

# Основы обработки текстов

## Лекция 9

### Машинный перевод

# План

- Применение машинного перевода
- Сложности перевода
  - Типология
  - Различия языков
- Классический подход
- Статистический подход
  - Модель зашумленного канала
  - Выравнивание
  - Тренировка моделей
  - Декодирование
  - Методы оценки

# Применение машинного перевода

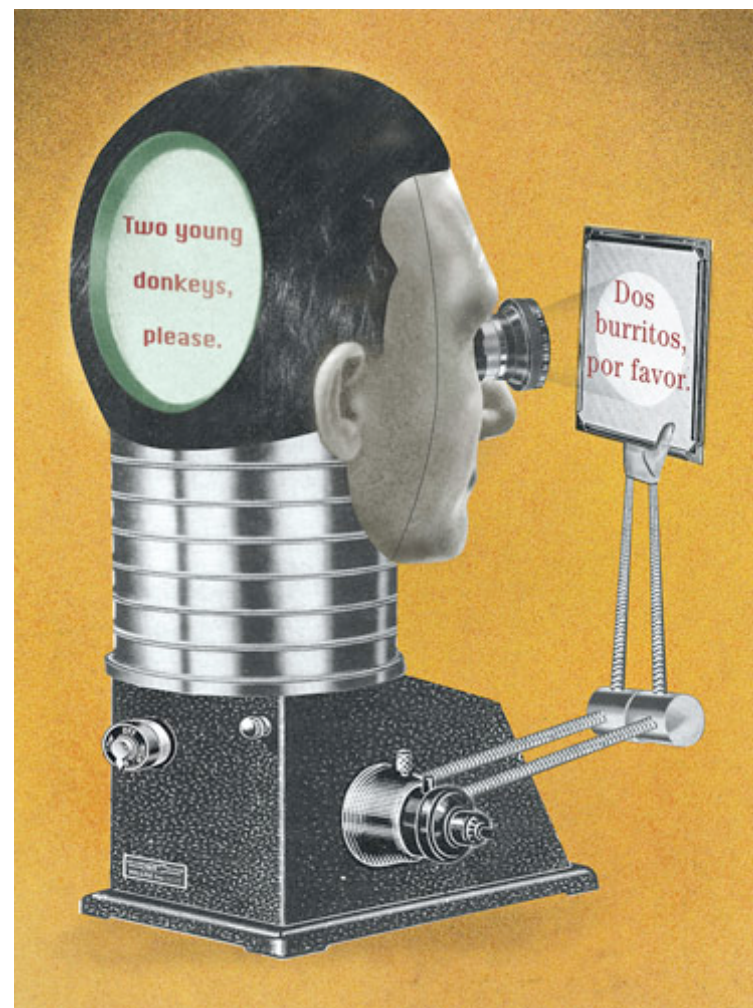
- Задачи, где достаточно грубого перевода
  - Задачи извлечения информации
  - Перевод Веб-страниц
  - e-mail
- Задачи, где результат перевода может быть исправлен
  - Помощь переводчику
- Перевод подмножеств языка
  - FANQT (Fully Automatic High Quality Translation)

# Где машинный перевод недостаточно хорош

- Художественная литература
- Разговорный язык
- Медицинский перевод в больницах
- Звонки в службу спасения

# Сложность перевода

- Некоторые аспекты языков схожи, некоторые различны
- Различия в
  - морфологии
  - лексике
  - структуре



# Морфология

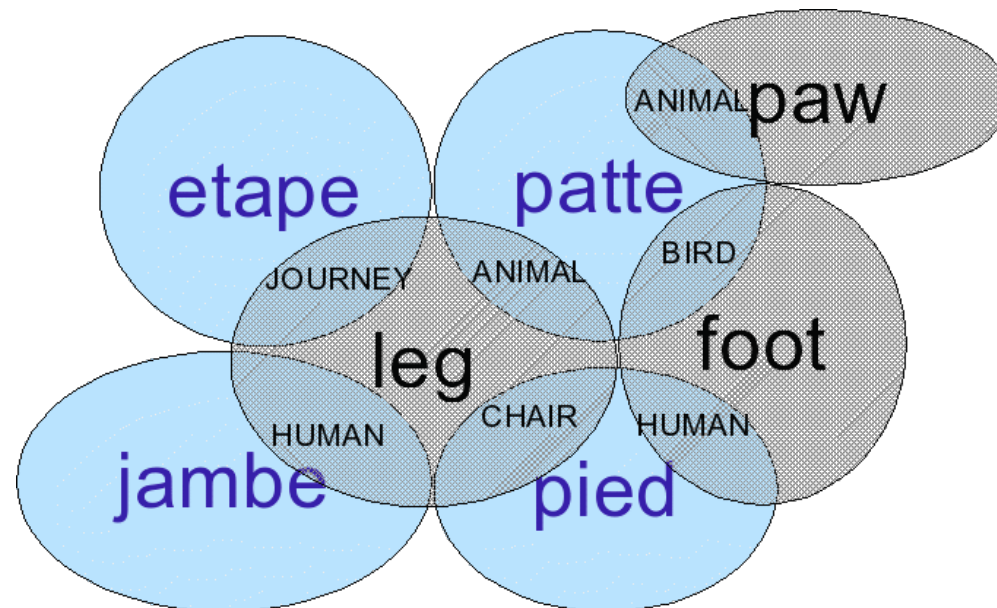
- Морфема
  - минимальная значимая единица языка
  - слово = морфема + морфема + морфема + ...
- Аффиксы
  - Префикс: **un**do
  - Суффикс: look**ing**
  - Инфикс: hingi (занимать) - h**um**ingi (заемщик)  
(Тагальский язык)
  - Циркумфикс: sagen (сказать) - **ge**sagt (сказал)  
(Немецкий)

# Морфологические различия

- Изолирующие языки
  - Каждое слово состоит из одной морфемы (Вьетнамский)
- Полисинтетические языки
  - слово состоит из множества морфем (Чукотский: **Тымэйн****ы****левтпыгтыркын** - У меня сильно болит голова)
- Агглютинативные
  - Морфемы несут определенные значения (Турецкий)
- Флективные
  - Морфемы имеют несколько значений (Русский: “хороший” - им. падеж, ед. число, муж. род)

# Лексические различия

- Семантические особенности:
  - В корейском нет слов брат/сестра, есть старший/младший брат/сестра
  - В чукотском около 20 слов для снега
- Английский vs французский





# Синтаксические различия

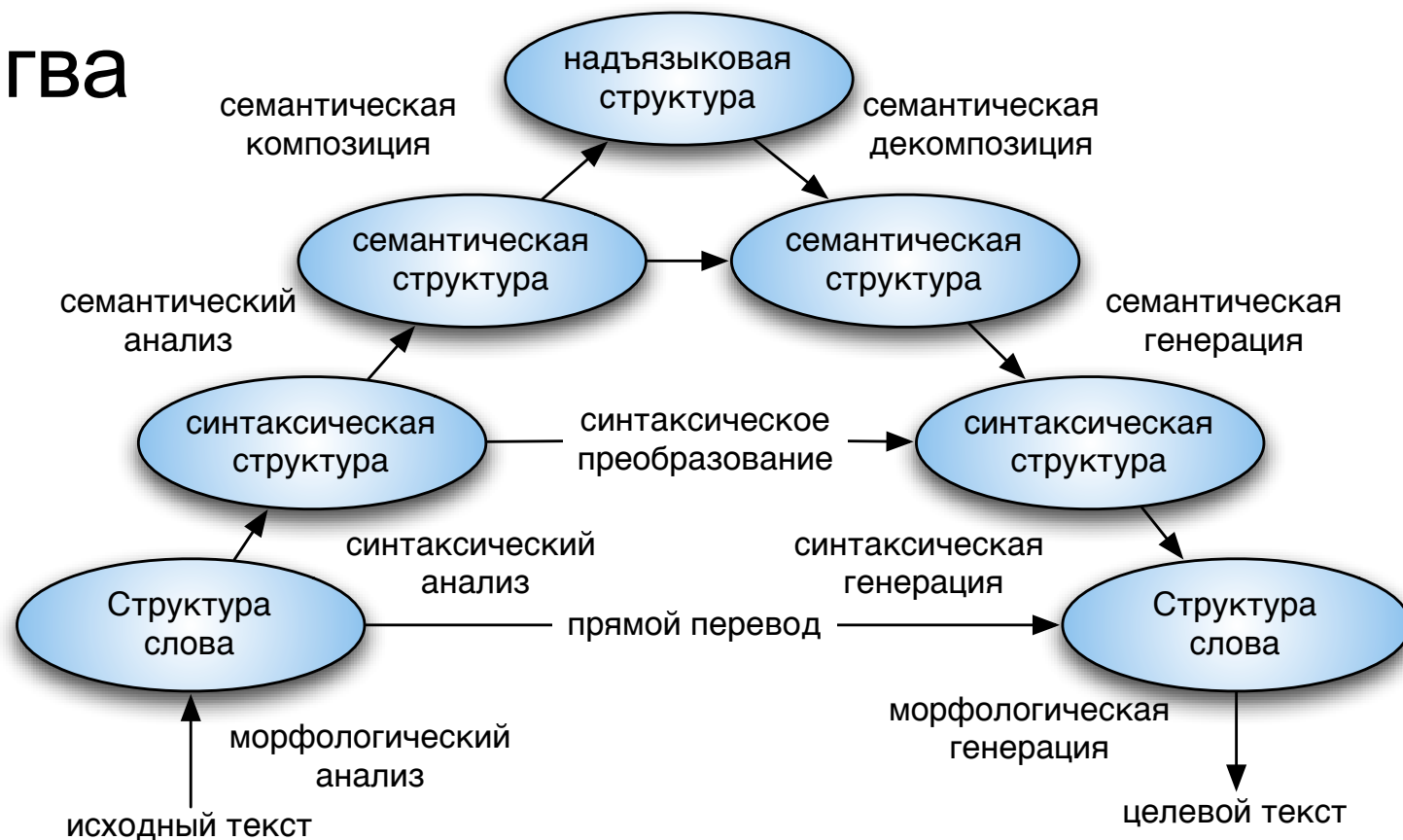
- СГО (Субъект-Глагол-Объект)
  - **Английский**, **Немецкий**
  - I am in Moscow
- СОГ
  - Японский, **Корейский**
  - 저는 **모스크바에** 있습니다 (**Я** в **Москве** **нахожусь**)
- ГСО
  - Ирландский, классический Арабский

# Границы

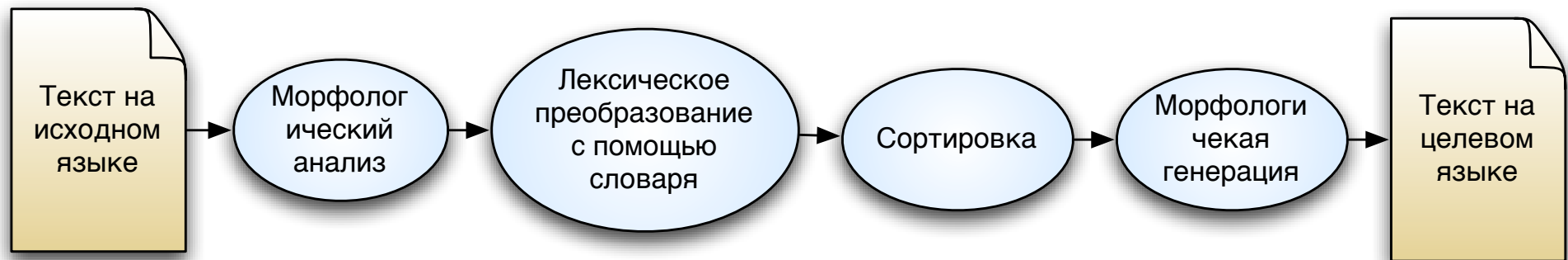
- Языки в которых не выделены границы слов:
  - Китайский, Японский, Тайский, Вьетнамский
- Предложения в некоторых языках больше похожи на параграфы
  - Китайский, современный Арабский

# Классические подходы

- Прямой перевод
- Преобразование
- Интерлингва



# Подход 1: Прямой перевод



- Последовательный перевод каждого слова
- Не используется никакие структуры кроме морфологии
- После перевода слов, делается сортировка

# Пример

|                           |   |
|---------------------------|---|
| Input:                    | Mary didn't slap the green witch                |
| After 1: Morphology       | Mary DO-PAST not slap the green witch           |
| After 2: Lexical Transfer | Maria PAST no dar una bofetada a la verde bruja |
| After 3: Local reordering | Maria no dar PAST una bofetada a la bruja verde |
| After 4: Morphology       | Maria no dió una bofetada a la bruja verde      |

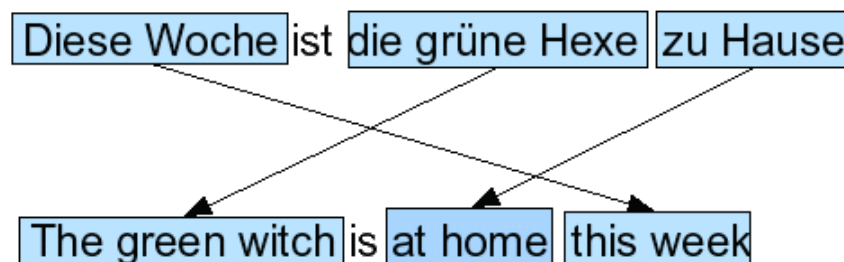
# Проблемы

- Сложные перестановки

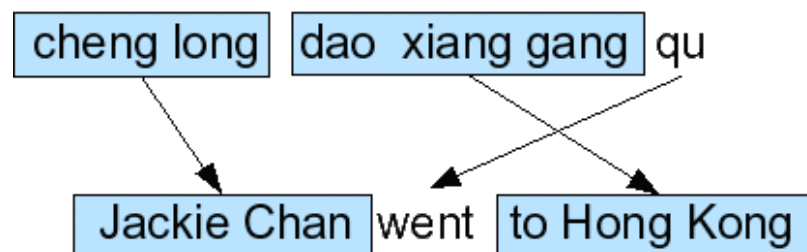
- термины

- длинные дистанции

- Немецкий



- Китайский

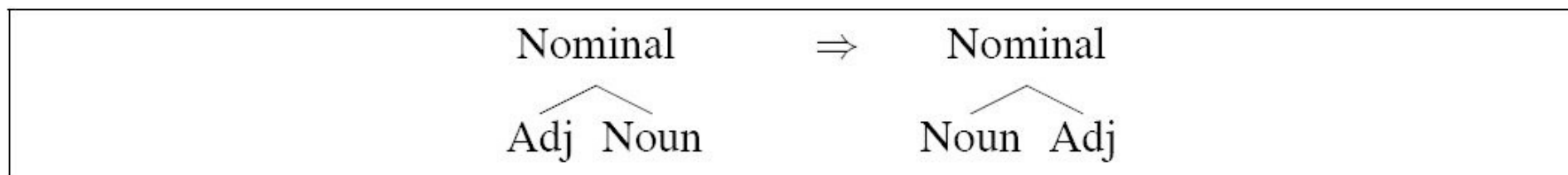


## Подход 2: Преобразование

- Применение знаний о различиях в языках
- Шаги
  - Анализ: синтаксический разбор исходного предложения
  - Преобразование: правила преобразования разбора в разбор на целевом языке
  - Генерация предложения на целевом языке

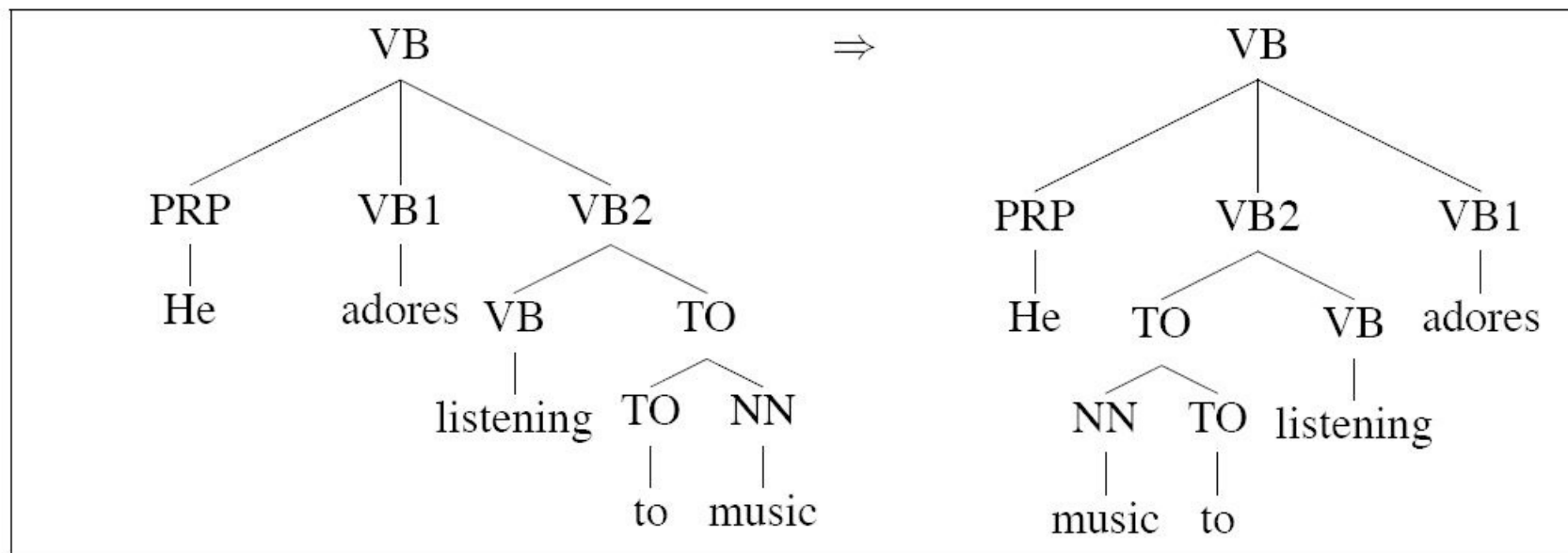
# Пример

- Английский: прилагательное существительное
- Французский: существительное прилагательное
- Не всегда
- Правило





# Правила преобразования



| English to Spanish:  |  |               |  |
|----------------------|--|---------------|--|
| 1.                   | $NP \rightarrow \text{Adjective}_1 \text{ Noun}_2$ | $\Rightarrow$ | $NP \rightarrow \text{Noun}_2 \text{ Adjective}_1$ |
| Chinese to English:  |  |               |  |
| 2.                   | $VP \rightarrow PP[+\text{Goal}] V$                | $\Rightarrow$ | $VP \rightarrow V PP[+\text{Goal}]$                |
| English to Japanese: |  |               |  |
| 3.                   | $VP \rightarrow V NP$                              | $\Rightarrow$ | $VP \rightarrow NP V$                              |
| 4.                   | $PP \rightarrow P NP$                              | $\Rightarrow$ | $PP \rightarrow NP P$                              |
| 5.                   | $NP \rightarrow NP_1 \text{ Rel. Clause}_2$        | $\Rightarrow$ | $NP \rightarrow \text{Rel. Clause}_2 NP_1$         |

# Systran: комбинирование ПОДХОДОВ

- Анализ
  - Морфологический, определение частей речи
  - Группировка (chunking)
  - Разбор некоторых зависимостей
- Преобразование
  - перевод идиом
  - Разрешение лексической многозначности
  - назначение предлогов на основе моделей управления глаголов
- Синтез
  - Применения большого двуязычного словаря
  - сортировка
  - морфологическая генерация

# Проблемы

- Грамматика и лексика содержат много специфики
- Трудно сделать и еще труднее поддерживать

# Интерлингва

- Пример системы: ABVYU Comreno
- Идея: Вместо использования правил преобразования между языками использовать значение предложения
- Шаги
  - Перевести исходное предложение в представление его значения
  - Сгенерировать целевое предложение из значения

# Интерлингва

Mary did not slap the green witch

|              |   |           |              |     |            |  |           |       |
|--------------|---|-----------|--------------|-----|------------|--|-----------|-------|
| EVENT        | SLAPPING  |           |              |     |            |  |           |       |
| AGENT        | MARY  |           |              |     |            |  |           |       |
| TENSE        | PAST  |           |              |     |            |  |           |       |
| POLARITY     | NEGATIVE  |           |              |     |            |  |           |       |
| THEME        | <table><tr><td>WITCH</td></tr><tr><td>DEFINITENESS</td><td>DEF</td></tr><tr><td>ATTRIBUTES</td><td><table><tr><td>HAS-COLOR</td><td>GREEN</td></tr></table></td></tr></table> | WITCH     | DEFINITENESS | DEF | ATTRIBUTES | <table><tr><td>HAS-COLOR</td><td>GREEN</td></tr></table> | HAS-COLOR | GREEN |
| WITCH        |   |           |              |     |            |  |           |       |
| DEFINITENESS | DEF   |           |              |     |            |  |           |       |
| ATTRIBUTES   | <table><tr><td>HAS-COLOR</td><td>GREEN</td></tr></table>  | HAS-COLOR | GREEN        |     |            |  |           |       |
| HAS-COLOR    | GREEN   |           |              |     |            |  |           |       |

# Проблемы

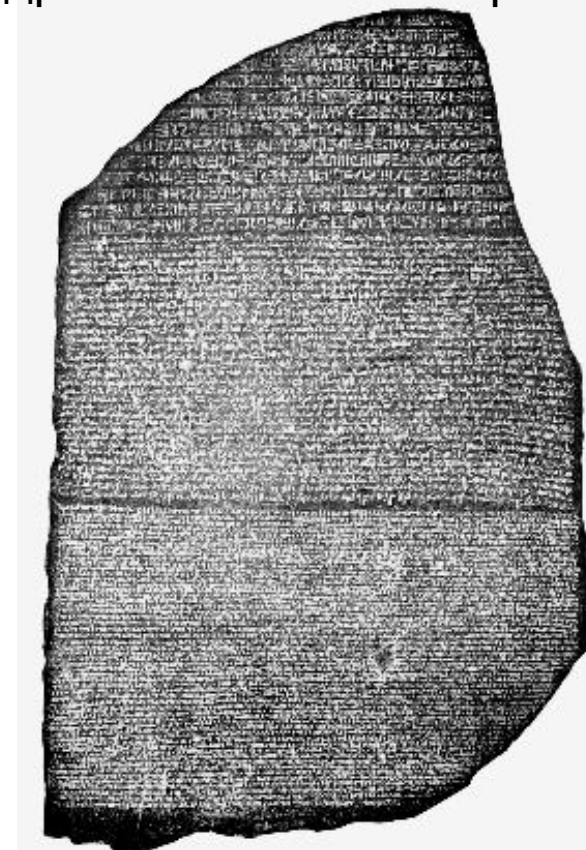
- Разные понятия в языках
  - 20 типов снега в Чукотском
  - Не нужны для англо-русского перевода
- Всесторонний анализ семантики и представление знаний
  - Возможно сделать только для специфичных подмножеств языка

# Статистический машинный перевод

- Идеи:
  - Использование параллельных текстов
  - Перевод по фразам
  - Сортировка результата

Розеттский камень:

- древнегреческий
- древнеегипетский
- древнеегипетские иероглифы



# Перевод на основе параллельных корпусов

Перевод с Центаврианского на Арктуранский [Knight, 1997]

Перевести: **farok crrrok hihok yorok klok kantok ok-yurp**

|   |  |
|---|--|
| 1a. ok-voon ororok sprok .              | 7a. lalok farok ororok lalok sprok izok enemok . |
| 1b. at-voon bichat dat .                | 7b. wat jjat bichat wat dat vat eneat .          |
| 2a. ok-drubel ok-voon anak plok sprok . | 8a. lalok brok anak plok nok .                   |
| 2b. at-drubel at-voon pippat rrat dat . | 8b. iat lat pippat rrat nnat .                   |
| 3a. erok sprok izok hihok ghrok .       | 9a. wiwok nok izok kantok ok-yurp .              |
| 3b. totat dat arrat vat hilat .         | 9b. totat nnat quat oloat at-yurp .              |
| 4a. ok-voon anak drok brok jok .        | 10a. lalok mok nok yorok ghrok klok .            |
| 4b. at-voon krat pippat sat lat .       | 10b. wat nnat gat mat bat hilat .                |
| 5a. wiwok farok izok stok .             | 11a. lalok nok crrrok hihok yorok zanzanok .     |
| 5b. totat jjat quat cat .               | 11b. wat nnat arrat mat zanzanat .               |
| 6a. lalok sprok izok jok stok .         | 12a. lalok rarok nok izok hihok mok .            |
| 6b. wat dat krat quat cat .             | 12b. wat nnat forat arrat vat gat .              |



# Перевод на основе параллельных корпусов

Перевод с Центаврианского на Арктуранский [Knight, 1997]

Перевести: **farok** crrrok hihok yorok klok kantok ok-yurp

|   |   |
|---|---|
| 1a. ok-voon ororok sprok .              | 7a. lalok <b>farok</b> ororok lalok sprok izok enemok . |
| 1b. at-voon bichat dat .                | 7b. wat <b>jjat</b> bichat wat dat vat eneat .          |
| 2a. ok-drubel ok-voon anak plok sprok . | 8a. lalok brok anak plok nok .                          |
| 2b. at-drubel at-voon pippat rrat dat . | 8b. iat lat pippat rrat nnat .                          |
| 3a. erok sprok izok hihok ghirok .      | 9a. wiwok nok izok kantok ok-yurp .                     |
| 3b. totat dat arrat vat hilat .         | 9b. totat nnat quat oloat at-yurp .                     |
| 4a. ok-voon anak drok brok jok .        | 10a. lalok mok nok yorok ghirok klok .                  |
| 4b. at-voon krat pippat sat lat .       | 10b. wat nnat gat mat bat hilat .                       |
| 5a. wiwok <b>farok</b> izok stok .      | 11a. lalok nok crrrok hihok yorok zanzanok .            |
| 5b. totat <b>jjat</b> quat cat .        | 11b. wat nnat arrat mat zanzanat .                      |
| 6a. lalok sprok izok jok stok .         | 12a. lalok rarok nok izok hihok mok .                   |
| 6b. wat dat krat quat cat .             | 12b. wat nnat forat arrat vat gat .                     |

# Перевод на основе параллельных корпусов

Перевод с Центаврианского на Арктуранский [Knight, 1997]

Перевести: **farok** **crrrok** hihok yorok klok kantok ok-yurp

|   |   |
|---|---|
| 1a. ok-voon ororok sprok .              | 7a. lalok farok ororok lalok sprok izok enemok .    |
| 1b. at-voon bichat dat .                | 7b. wat jjat bichat wat dat vat eneat .             |
| 2a. ok-drubel ok-voon anak plok sprok . | 8a. lalok brok anak plok nok .                      |
| 2b. at-drubel at-voon pippat rrat dat . | 8b. iat lat pippat rrat nnat .                      |
| 3a. erok sprok izok hihok ghrok .       | 9a. wiwok nok izok kantok ok-yurp .                 |
| 3b. totat dat arrat vat hilat .         | 9b. totat nnat quat oloat at-yurp .                 |
| 4a. ok-voon anak drok brok jok .        | 10a. lalok mok nok yorok ghrok klok .               |
| 4b. at-voon krat pippat sat lat .       | 10b. wat nnat gat mat bat hilat .                   |
| 5a. wiwok farok izok stok .             | 11a. lalok nok <b>crrrok</b> hihok yorok zanzanok . |
| 5b. totat jjat quat cat .               | 11b. wat nnat arrat mat zanzanat .                  |
| 6a. lalok sprok izok jok stok .         | 12a. lalok rarok nok izok hihok mok .               |
| 6b. wat dat krat quat cat .             | 12b. wat nnat forat arrat vat gat .                 |

# Перевод на основе параллельных корпусов

Перевод с Центаврианского на Арктуранский [Knight, 1997]

Перевести: **farok** crrrok **hihok** yorok klok kantok ok-yurp

|   |   |
|---|---|
| 1a. ok-voon ororok sprok .                | 7a. lalok farok ororok lalok sprok izok enemok .    |
| 1b. at-voon bichat dat .                  | 7b. wat jjat bichat wat dat vat eneat .             |
| 2a. ok-drubel ok-voon anak plok sprok .   | 8a. lalok brok anak plok nok .                      |
| 2b. at-drubel at-voon pippat rrat dat .   | 8b. iat lat pippat rrat nnat .                      |
| 3a. erok sprok izok <b>hihok</b> ghirok . | 9a. wiwok nok izok kantok ok-yurp .                 |
| 3b. totat dat <b>arrat</b> vat hilat .    | 9b. totat nnat quat oloat at-yurp .                 |
| 4a. ok-voon anak drok brok jok .          | 10a. lalok mok nok yorok ghirok klok .              |
| 4b. at-voon krat pippat sat lat .         | 10b. wat nnat gat mat bat hilat .                   |
| 5a. wiwok farok izok stok .               | 11a. lalok nok crrrok <b>hihok</b> yorok zanzanok . |
| 5b. totat jjat quat cat .                 | 11b. wat nnat <b>arrat</b> mat zanzanat .           |
| 6a. lalok sprok izok jok stok .           | 12a. lalok rarok nok izok <b>hihok</b> mok .        |
| 6b. wat dat krat quat cat .               | 12b. wat nnat forat <b>arrat</b> vat gat .          |

# Перевод на основе параллельных корпусов

Перевод с Центаврианского на Арктуранский [Knight, 1997]

Перевести: **farok** crrrok **hihok** **yorok** klok kantok ok-yurp

|   |  |
|---|--|
| 1a. ok-voon ororok sprok .              | 7a. lalok farok ororok lalok sprok izok enemok . |
| 1b. at-voon bichat dat .                | 7b. wat jjat bichat wat dat vat eneat .          |
| 2a. ok-drubel ok-voon anak plok sprok . | 8a. lalok brok anak plok nok .                   |
| 2b. at-drubel at-voon pippat rrat dat . | 8b. iat lat pippat rrat nnat .                   |
| 3a. erok sprok izok hihok ghrok .       | 9a. wiwok nok izok kantok ok-yurp .              |
| 3b. totat dat arrat vat hilat .         | 9b. totat nnat quat oloat at-yurp .              |
| 4a. ok-voon anak drok brok jok .        | 10a. lalok mok nok yorok ghrok klok .            |
| 4b. at-voon krat pippat sat lat .       | 10b. wat nnat gat mat bat hilat .                |
| 5a. wiwok farok izok stok .             | 11a. lalok nok crrrok hihok yorok zanzanok .     |
| 5b. totat jjat quat cat .               | 11b. wat nnat arrat mat zanzanat .               |
| 6a. lalok sprok izok jok stok .         | 12a. lalok rarok nok izok hihok mok .            |
| 6b. wat dat krat quat cat .             | 12b. wat nnat forat arrat vat gat .              |

# Перевод на основе параллельных корпусов

Перевод с Центаврианского на Арктуранский [Knight, 1997]

Перевести: **farok** crrrok **hihok** yorok **clock** kantok ok-yurp

|   |  |
|---|--|
| 1a. ok-voon ororok sprok .              | 7a. lalok farok ororok lalok sprok izok enemok . |
| 1b. at-voon bichat dat .                | 7b. wat jjat bichat wat dat vat eneat .          |
| 2a. ok-drubel ok-voon anak plok sprok . | 8a. lalok brok anak plok nok .                   |
| 2b. at-drubel at-voon pippat rrat dat . | 8b. iat lat pippat rrat nnat .                   |
| 3a. erok sprok izok hihok ghrok .       | 9a. wiwok nok izok kantok ok-yurp .              |
| 3b. totat dat arrat vat hilat .         | 9b. totat nnat quat oloat at-yurp .              |
| 4a. ok-voon anak drok brok jok .        | 10a. lalok mok nok yorok ghrok <b>clock</b> .    |
| 4b. at-voon krat pippat sat lat .       | 10b. wat nnat gat mat bat hilat .                |
| 5a. wiwok farok izok stok .             | 11a. lalok nok crrrok hihok yorok zanzanok .     |
| 5b. totat jjat quat cat .               | 11b. wat nnat arrat mat zanzanat .               |
| 6a. lalok sprok izok jok stok .         | 12a. lalok rarok nok izok hihok mok .            |
| 6b. wat dat krat quat cat .             | 12b. wat nnat forat arrat vat gat .              |

# Перевод на основе параллельных корпусов

Перевод с Центаврианского на Арктуранский [Knight, 1997]

Перевести: **farok** crrrok **hihok** yorok **clock** kantok ok-yurp

|   |  |
|---|--|
| 1a. ok-voon ororok sprok .              | 7a. lalok farok ororok lalok sprok izok enemok . |
| 1b. at-voon bichat dat .                | 7b. wat jjat bichat wat dat vat eneat .          |
| 2a. ok-drubel ok-voon anak plok sprok . | 8a. lalok brok anak plok nok .                   |
| 2b. at-drubel at-voon pippat rrat dat . | 8b. iat lat pippat rrat nnat .                   |
| 3a. erok sprok izok hihok ghrok .       | 9a. wiwok nok izok kantok ok-yurp .              |
| 3b. totat dat arrat vat hilat .         | 9b. totat nnat quat oloat at-yurp .              |
| 4a. ok-voon anak drok brok jok .        | 10a. lalok mok nok yorok ghrok <b>clock</b> .    |
| 4b. at-voon krat pippat sat lat .       | 10b. wat nnat gat mat bat hilat .                |
| 5a. wiwok farok izok stok .             | 11a. lalok nok crrrok hihok yorok zanzanok .     |
| 5b. totat jjat quat cat .               | 11b. wat nnat arrat mat zanzanat .               |
| 6a. lalok sprok izok jok stok .         | 12a. lalok rarok nok izok hihok mok .            |
| 6b. wat dat krat quat cat .             | 12b. wat nnat forat arrat vat gat .              |

# Перевод на основе параллельных корпусов

Перевод с Центаврианского на Арктуранский [Knight, 1997]

Перевести: **farok** crrrok **hihok** **yorok** **clock** kantok ok-yurp

|   |  |
|---|--|
| 1a. ok-voon ororok sprok .              | 7a. lalok farok ororok lalok sprok izok enemok . |
| 1b. at-voon bichat dat .                | 7b. wat jjat bichat wat dat vat eneat .          |
| 2a. ok-drubel ok-voon anak plok sprok . | 8a. lalok brok anak plok nok .                   |
| 2b. at-drubel at-voon pippat rrat dat . | 8b. iat lat pippat rrat nnat .                   |
| 3a. erok sprok izok hihok ghirok .      | 9a. wiwok nok izok kantok ok-yurp .              |
| 3b. totat dat arrat vat hilat .         | 9b. totat nnat quat oloat at-yurp .              |
| 4a. ok-voon anak drok brok jok .        | 10a. lalok mok nok yorok ghirok clock .          |
| 4b. at-voon krat pippat sat lat .       | 10b. wat nnat gat mat bat hilat .                |
| 5a. wiwok farok izok stok .             | 11a. lalok nok crrrok hihok yorok zanzanok .     |
| 5b. totat jjat quat cat .               | 11b. wat nnat arrat mat zanzanat .               |
| 6a. lalok sprok izok jok stok .         | 12a. lalok rarok nok izok hihok mok .            |
| 6b. wat dat krat quat cat .             | 12b. wat nnat forat arrat vat gat .              |

# Перевод на основе параллельных корпусов

Перевод с Центаврианского на Арктуранский [Knight, 1997]

Перевести: **farok** crrrok **hihok** **yorok** **clock** kantok ok-yurp

|   |  |
|---|--|
| 1a. ok-voon ororok sprok .              | 7a. lalok farok ororok lalok sprok izok enemok . |
| 1b. at-voon bichat dat .                | 7b. wat jjat bichat wat dat vat eneat .          |
| 2a. ok-drubel ok-voon anak plok sprok . | 8a. lalok brok anak plok nok .                   |
| 2b. at-drubel at-voon pippat rrat dat . | 8b. iat lat pippat rrat nnat .                   |
| 3a. erok sprok izok hihok ghirok .      | 9a. wiwok nok izok kantok ok-yurp .              |
| 3b. totat dat arrat vat hilat .         | 9b. totat nnat quat oloat at-yurp .              |
| 4a. ok-voon anak drok brok jok .        | 10a. lalok mok nok yorok ghirok <b>clock</b> .   |
| 4b. at-voon krat pippat sat lat .       | 10b. wat nnat gat mat bat hilat .                |
| 5a. wiwok farok izok stok .             | 11a. lalok nok crrrok hihok yorok zanzanok .     |
| 5b. totat jjat quat cat .               | 11b. wat nnat arrat mat zanzanat .               |
| 6a. lalok sprok izok jok stok .         | 12a. lalok rarok nok izok hihok mok .            |
| 6b. wat dat krat quat cat .             | 12b. wat nnat forat arrat vat gat .              |

МЕТОДОМ  
ИСКЛЮЧЕНИЯ



# Перевод на основе параллельных корпусов

Перевод с Центаврианского на Арктуранский [Knight, 1997]

Перевести: **farok** **crrrok** **hihok** **yorok** **clock** **kantok** **ok-yurp**

|   |  |
|---|--|
| 1a. ok-voon ororok sprok .              | 7a. lalok farok ororok lalok sprok izok enemok .           |
| 1b. at-voon bichat dat .                | 7b. wat jjat bichat wat dat vat eneat .                    |
| 2a. ok-drubel ok-voon anak plok sprok . | 8a. lalok brok anak plok nok .                             |
| 2b. at-drubel at-voon pippat rrat dat . | 8b. iat lat pippat rrat nnat .                             |
| 3a. erok sprok izok hihok ghrok .       | 9a. wiwok nok izok kantok ok-yurp .                        |
| 3b. totat dat arrat vat hilat .         | 9b. totat nnat quat oloat at-yurp .                        |
| 4a. ok-voon anak drok brok jok .        | 10a. lalok mok nok yorok ghrok clock .                     |
| 4b. at-voon krat pippat sat lat .       | 10b. wat nnat gat mat bat hilat .                          |
| 5a. wiwok farok izok stok .             | 11a. lalok nok <b>crrrok</b> hihok yorok <b>zanzanok</b> . |
| 5b. totat jjat quat cat .               | 11b. wat nnat arrat mat <b>zanzanat</b> .                  |
| 6a. lalok sprok izok jok stok .         | 12a. lalok rarok nok izok hihok mok .                      |
| 6b. wat dat krat quat cat .             | 12b. wat nnat forat arrat vat gat .                        |

Родственное  
слово?

# Перевод на основе параллельных корпусов

Перевод с Центаврианского на Арктуранский [Knight, 1997]

Задание: упорядочить: {jjat, arrat, mat, bat, oloat, at-yurp}

|   |  |
|---|--|
| 1a. ok-voon ororok sprok .              | 7a. lalok farok ororok lalok sprok izok enemok . |
| 1b. at-voon bichat dat .                | 7b. wat jjat bichat wat dat vat eneat .          |
| 2a. ok-drubel ok-voon anak plok sprok . | 8a. lalok brok anak plok nok .                   |
| 2b. at-drubel at-voon pippat rrat dat . | 8b. iat lat pippat rrat nnat .                   |
| 3a. erok sprok izok hihok ghrok .       | 9a. wiwok nok izok kantok ok-yurp .              |
| 3b. totat dat arrat vat hilat .         | 9b. totat nnat quat oloat at-yurp .              |
| 4a. ok-voon anak drok brok jok .        | 10a. lalok mok nok yorok ghrok klok .            |
| 4b. at-voon krat pippat sat lat .       | 10b. wat nnat gat mat bat hilat .                |
| 5a. wiwok farok izok stok .             | 11a. lalok nok crrrok hihok yorok zanzanok .     |
| 5b. totat jjat quat cat .               | 11b. wat nnat arrat mat zanzanat .               |
| 6a. lalok sprok izok jok stok .         | 12a. lalok rarok nok izok hihok mok .            |
| 6b. wat dat krat quat cat .             | 12b. wat nnat forat arrat vat gat .              |

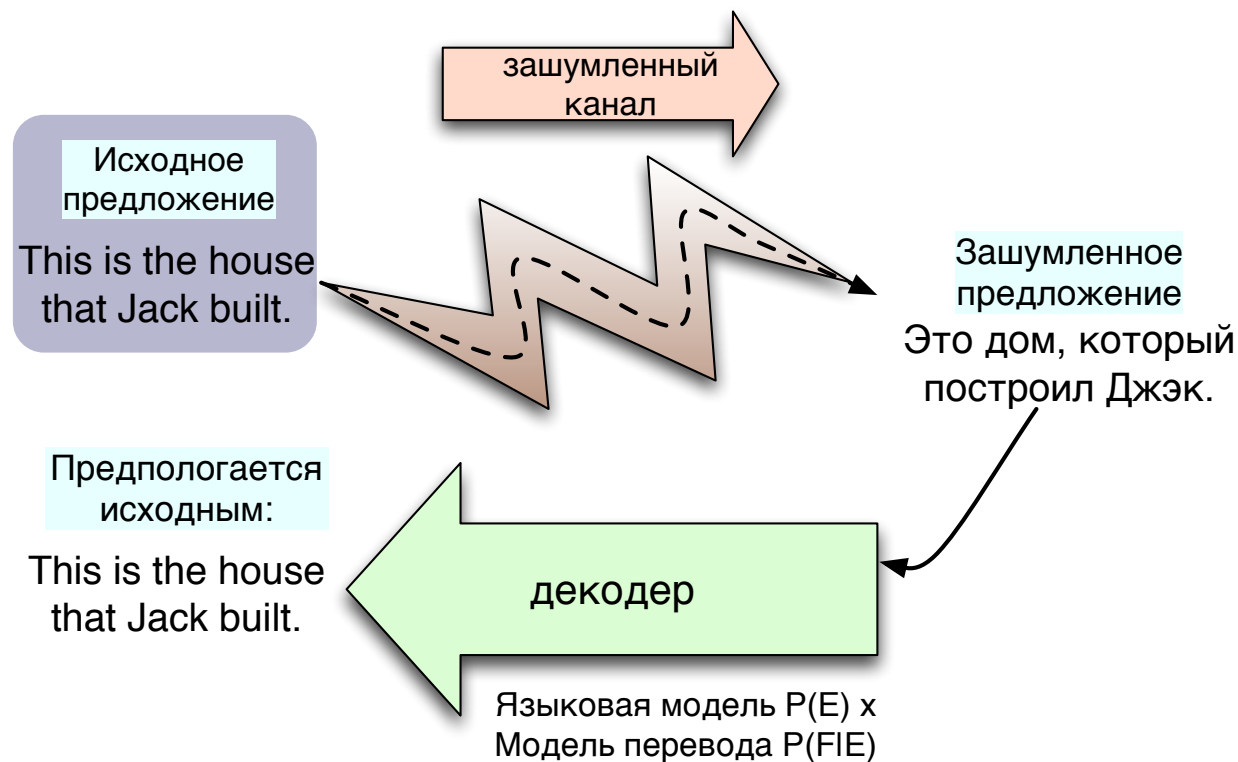
# Перевод на основе параллельных корпусов

В действительности это англо-испанский перевод

**Clients do not sell pharmaceuticals in Europe => Clientes no venden medicinas en Europa**

|  |   |
|--|---|
| 1a. Garcia and associates .<br>1b. Garcia y asociados .                              | 7a. the clients and the associates are enemies .<br>7b. los clients y los asociados son enemigos .          |
| 2a. Carlos Garcia has three associates .<br>2b. Carlos Garcia tiene tres asociados . | 8a. the company has three groups .<br>8b. la empresa tiene tres grupos .                                    |
| 3a. his associates are not strong .<br>3b. sus asociados no son fuertes .            | 9a. its groups are in Europe .<br>9b. sus grupos estan en Europa .  |
| 4a. Garcia has a company also .<br>4b. Garcia tambien tiene una empresa .            | 10a. the modern groups sell strong pharmaceuticals .<br>10b. los grupos modernos venden medicinas fuertes . |
| 5a. its clients are angry .<br>5b. sus clientes estan enfadados .                    | 11a. the groups do not sell zenzanine .<br>11b. los grupos no venden zanzanina .                            |
| 6a. the associates are also angry .<br>6b. los asociados tambien estan enfadados .   | 12a. the small groups are not modern .<br>12b. los grupos pequenos no son modernos .                        |

# Модель зашумленного канала



- Байесовская модель

$$\hat{E} = \arg \max_{E \in English} P(F|E)P(E)$$

Модель  
перевода

Языковая  
модель

# Машинный перевод

- Языковая модель
  - N-граммы
  - СКС грамматики
  - Нейронная сеть
- Модель перевода
- Декодер

# Модель перевода на основе фраз

- $P(F|E)$
- Разбиваем  $E$  на фразы
- Переводим каждую фразу из  $E$  во фразу на другом языке, запоминая **вероятность перевода**
- Находим наиболее вероятную последовательность фраз  $F$

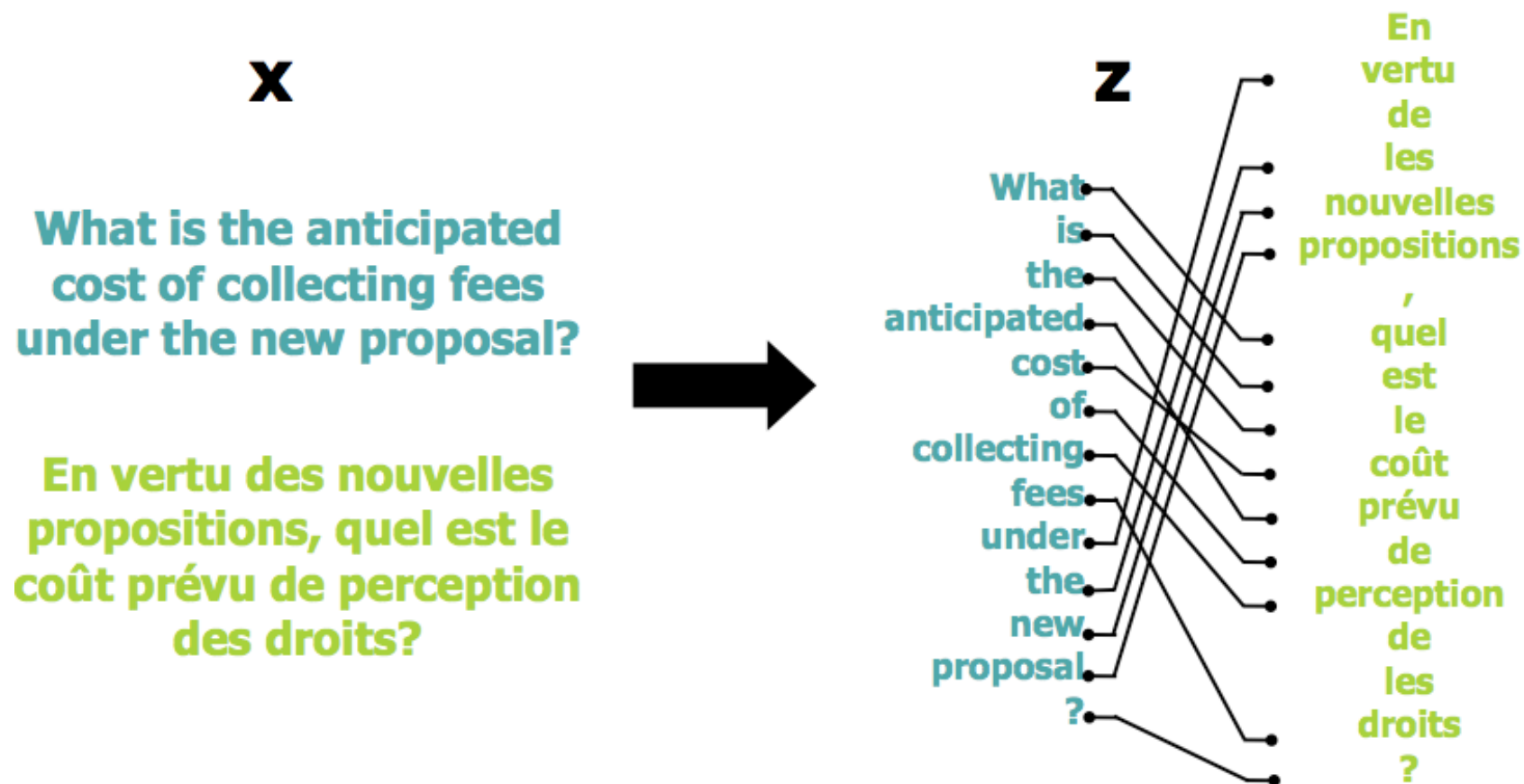
# Вероятность перевода

- Пусть есть параллельные тексты с указанием соответствия между фразами в  $E$  и  $F$  (см. далее).
- Тогда вероятность перевода можно оценить на основе метода максимального правдоподобия

$$\phi(\bar{f}, \bar{e}) = \frac{\text{count}(\bar{f}, \bar{e})}{\sum_{\bar{f}} \text{count}(\bar{f}, \bar{e})}$$

# Выравнивание слов

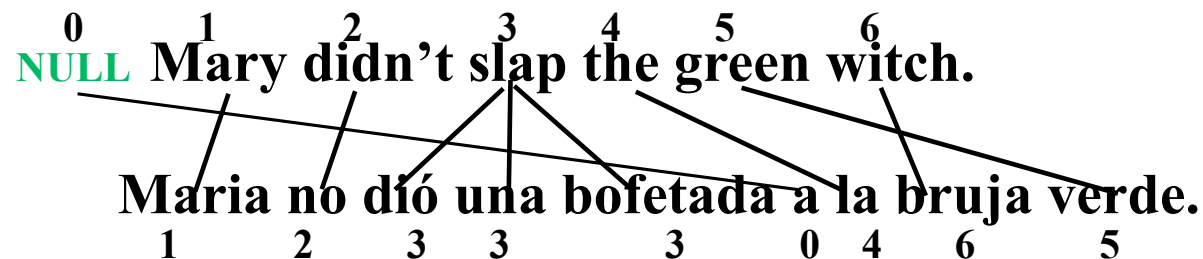
- Сначала выравнивают слова
- Вход: пары предложение-перевод





# Выравнивание один ко многим

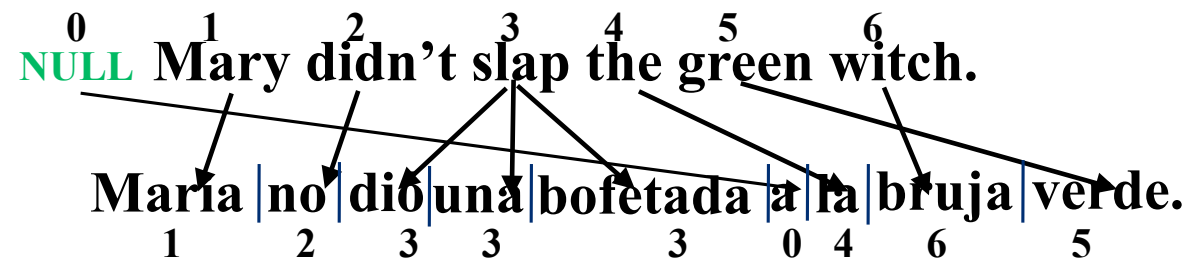
- Для простоты предположим что
  - слово из  $F$  соответствует одному слову в  $E$
  - но слово из  $E$  может соответствовать нескольким словам в  $F$
- Некоторые слова в  $F$  могут соответствовать элементу NULL в  $E$
- Тогда выравнивание можно задавать вектором



# Модель IBM Model 1

- Первая самая простая модель предложенная в основополагающей статье [Brown et. al. 1993]
- Предлагает простую генерирующую модель для получения  $F$  из  $E = e_1, e_2, \dots, e_I$ 
  - Выбрать длину  $J$  предложения  $F = f_1, f_2, \dots, f_J$
  - Выбрать выравнивание  $A = a_1, a_2, \dots, a_J$
  - Получить  $F$  из  $E$

# Пример



# Подсчет $P(F|E)$

- Обозначения

- Длина предложения на языке  $E$  равна  $I$
- Длина предложения на языке  $F$  равна  $J$
- Вероятность длины предложения в  $F$  равна  $P(J|E)$

- Model 1: Предположим что все выравнивания  $A$  равновероятны (их  $(I + 1)^J$ )

- Тогда условная вероятность (выравниваний):

$$P(A | E) = P(A | E, J)P(J | E) = \frac{P(J | E)}{(I + 1)^J}$$

# Подсчет $P(F|E)$

- Пусть  $t(f_x, e_y)$  вероятность перевода слова  $e_y$  в слово  $f_x$
- Определим  **$P(F|E)$**

$$P(F | E) = \sum_A P(F | E, A) P(A | E) = \sum_A \frac{P(J | E)}{(I + 1)^J} \prod_{j=1}^J t(f_j, e_{a_j})$$

# Как получить выравнивание?

- Цель: найти наиболее вероятное выравнивание

$$\begin{aligned}\hat{A} &= \operatorname{argmax}_A P(F, A | E) \\ &= \operatorname{argmax}_A \frac{P(J | E)}{(I + 1)^J} \prod_{j=1}^J t(f_j, e_{a_j}) \\ &= \operatorname{argmax}_A \prod_{j=1}^J t(f_j, e_{a_j})\end{aligned}$$

- Так как различные переводы для каждой позиции  $j$  независимы, максимум произведения достигается при максимуме каждого термина

$$a_j = \operatorname{argmax}_{0 \leq i \leq I} t(f_j, e_i) \quad 1 \leq j \leq J$$

# ЕМ-алгоритм

- Выбираем начальные параметры
- Пока не сойдется выполняем:
  - Е-шаг: Вычисляем вероятность всех выравниваний с помощью текущей модели
  - М-шаг: Используем эти вероятности для переоценки значений всех параметров модели

Для сокращения времени работы используется алгоритм динамического программирования

# Пример

**Тренировочный  
корпус**

green house  
casa verde

the house  
la casa

**Вероятности  
перевода**

|       | verde | casa | la  |
|-------|-------|------|-----|
| green | 1/3   | 1/3  | 1/3 |
| house | 1/3   | 1/3  | 1/3 |
| the   | 1/3   | 1/3  | 1/3 |

**Предполагаем начальные  
вероятности равными**

**Вычисляем  
вероятности  
выравнивания  
 $P(A, F | E)$**

green house  
casa verde

$$1/3 \times 1/3 = 1/9$$

~~green house~~  
casa verde

$$1/3 \times 1/3 = 1/9$$

the house  
la casa

$$1/3 \times 1/3 = 1/9$$

~~the house~~  
la casa

$$1/3 \times 1/3 = 1/9$$

**Нормализуем  
 $P(A | F, E)$**

$$\frac{1/9}{2/9} = \frac{1}{2}$$

$$\frac{1/9}{2/9} = \frac{1}{2}$$

$$\frac{1/9}{2/9} = \frac{1}{2}$$

$$\frac{1/9}{2/9} = \frac{1}{2}$$

$$P(A|E, F) = \frac{P(A, F|E)}{\sum_A P(A, F|E)}$$



# Пример

green house  
casa verde  
1/2

~~green house~~  
~~casa verde~~  
1/2

the house  
la casa  
1/2

~~the house~~  
~~la casa~~  
1/2

Считаем  
веса  
переводов

|       | verde | casa      | la  |
|-------|-------|-----------|-----|
| green | 1/2   | 1/2       | 0   |
| house | 1/2   | 1/2 + 1/2 | 1/2 |
| the   | 0     | 1/2       | 1/2 |

Нормализуем  
и получаем  
 $P(f | e)$

|       | verde | casa | la  |
|-------|-------|------|-----|
| green | 1/2   | 1/2  | 0   |
| house | 1/4   | 1/2  | 1/4 |
| the   | 0     | 1/2  | 1/2 |

# Пример

Вероятности  
перевода

|       | verde | casa | la  |
|-------|-------|------|-----|
| green | 1/2   | 1/2  | 0   |
| house | 1/4   | 1/2  | 1/4 |
| the   | 0     | 1/2  | 1/2 |

Пересчитываем  
вероятности  
выравнивания  
 $P(A, F | E)$

|                        |                        |                        |                        |
|------------------------|------------------------|------------------------|------------------------|
| green house            | <del>green house</del> | <del>the house</del>   | the house              |
| casa verde             | casa verde             | la casa                | la casa                |
| $1/2 \times 1/4 = 1/8$ | $1/2 \times 1/2 = 1/4$ | $1/2 \times 1/2 = 1/4$ | $1/2 \times 1/4 = 1/8$ |

$$P(A, F | E) = \prod_{j=1}^J t(f_j | e_{a_j})$$

Нормализуем  
и получаем  
 $P(A | F, E)$

$$\frac{1/8}{3/8} = \frac{1}{3} \quad \frac{1/4}{3/8} = \frac{2}{3} \quad \frac{1/4}{3/8} = \frac{2}{3} \quad \frac{1/8}{3/8} = \frac{1}{3}$$

Продолжаем алгоритм до сходимости  
или ограниченное число шагов

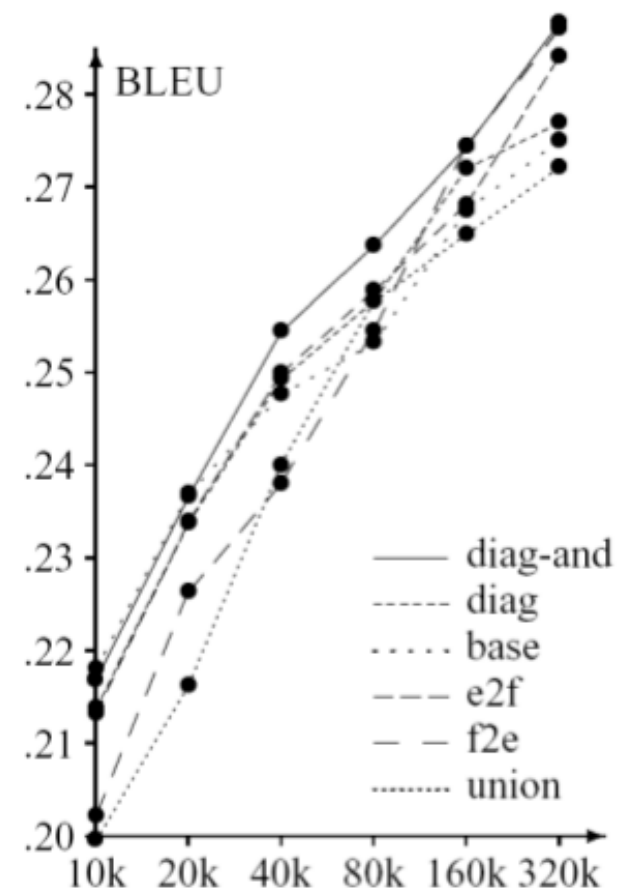
# Выравнивание фраз

- Мы обсудили как выравнивать слова и переводить текст по словам
- Теперь перейдем к фразам



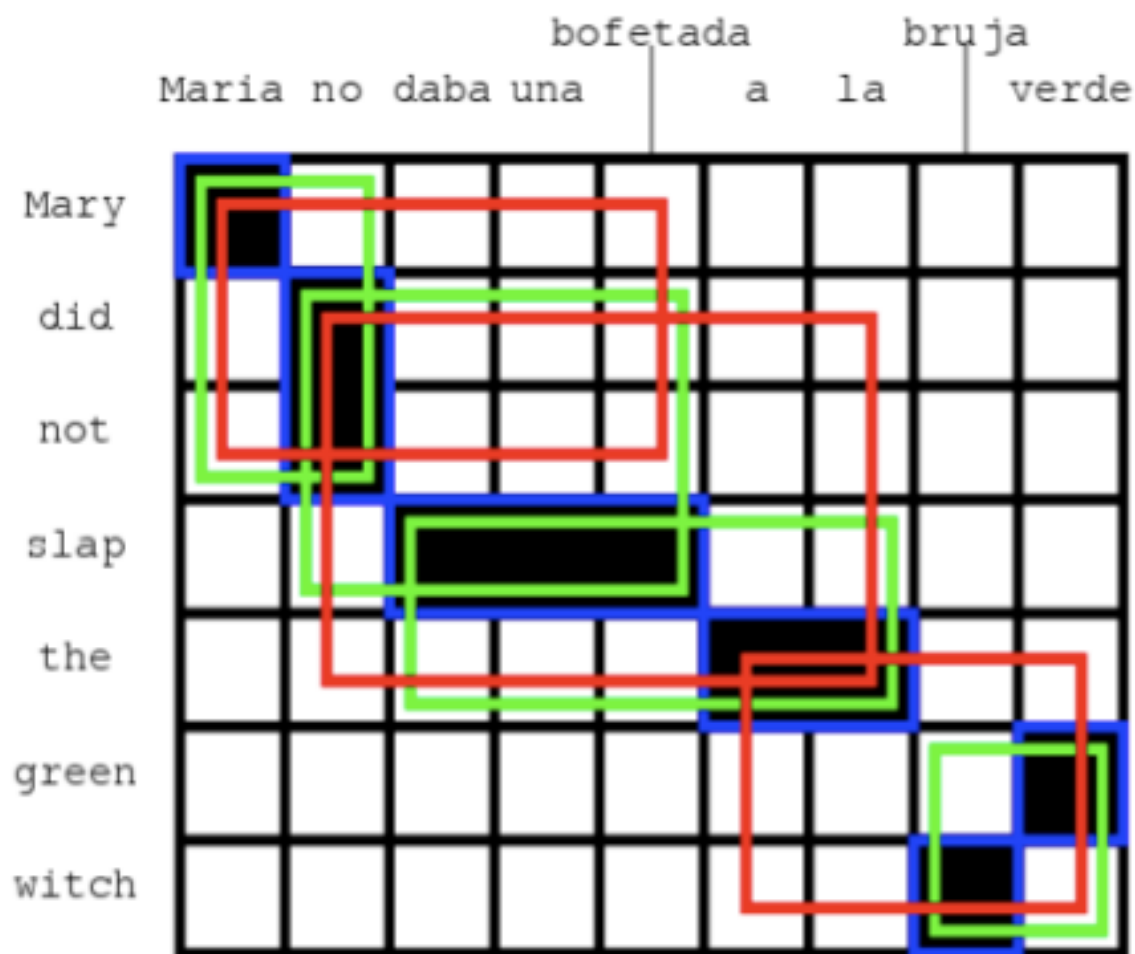
# Выравнивание фраз

- Существует несколько эвристических методов **выравнивания фраз** по матрице пересечений



# Извлечение фраз

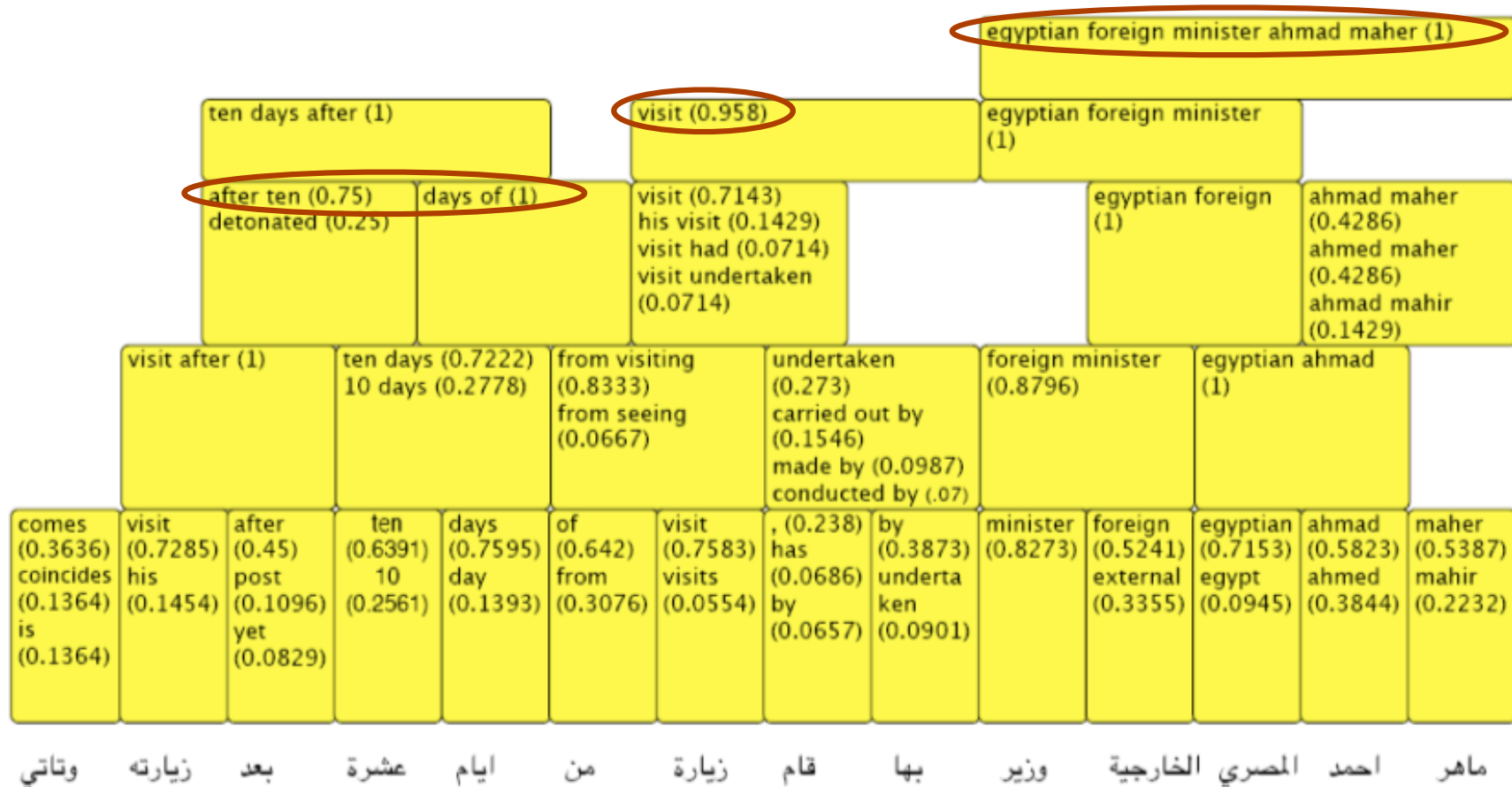
Google (Slav Petrov, SYRCoDIS'11):



Выбираем все консистентные выравнивания

# Декодирование

- Аналог Витерби: выбрать предложение е максимизирующее  $P(e) \times P(f | e)$



# Оценка моделей

- Оценка людьми
  - плавность
  - достоверность
    - адекватность (по фиксированной шкале)
    - информативность (ответ на вопрос по переводу)
- Автоматическая оценка
  - сравнение с одним или несколькими экспертными переводами
  - Меры качества
    - BLEU (Bilingual evaluation understudy)
    - NIST
    - TER
    - METEOR



# Оценка моделей: BLEU

- Определить число N-грамм из машинного перевода в экспертных переводах
- Вычислить модифицированную меру точности

# Оценка моделей: BLEU

Cand 1: **Mary** **no** **slap** **the** **witch** **green**

Cand 2: Mary did not give a smack to a green witch.

Ref 1: **Mary** did not **slap** **the** **green** **witch.**

Ref 2: **Mary** did not smack **the** **green** **witch.**

Ref 3: **Mary** did not hit a **green** sorceress.

Cand 1 точность по 1-граммам: 5/6

# Оценка моделей: BLEU

Cand 1: Mary no slap the witch green.

Cand 2: Mary did not give a smack to a green witch.

Ref 1: Mary did not slap the green witch.

Ref 2: Mary did not smack the green witch.

Ref 3: Mary did not hit a green sorceress.

Cand 1 точность по 2-граммам: 1/5

# Оценка моделей: BLEU

Cand 1: Mary no slap the witch green.

Cand 2: Mary did not give a smack to a green witch.

Ref 1: Mary did not slap the green witch.

Ref 2: Mary did not smack the green witch.

Ref 3: Mary did not hit a green sorceress.

Для каждой N-граммы счетчик не должен превышать максимального количества этой n-граммы в любом предложении

Cand 2 точность 1-грамм: 7/10

# Оценка моделей: BLEU

Cand 1: Mary no slap the witch green.

Cand 2: Mary did not give a smack to a green witch.

Ref 1: Mary did not slap the green witch.

Ref 2: Mary did not smack the green witch.

Ref 3: Mary did not hit a green sorceress.

Cand 2 точность 2-грамм: 4/9

# Модифицированная точность

- Среднее геометрическое всех N-граммам (обычно  $N < 5$ )

$$p_n = \frac{\sum_{C \in \text{corpus}} \sum_{n\text{-gram} \in C} \text{count}_{\text{clip}}(n\text{-gram})}{\sum_{C \in \text{corpus}} \sum_{n\text{-gram} \in C} \text{count}(n\text{-gram})}$$

$$p = \sqrt[N]{\prod_{n=1}^N p_n}$$

**Cand 1:**  $p = \sqrt[2]{\frac{5}{6} \frac{1}{5}} = 0.408$

**Cand 2:**  $p = \sqrt[2]{\frac{7}{10} \frac{4}{9}} = 0.558$

# Штраф за краткость

- Сложно посчитать полноту (recall) из-за нескольких экспертных мнений
- Вместо этого используется штраф за краткость
- Пусть  $r$  - длина экспертного предложения с наибольшим количеством совпадающих N-грамм. Пусть  $c$  - длина машинного перевода

$$BP = \begin{cases} 1 & \text{if } c > r \\ e^{(1-r/c)} & \text{if } c \leq r \end{cases}$$

# Подсчет BLEU

- В итоге:  $BLEU = BP \times p$

**Cand 1:** Mary no slap the witch green.

**Best Ref:** Mary did not slap the green witch.

$$c = 6, \quad r = 7, \quad BP = e^{(1-7/6)} = 0.846$$

$$BLEU = 0.846 \times 0.408 = 0.345$$

**Cand 2:** Mary did not give a smack to a green witch.

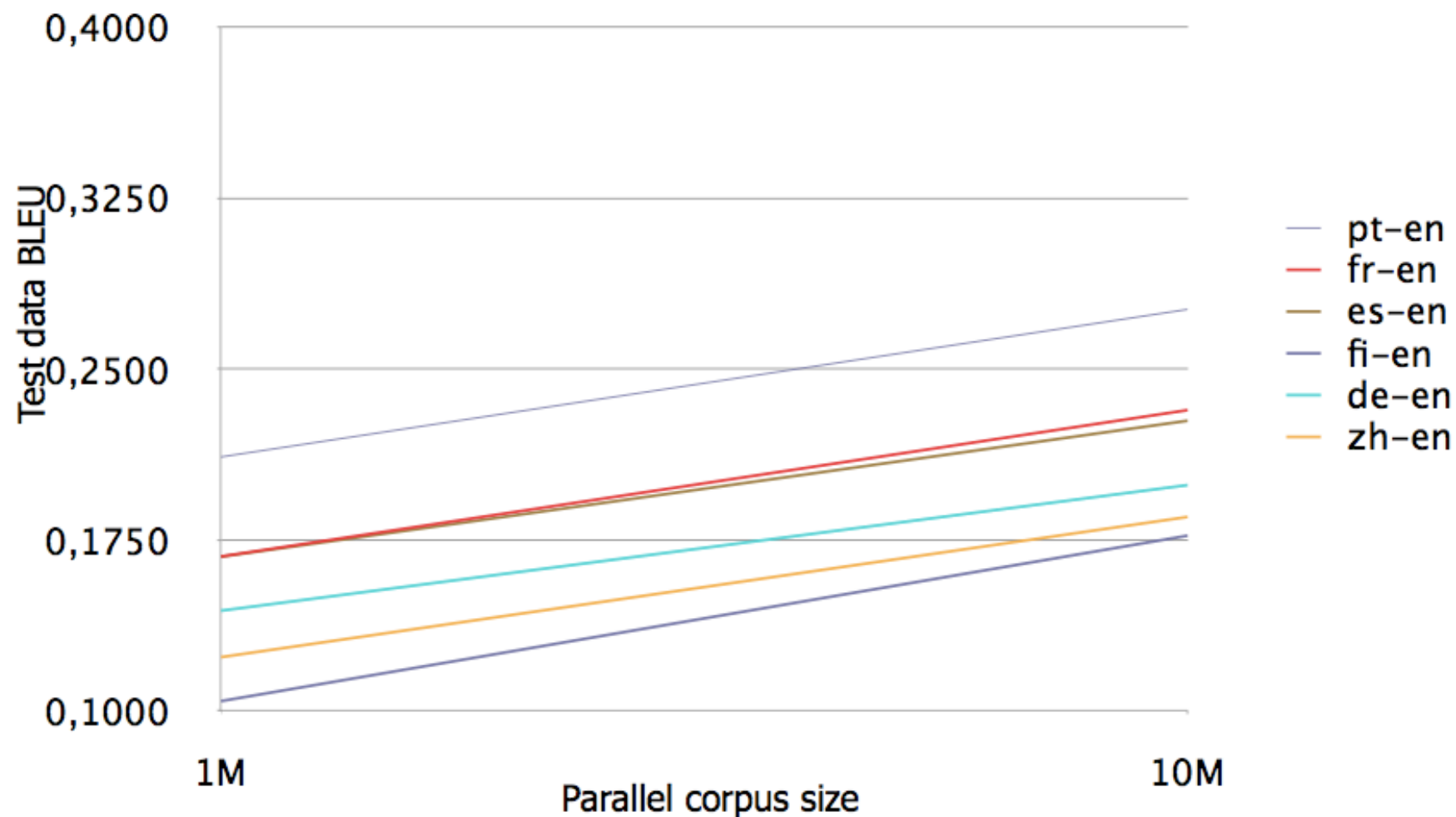
**Best Ref:** Mary did not smack the green witch.

$$c = 10, \quad r = 7, \quad BP = 1$$

$$BLEU = 1 \times 0.558 = 0.558$$

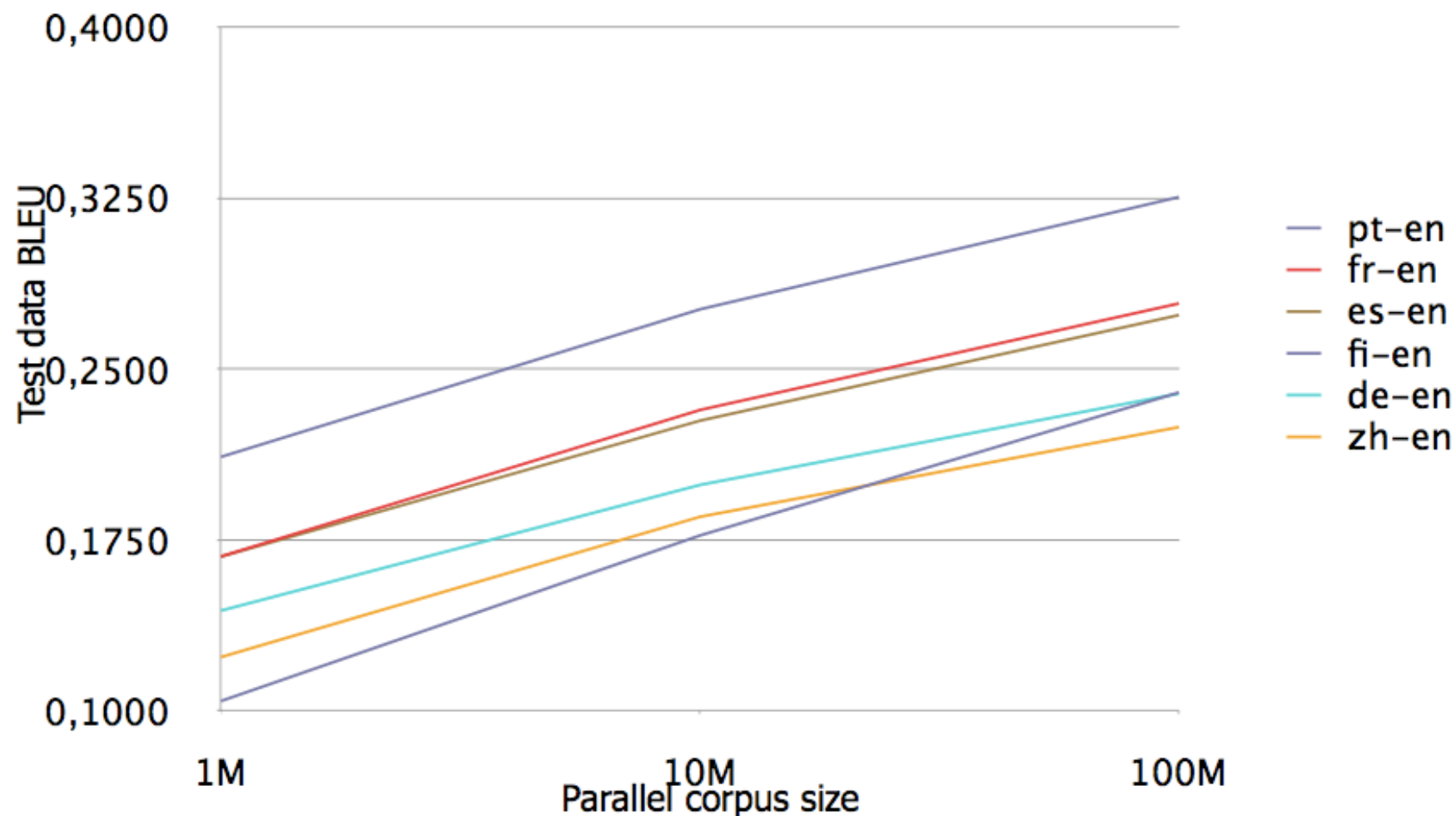


# Лучшие данные - много данных



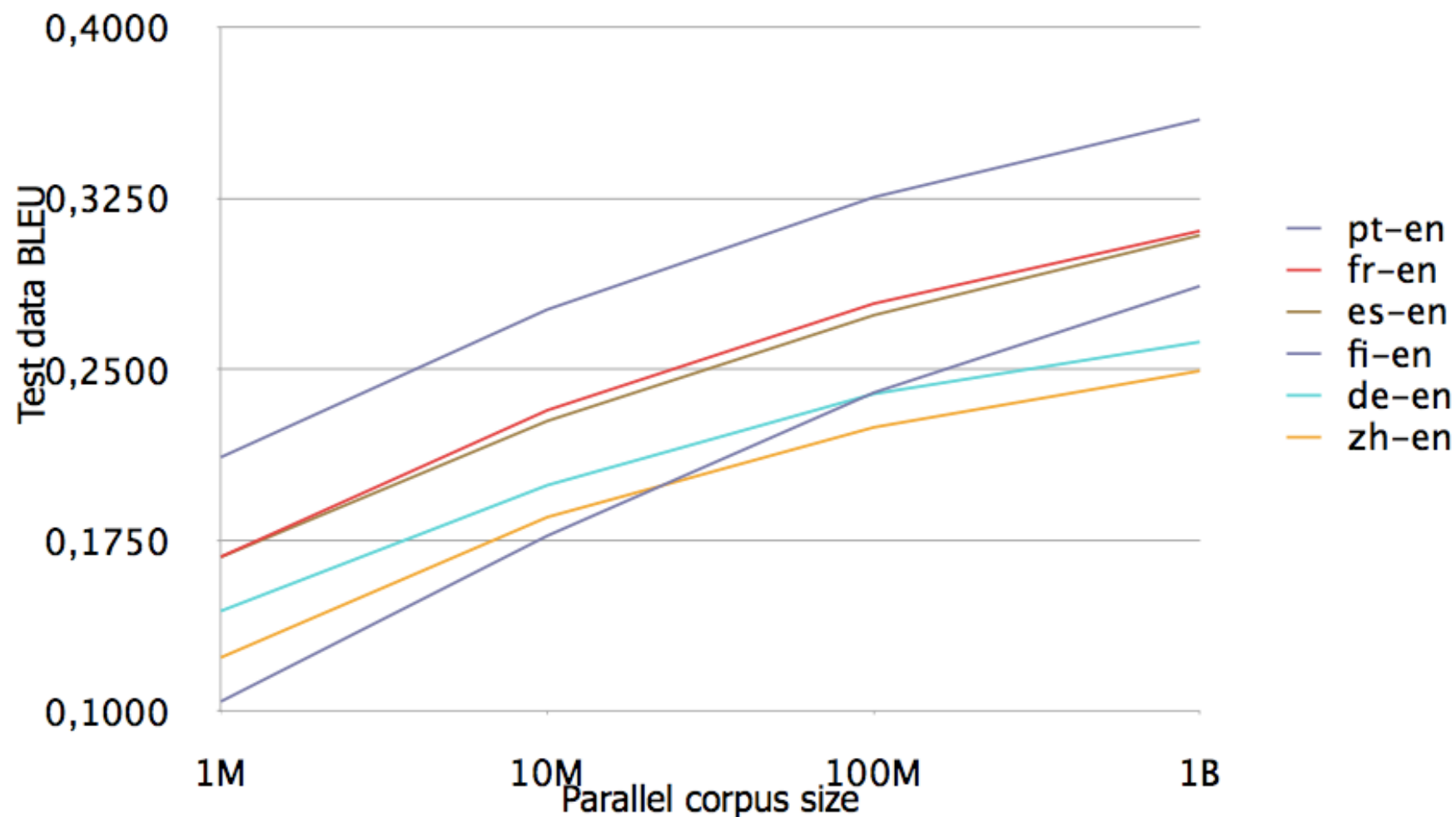
by Google

# Лучшие данные - много данных



by Google

# Лучшие данные - много данных



by Google

# Заключение

- Трудность перевода заключается в существенных **различиях** между языками
- Классические подходы: **прямой перевод, преобразование, интерлингва**
- Для статистического машинного перевода применяется **модель зашумленного канала, комбинирующая модель перевода и языковую модель**
- Для **выравнивания** слов в двуязычных корпусах применяются формальные модели, например, **IBM Model 1**
- Для оценки систем используются различные метрики: **BLEU, TER, METEOR**.

# Следующая лекция

- Языковые модели на основе нейронных сетей