Яндекс

Яндекс

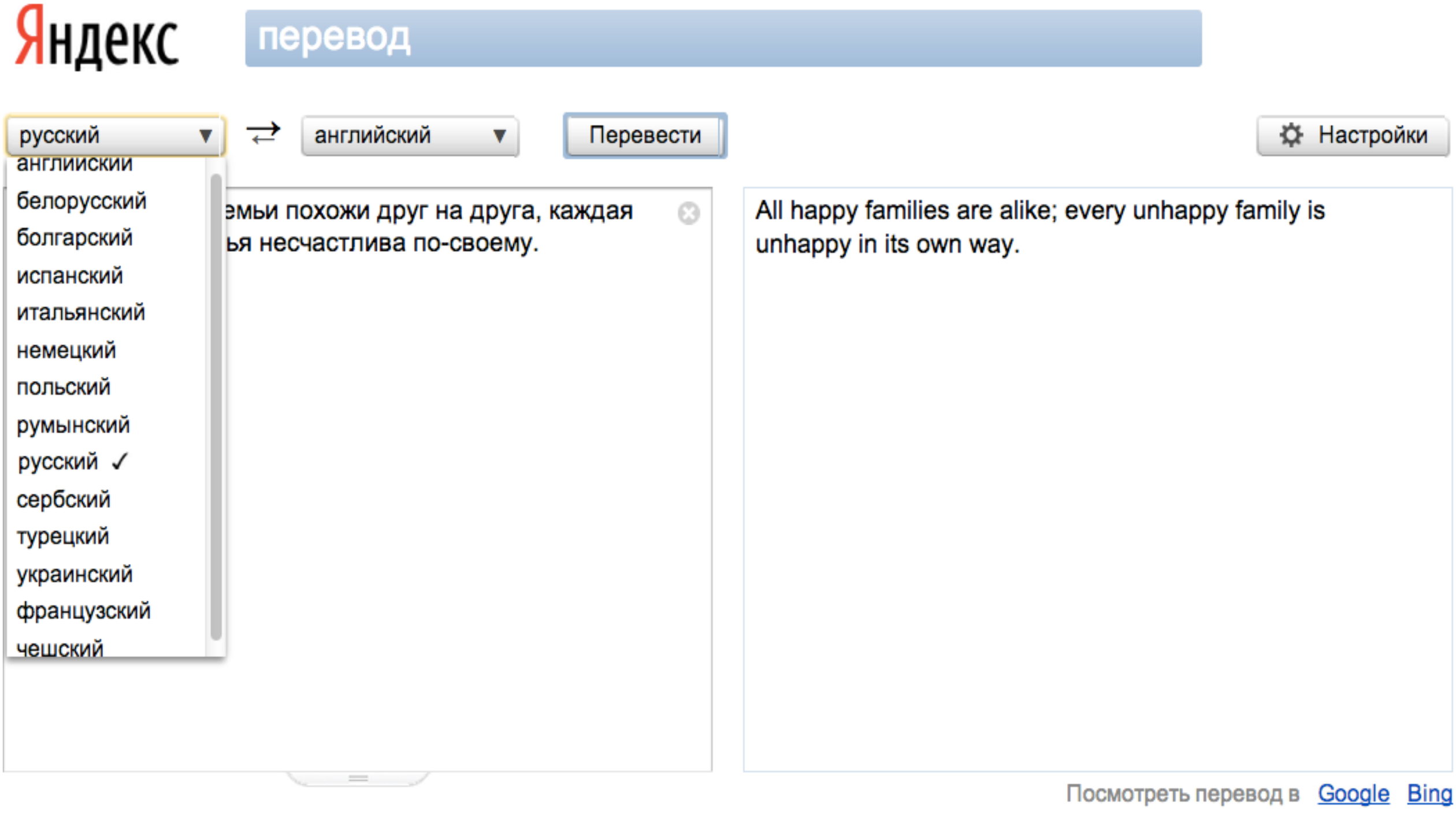
Влияние различных типов орфографических ошибок на качество статистического машинного перевода

Елена Мещерякова, Ирина Галинская,
Валентин Гусев, Мария Шматова

Отдел лингвистических технологий

Яндекс.Перевод

- http://translate.yandex.ru



Предредактирование (pre-editing) входного текста в МТ

- нормализация (normalizing)
- упрощенный язык (controlled/standard language)
- изменение порядка слов (reordering)
- исправление ошибок (correction of errors)

Онлайн-сервисы МП и исправление ошибок

- на перевод поступают тексты различного качества
- много запросов с опечатками, неверной пунктуацией, отсутствием прописных букв и диакритики
- некоторые онлайн-системы предлагают варианты исправления
- такими подсказками трудно воспользоваться, не зная языка, с которого осуществляется перевод

Задачи исследования

- исследовать типы ошибок в запросах к сервису МП
- определить степень влияния различных типов ошибок на перевод
- понять, какие ошибки следует исправлять в первую очередь

Типичные ошибки: опечатки

Russin → Russian; countri → country

Frenh → French; theey → they

mit disem → mit diesem; wihctig → wichtig

Warszwa → Warszawa

Влияние ошибок на перевод: опечатки

Der Vater arbeitete als Buchhalter in einem **Krakenhaus**.

C1 Отец работал бухгалтером **в Krakenhaus**.
C2 Отец работал бухгалтером **в доме осьминога**.
C3 Отец работал бухгалтером **в больнице**.

Der Vater arbeitete als Buchhalter in einem **Krankenhaus**.

C1 Отец работал бухгалтером **в больнице**.
C2 Отец работал бухгалтером **в больнице**.
C3 Отец работал бухгалтером **в больнице**.

Типичные ошибки: диакритика

*ber**u**hmter → ber**ü**hmter*

*A**n**derungen → Ä**n**derungen*

*zo**l**te z**l**oto → **ż**ó**ł**te z**ł**oto*

*to**z**samo**s**ci → to**ż**samo**ś**ci*

Влияние ошибок на перевод: диакритика

Alan Marschall ist ein **berUhmter** australischer Schriftsteller.

C1 Алан Маршал является **berUhmter** австралийский писатель.
C2 Алан Маршалл является **известный** австралийский писатель.
C3 Алан Маршалл — **известный** австралийский писатель.

Alan Marschall ist ein **berühmter** australischer Schriftsteller.

C1 Алан Маршал **известный** австралийский писатель.
C2 Алан Маршалл является **известный** австралийский писатель.
C3 Алан Маршалл — **известный** австралийский писатель.

Типичные ошибки: склейка/разрезание

I saw him yesterday.He said… → I saw him yesterday. He said…

***In**der Stadt → **In der** Stadt*

*nr dok**u m**entu → nr dok**um**entu*

Влияние ошибок на перевод: склейка/разрезание

We can see wild animals at the **z oo**.

C1 Мы можем увидеть диких животных на **z oo**.
C2 Мы можем увидеть диких животных в **Z OO**.
C3 Мы можем увидеть диких животных на **z oo**.

We can see animals at the **zoo**.

C1 Мы можем увидеть диких зверей в **зоопарке**.
C2 Мы можем увидеть диких животных в **зоопарке**.
C3 Мы можем увидеть диких животных в **зоопарке**.

Типичные ошибки: капитализация, пунктуация

i → ***I***; ***m****oscow* → ***M****oscow*

health friends family → *health****,*** *friends****,*** *family*

zwei ***b****rüder* → *zwei* ***B****rüder*

a ja z Swidnicy która jest w ***p****olsce* →

a ja z Swidnicy, która jest w ***P****olsce*

Влияние ошибок на перевод: пунктуация

Do you give up

C1 Ты сдаешься
C2 Вы отказаться от
C3 Вы даете

Do you give up**?**

C1 Ты сдаешься?
C2 Ты сдаешься?
C3 Вы даете?

Влияние ошибок на перевод: капитализация

i was lucky. most people fail.

C1 я был lucky. большинство людей терпят неудачу.
C2 мне повезло. большинство людей терпят неудачу.
C3 я был повезло. Большинство людей не.

I was lucky. Most people fail.

C1 Мне повезло. Большинство людей терпят неудачу.
C2 Мне повезло. Большинство людей не.
C3 Я был повезло. Большинство людей не.

Тестовые наборы: общая характеристика

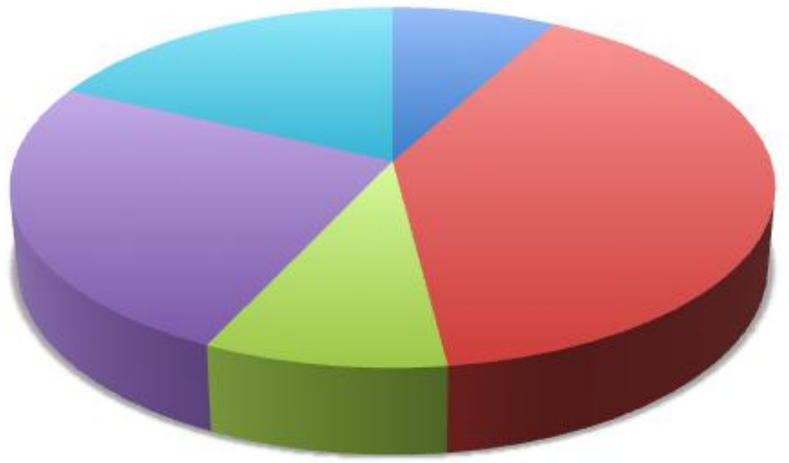
- 500 случайных запросов к сервису Яндекс.Перевод
- направления перевода: **en-ru**, **de-ru**, **pl-ru**
- длина запросов не более 1000 символов

Язык	**Средняя длина запроса (в словах)**
Английский	17
Немецкий	23
Польский	20

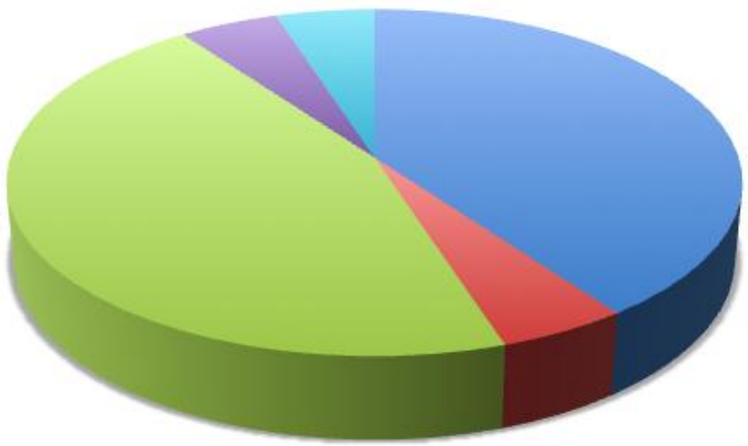
Тестовые наборы: тематика

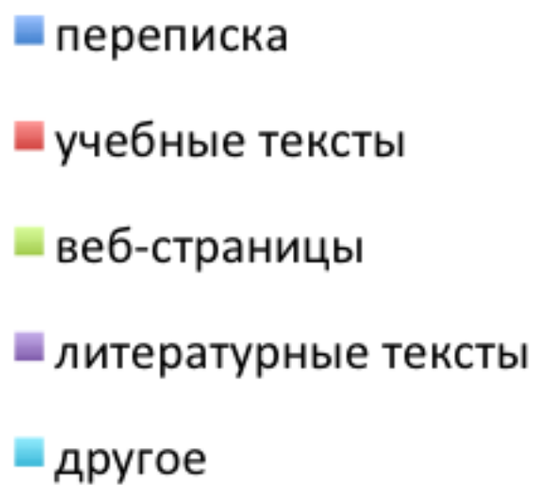
Английский язык

Немецкий язык



Польский язык





Тестовые наборы: ошибки

Язык	Английский	Немецкий	Польский
Диакритика	**-**	**5%**	**12%**
Опечатки	**33%**	**40%**	**37%**
Капитализация + пунктуация	**38%**	**49%**	**62%**
Все ошибки	**53%**	**67%**	**71%**

Эксперимент

- для каждого языка последовательно исправили все типы ошибок

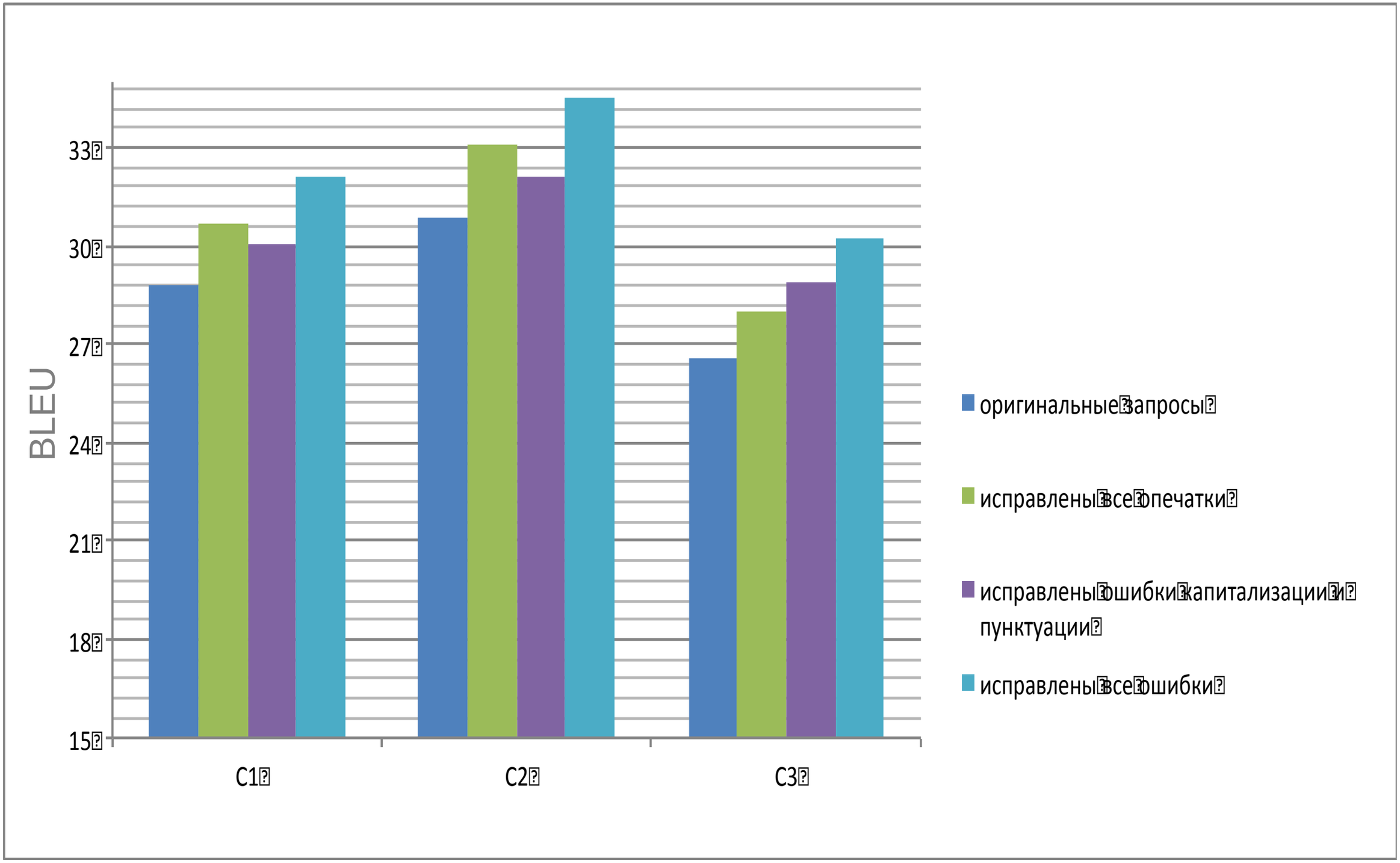
- подготовили эталоны переводов

- тестировались 3 бесплатных онлайн-системы машинного перевода (С1, С2, С3)

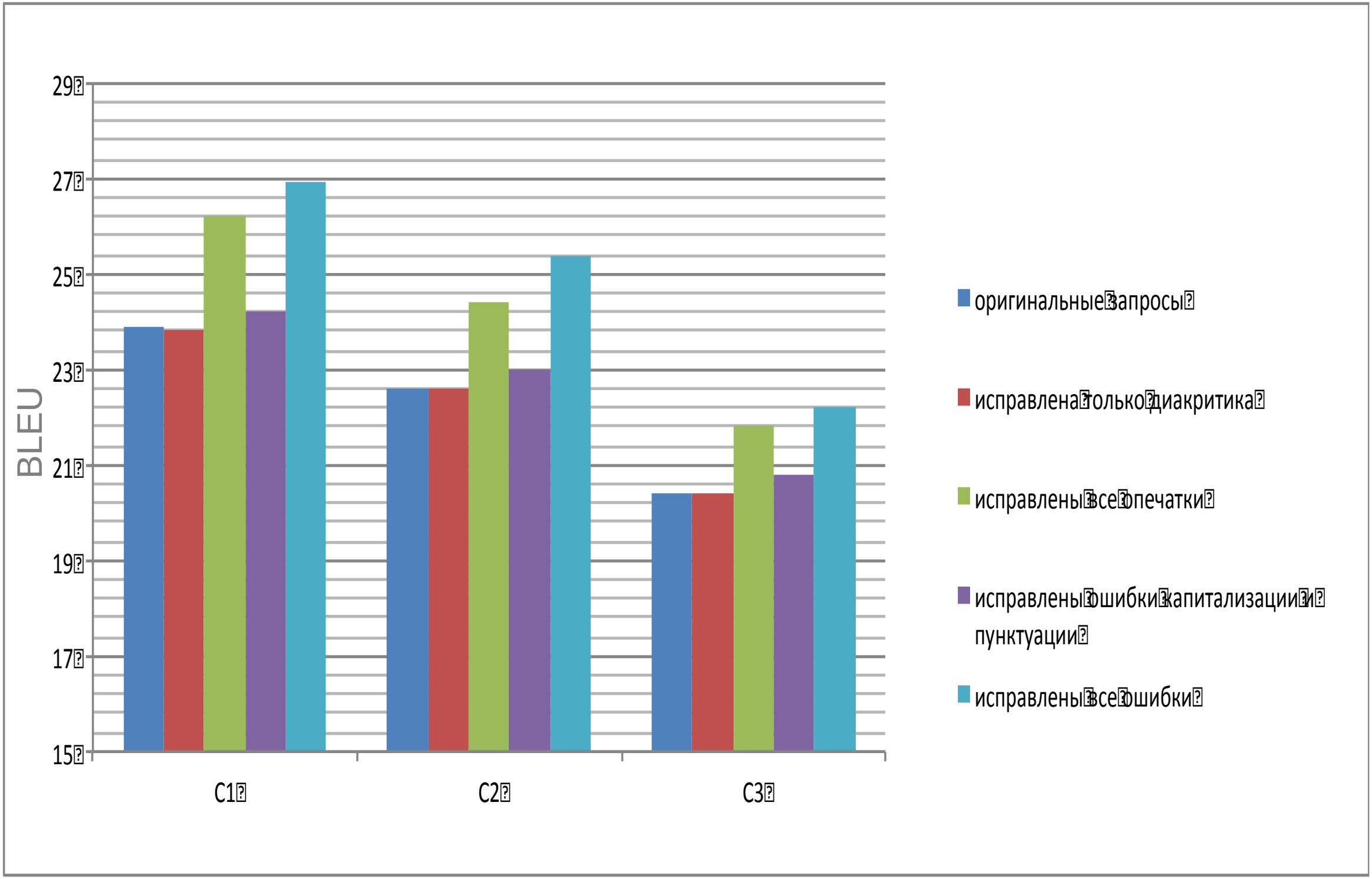
Эксперимент: оценка качества МП

- три варианта автоматической метрики BLEU:
  - с учетом капитализации и пунктуации
  - с учетом только пунктуации
  - без учета капитализации и пунктуации

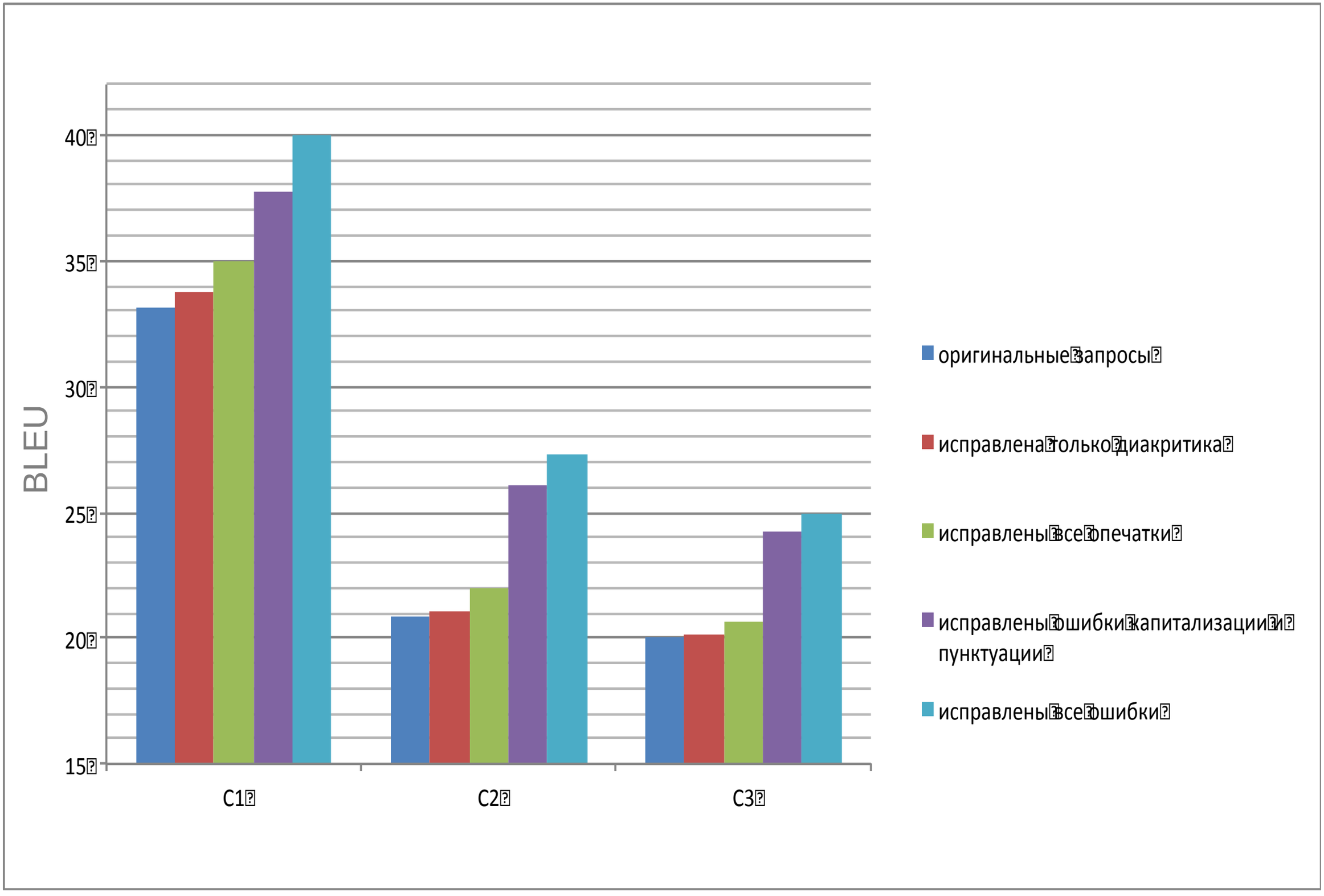
Результаты: английский язык



Результаты: немецкий язык



Результаты: польский язык



Результаты

- исправление опечаток приводит к росту BLEU для всех трех языковых пар
- прирост качества составляет 10-15%
- наибольший вклад вносит исправление опечаток и ошибок капитализации / пунктуации
- диакритика практически не влияет на результаты

Яндекс

Спасибо за внимание!