributes
Apply Undo changes
to front Auto-apply
Auto-apply is OFF

Создание референциальной разметки

- О. Красавина
- А. Антонова
- Д. Залманов
- А. Линник
- М. Худякова
- Студенты-практиканты ОТиПЛ

Состояние референциальной разметки корпуса RefRhet

- Размечен на 2/3
- Дальнейшие результаты основаны на следующих данных
  - 247 текстов
  - 110 тыс. словоупотреблений
  - 26 024 маркабул
    - 7097 имен собственных
    - 8560 определенных дескрипций
    - 1797 местоимений 3 лица
  - 3756 надежных пар «анафор – антецедент»
    - имена собственные — 1623 (43%)
    - определенные дескрипции — 971 (26%)
    - местоимения — 1162 (31%)

Факторы референциального выбора

- Признаки референта:
  - первое/непервое упоминание в дискурсе (referentiality)
  - одушевленность (animacy)
  - протагонизм
- Признаки антецедента:
  - Тип синтаксической группы (phrase_type)
  - Грамматическая роль (gramm_role)
  - Референциальная форма (np_form, def_np_form)
  - Входит ли в состав прямой речи (dir_speech)

Факторы референциального выбора

- Признаки анафора:
  - Тип синтаксической группы (phrase_type)
  - Грамматическая роль (gramm_role)
  - Входит ли в состав прямой речи (dir_speech)
- Расстояния между анафором и антецедентом:
  - Расстояние в словах
  - Расстояние в маркабулах
  - Линейное расстояние в клаузах
  - Иерархическое расстояние в элементарных дискурсивных единицах

Постановка задачи машинного обучения

- Зависимая переменная:
  - Референциальная форма (np_form)
- Двуклассовая задача:
  - полная ИГ vs. местоимение
- Трехклассовая задача:
  - определенная дескрипция vs. имя собственное vs. местоимение
- Максимизируем аккуратность:
  - отношение правильных случаев предсказания к общему количеству

Методы машинного обучения (Weka)

- Легко интерпретируемые методы:
  - Логические алгоритмы
    - Деревья решений (C4.5)
    - Решающие правила (JRip)
  - Более высокое качество:
    - Логистическая регрессия
- Контроль качества – метод скользящего контроля

Примеры правил, порождаемых алгоритмом JRip

- *(Грамматическая роль антецедента = подлежащее) И*
  (Иерархическое расстояние ≤ 1.5) И
  (Расстояние в словах ≤ 7)
  => местоимение

- *(Одушевленный) И*
  (Расстояние в маркабулах ≥ 2) И
  (Расстояние в словах ≤ 11)
  => местоимение

Основные результаты

- Аккуратность
- Двуклассовая задача:
  - логистическая регрессия - 86.1%
  - логические алгоритмы - 85%
- Трехклассовая задача:
  - логистическая регрессия - 74%
  - логические алгоритмы - 72%

Многофакторность выбора

Признак	Трехклассовая задача	Двуклассовая задача
Наибольший класс	43%	69%
Расстояние в словах	55%	76%
Иерархическое расстояние	53.5%	74.8%
Грамматическая роль анафора	45.2%	70%
Анафор в прямой речи	43.8%	70%
Одушевленный	47.3%	71.5%
Комбинация факторов	74%	86.1%

Референциальный выбор – вероятностный процесс

- По данным Kibrik 1999

Потенциальные референциальные выражения	Фактические реф. выражения
Только полная ИГ (19%)	Полная ИГ (49%)
Полная ИГ, ?местоимение (21 %)	
Местоимение или полная ИГ (28%)	Местоимение (51%)
Местоимение, ?полная ИГ (23%)	
Только местоимение (9%)	

Перспективы вероятностной модели

- Предсказание реф. выбора не может быть полностью детерминированным
- Есть часть случаев, когда реф. выбор является произвольным
- Важно настроить модель так, чтобы она обрабатывала такие случаи особым образом
- Это задача для дальнейших исследований
- Логистическая регрессия выдает оценки вероятности для каждой из опций референциального выбора

Вероятностная многофакторная модель референциального выбора

Выводы

- Большой корпус для референциальных исследований
- Многофакторность
- Уже достигнут высокий уровень правильного предсказания реф. выбора
- И это еще не предел
- Вероятностный характер реф. выбора
- Возможно, вероятностную оценку можно проинтерпретировать как коэффициент активации из когнитивной модели
- Применимость для широкого круга языковых выборов