Международная конференция «Диалог 2009»

Круглый стол «Поисковые системы и коррекция грамотности» (ведущий М.А. Кронгауз)

М.А. КРОНГАУЗ (РГГУ): Я сейчас произнесу для кого-то банальные вещи, а для кого-то нет. Насколько я понял, название проведенного мной круглого стола «Народ против Яндекса» было не всеми понято. Это не значит, что какой-то мифический народ выступает против деятельности Яндекса. Это стандартная форма американской юриспруденции, которая означает возбуждение дела при отсутствии истца, истцом в данном случае является государство. Речь шла о том, что я в качестве народа народ в качестве пользователей инициирую некое дело, если хотите, связанное с поисковыми системами и вообще влиянием компьютерных систем, компьютерных программ на нашу грамотность. Яндекс относительно случайно попал под руку: во-первых, у него есть знаменитый сервис Народ.Ру, поэтому здесь имела место рифма, а с другой стороны, Яндекс является наиболее продвинутым сервисом, самой лучшей поисковой системой — с одной стороны, лучшей, а с другой — именно раздражающей своей продвинутостью. Не тем, что она продвинута, а тем, что она подсказывает лишнее.

Я приведу один пример, не имеющий прямого отношения к Яндексу. У нас в семье происходит постоянная битва, полем которой является домашний компьютер, потому что один из членов семьи требует, чтобы на него была поставлена программа Punto Switcher, а другой категорически против. Punto Switcher автоматически переключает регистр, если считает, что на кириллице или латинице написано бессмысленно, он переводит в несколько осмысленное в другом регистре, то есть делает некоторую работу за пользователя. Вот есть много разных программ, которые работают за нас, в частности спелл-чекеры, которые правят строчные буквы на прописные, которые подчеркивают, то есть менее агрессивно подсказывают, но подчеркивают и обращают наше внимание на то, что нечто надо исправить. Мы сегодня как раз вспоминали старый случай, нашумевший, почти анекдотический, о том, как Word стал подчеркивать слова *жид*, *негр*, *голубой* и *розовый*. Но после некоторого всплеска эмоций в блогосфере это все-таки было ликвидировано. То есть некоторое такое толерантное показание, исходящее от начальства не знаю чего — Майкрософта или чего-то там — было выполнено технически в рамках спелл-чекера. Ну понятно, что это некоторое насилие, потому что запретить нам использовать слово *голубой шарик* или *голубое небо*… Понятно, что это привлечение внимания к слову.

Параллельно я хочу сказать, что обратил внимание (правда, не сейчас, а несколько раньше), что грамотность дипломных работ резко снизилась с переходом на компьютеры и появлением спелл-чекеров. Действительно, происходит активное взаимодействие человека и компьютера, причем компьютер становится все активнее и дает нам советы, выполняет за нас некоторую интеллектуальную работу, чему, в общем, иногда пользователь рад, иногда не рад, — это зависит и от пользователя, и от работы. Ну и — о чем я говорил вчера — есть и третья сторона: лингвисты, и рекомендации, которые нам дают разные программы, часто противоречат рекомендациям лингвистов, если нам удается найти слово в словаре, что бывает редко, и поэтому искать в словаре часто

бесперспективно — проще найти в поисковой системе. И, кроме того, рекомендации разных компьютерных программ часто противоречат друг другу. Например, спелл-чекер дает рекомендации не такие, как поисковые системы.

Вы понимаете, что и вчерашний мой доклад, и сегодняшнее выступление — это выступление со стороны того самого народа, который либо подчиняется, либо бунтует, но все равно не понимает, что с ним происходит. А сегодня здесь специалисты, представители «Яндекса» — той самой наиболее продвинутой системы… я могу это говорить открыто, потому что «Гугл» отказался прийти, за это я награжден бонусом говорить правду. Мы честно звали, вели некоторые смешные переговоры, показавшие, что «Гугл» не пользуется в переписке спелл-чекерами [смех в зале], но это не самое главное.
И теперь здесь представители «Яндекса», я их буду представлять по ходу дела, вот в президиуме два официальных выступающих, и в зале сидят более молодые коллеги — если они захотят выступить, то, естественно, тоже мы представим. Но сначала слово Илье Сегаловичу — от «Яндекса».

И.В. СЕГАЛОВИЧ (Компания «Яндекс»): Я сначала попытаюсь сформулировать некий стратегический кусочек нашего рассказа про исправление опечаток. Понятно, что мы как компания преследуем какую-то цель, что в этом сервисе мы хотим сделать и почему мы это делаем. И Алексей Байтин расскажет, как мы это делаем. Мы пытаемся помочь пользователю сразу найти искомую информацию. Что это значит, я чуть позже скажу. И второе — мы пытаемся действовать по принципу наименьшего изумления. Значит, такой хороший принцип — есть ссылочка в Википедии, где дается описание этого принципа, — он важен, вообще говоря, вот в каком смысле. Малый процент пользователей может очень сильно обидеться на поисковую систему, очень сильно удивиться и уйти. Мы хотим избежать этого сценария. Построена система, с одной стороны, чтобы обеспечить большинству комфортный и быстрый поиск, а с другой стороны, не обидеть меньшинство. Ну и, да, мы уважаем правила русского языка.

Вот чуть-чуть поподробней. Вот интересная тут вещь написана [на презентации], что лишние две секунды для пользователей уменьшают число кликов на 4%. Это исследование не в рамках данной работы делалось, но оно касается именно скорости ответов и недавно было опубликовано, по-моему, то ли Yahoo, то ли Гуглом. Две секунды для человека — очень важная вещь. Время, которое проходит от запроса до первого клика, среднее время, — примерно 20—25 секунд. А если говорить о медиане, то есть о настоящем среднем времени, то есть исключить тех, кто слишком долго размышляет, медитирует над страницей, то эта цифра — в районе 7—8 секунд. То есть 7 секунд человеку достаточно, чтобы что-то такое выхватить из странички и перейти. Поэтому каждые 2 секунды задержки значимы. Поэтому если примерно в 15% запросов люди делают ошибки, мы, игнорируя их нужды, заставляем их еще какие-то действия производить, то мы не выполняем существенную часть работы, которую мы обязаны выполнить, — мы обязаны помочь людям найти информацию. И когда мы это делаем, очевидно, мы чувствуем уверенность, в статистическом смысле: набирая запрос, человек ошибся, и мы должны ему помочь. Ну и в целом, мы стараемся в идеале быстро помочь тем, кто ошибся, а тех, кто намеренно эту штуку написал, мы должны не обидеть, мы должны дать им возможность получить исходный запрос или кусок результата по исходному запросу и для них тоже оставить лазейку какого-то варианта поиска.

Ну и про принцип наименьшего изумления — это конкретно про ту аудиторию, которая на самом деле имела в виду это редкое написание, необычное, которое мы посчитали, что надо исправить. Должно быть ясно, как система исправила запрос, какой результат у человека — исходный и исправленный. То есть это должно быть не спрятано где-то, а должно быть внятно написано в первой строчке над выдачей. И должна быть, понятно, возможность поправить систему, выбрать исходный или измененный вариант. Очень много способов, играясь с тем, как написать, как объяснить человеку, что запрос изменен, как не переборщить в этом месте…

Вот, например, у нас был момент, когда мы показывали исходный неисправленный вариант и говорили: «Вот это у вас такой ужасный исходный неисправленный вариант, а на самом деле мы вам его уже как бы исправили». И формат этого текста был таков, что люди воспринимали его как: «Вы сделали ошибку, а вот правильный вариант». Они туда кликали и получали совсем пустую выдачу, неправильную. А мы как раз имели в виду, что: «Вы могли бы вернуться и получить исходный вариант, но там ничего не найдено». Тем не менее это очень трудно донести — что бы вы ни написали на экране люди… надо угадать их намерения. Слова не важны. Если вы угадали их намерения, то вы получили правильный отклик от пользователя.

И, наконец, специально для слушателей конференции «Диалог» я нарисовал слайд «Мы уважаем правила русского языка». Это означает, что мы используем все словари, до которых смогли дотянуться, и придаем им существенный вес, но в то же время мы не боимся исправлять формально грамматически корректные слова, если нам понятно, что вероятность ошибки в частом слове больше вероятности поиска редкого термина. Вот Алексей расскажет, почему мы в этом так уверены, почему нам это понятно.

М.А. КРОНГАУЗ: Илья сказал сейчас очень важные вещи, и Алексей сейчас подробнее об этом расскажет, что работа поисковой системы направлена прежде всего на благоденствие статистического большинства. Потому что меньшинство может быть (подчеркиваю — не должно, а может быть) более квалифицированно, более аккуратно, не будет опечаток, или знают, что ищут какое-то маргинальное явление, тем не менее, меньшинством быть неудобно. Теперь слово Алексею Байтину. Мне кажется важным сказать, что Алексей до Яндекса делал спелл-чекер для Word’а.

А.В. БАЙТИН (Компания «Яндекс»): Исправляются ошибки, Илья уже рассказал. Я скажу, как мы их исправляем. На Яндексе существуют три крупных блока. Они между собой каким-то образом связаны, но по большому счету это совершенно независимые сервисы. Первый — это самый массовый поиск, «Яндекс.Поиск», на нем мы исправляем опечатки поисковых запросов. В день бывает 50 миллионов запросов на Яндексе, вы можете оценить аудиторию. Второй по важности для Яндекса — набор непоисковых сервисов. Это «Яндекс.Почта» и многие другие, я их перечислять не буду, для вас самым важным, наверно, будет почтовый сервис. И есть такой продукт у Яндекса — бесплатное приложение-расширение браузера, «Яндекс.Бар», который позволяет производить проверку орфографии, исправление ошибок на любой веб-странице. И последний сервич — для самых, скажем так, упорных, которые не удовлетворились первыми двумя, — существуют словари русского языка. Это тоже отдельная служба, на них можно зайти с

главной страницы Яндекса и посмотреть информацию по интересующему слову и, может быть, даже проверить свою грамотность.

Что такое исправление запросов? Можно запрос вообще не исправлять. Поисковая система может просто принять от вас некий запрос и показать выдачу. Известно, что минимум в 12% запросов есть опечатки, значит, можете опять посчитать — из 50 миллионов возьмите 12%, и получите большую цифру. Некоторые поисковые системы до сих пор не исправляют опечаток в запросах, но надо признать, что благодаря улучшению вычислительной техники и развитию искусственного интеллекта за последние 2-3 года произошел большой рывок, качественный сдвиг в области исправления запросов, и сейчас человек уже не должен трястись от страха, что он опять сделал опечатку. Вы понимаете, что большинство наших граждан не очень хорошо владеют русским языком, и в типичном запросе «Туры в Египет» в слове «Египет» может быть 2-3 ошибки. Понятно, что если эти ошибки не исправлять, вы, возможно, найдете туры, но совершенно не те, которые вы искали, и можете нарваться на совершенно обсценные тексты, то есть не то, что ищете. То есть будем считать, что мы переходим на совершенно новый уровень качества обслуживания и пытаемся пользователей не отпускать с самого начала и даже корректировать его грамотность, невзирая на его мнение на этот счет. Так иногда случается, об этих ужасных случаях я тоже отдельно расскажу.

Но сейчас общий случай. Вы начинаете что-то набирать в Яндексе. Допустим, вы набрали «адн». Я не знаю, что вы набирали, но мы-то уже знаем, что вы ищете «одноклассники». И мы их вам подсказываем. Вы можете набрать «одноклассники» с одной «с» или «адноклассники». Скорее всего, мы начнем вам навязывание так называемой бесплатной услуги. Вот, представьте себе, что вы нас не слушаетесь. Следующим шагом скорее всего будет автоматическое исправление вашего запроса. Вот, допустим, вы набирали «упражнение для культуристов» и сделали ошибки и в слове «упражнение», и в слове «культуристов». Мы автоматически исправим за вас этот запрос и сразу покажем выдачу на правильное написание. В некоторых случаях мы это делаем, не будучи стопроцентно уверены, что это правильное исправление. Допустим, человек набрал: «чемпионат России по хоккею 2009 года итоговая таблица», наделал там кучу ошибок, Яндекс ничего не нашел, и, не будучи на сто процентов уверен, что исправление правильное, все-таки произвел автозамену запроса и указал на самом деле правильную выдачу.

Зачастую люди не доверяют Яндексу. В исходном запросе это ссылка, то есть вы можете на нее ткнуть и попасть на опечаточное, скажем так, написание. Статистика у нас такая: многие нам не доверяют, мы показываем такие формы в день 5 миллионов раз. 30 000 раз люди кликают в исходный запрос. 90%, насколько я помню, из них, посмотрев выдачу на то, что они считают правильной формулировкой, все-таки возвращаются на наше, то есть убеждаются, что мы им все-таки помогли.

Если у нас нет 99% уверенности, что мы можем исправить ваш запрос, мы показываем обычную подсказку. Если вы сделали ошибки в словах «агентство» и «недвижимости», мы показываем вам варианты исправлений. Вы можете кликнуть в эту ссылочку, и вы получите правильную выдачу. Выдача, которую видит человек на этом экране, зачастую весьма убогая. Вы видите результаты поиска

[на «агенство недвижимости»] — там 117 000 страниц. Если человек ткнет в «агентство недвижимости», я думаю, он получит 10 миллионов страниц.

Теперь можно рассказать о более интересных для вас вещах — это статистика по ошибкам. Основная группа ошибок — это ошибки в словах. Они делятся на ошибки случайные, как в слове «прокуратруа», то есть просто перепутаны две буквы на клавиатуре, и в слове «тиливизор» — это так называемая когнитивная ошибка: человек уверен, что «телевизор» пишется именно так. Крупная группа запросов — это запросы, связанные со слитно-раздельным написанием. Люди часто ошибаются с установкой пробела — опять же, либо случайно, либо в когнитивном плане. И, может быть, самая интересная группа — это контекстные ошибки, когда хорошее слово является на самом деле опечаткой другого хорошего слова, например, «меховой слон». Само по себе слово «слон» хорошее, его не надо ни на что исправлять. Но в этом контексте, со словом «меховой» — очевидно, это «меховой салон». И совсем небольшая группа ошибок — это когда люди забывают переключить раскладку клавиатуры, печатают всякую тарабарщину, и когда люди печатают так называемый транслит. Вот со всеми этими опечатками мы и боремся, пытаемся человеку запрос исправить.

Для вас, вероятно, будет интересной статистика распределения случайных и когнитивных ошибок. По нашей статистике, они в общем равны, 50 на 50, как ни странно. Казалось бы, случайных ошибок должно быть гораздо больше, но почему-то это не так. То есть неграмотность имеет место быть.

И интересная статистика частот для соотношения, скажем, когнитивных ошибок и для случайных. Первая — это ошибка в слове «скачать» пропущена буква «а»: в 900 раз встречается чаще правильное написание, чем опечаточное. Для слова «одноклассники» без буквы «с» уже чаще встречается соотношение: 1 к 50. А вот для такой стопроцентно когнитивной ошибки, как пропущенная буква «т» в слове «агентство» соотношение всего 1 к 3. Это значит, что, грубо говоря, из четырех раз один раз написали слово без буквы «т».

[Демонстрирует слайд с неправильными написаниями слов.] Вот как вы считаете, эти запросы надо было исправлять автоматически, или надо было дать подсказку? На самом деле, мы все эти запросы исправляем автоматически. Единственным подозрительным выглядит слово «Постернак», потому что, может быть, есть человек по фамилии Постернак. Но дальше вы увидите, почему машина считает, что это все-таки не «Постернак», а «Пастернак».

Вот еще статистика, которую можно из запросов Яндекса почерпнуть. Это распределение частот нормальных, правильных слов и их опечаток. Чемпионом по опечаткам являются опять те же «одноклассники» — 479 уникальных опечаток в день. То есть в слове «одноклассники» люди умудряются делать почти 500 разных опечаток, и делают это 22 000 раз. На самом деле, в слове «одноклассники» еще больше — 1500, даже 1700 различных вариантов опечаток. Это за месяц. Если мы посчитаем за год это количество, там, наверно, будет 2 с половиной тысячи. Можно приклеить много букв, можно переставлять их как угодно, то есть комбинаторика здесь… По-моему, на слово «как» у нас несколько сотен опечаток. У нас есть специальный сервис, который для слов показывает все их опечаточные варианты, которые мы детектировали. И самое интересное: частота опечаток в слове «агентство» по отношению к частоте самого слова — три. Вот как работать с таким словом? Наверное, вы теперь понимаете, почему

мы для слова «агентство недвижимости» не сделали автоматического исправления — у нас просто не хватило уверенности.

Я не буду пытаться рассказать подробно о механизмах, которые используются для исправления запросов, потому что это 2-3 темы, каждая, наверное, минут на 40. Значит, у нас есть некоторые постулаты, которые мы считаем постулатами. Вы их можете увидеть на слайде, и они полностью подтверждаются экспериментом, на яндексовском логе запросов все эти постулаты проверены. И благодаря, скажем так, логу запросов Яндекса, мы можем исправлять эти запросы, то есть лечим запросы при помощи самих же запросов. Что нам это позволяет делать? У нас есть три фундаментальных источника информации. Самое главное — это частотность слов. Мы знаем частоты употребления практически всех слов русского языка за очень продолжительный период, потому что трудно найти слово, которое кто-то не набирал в строке запроса Яндекса. И при таком трафике вы понимаете, какие там частоты и какая точность при статистике. Второе очень важное знание, которое мы обретаем из лога запросов, — это сочетаемость слов. Потому что самого по себе слова, в общем-то, недостаточно. У нас вообще нет ни хороших, ни плохих слов. В некоторых контекстах они могут быть плохими — вот «меховой слон»: «слон» хорошее слово, но с «меховым» никак не состыкуется. И таких у нас примеров миллионы. И очень интересная у нас есть информация — это переформулировки запросов. Это анализ пользовательских сессий. То есть люди, делая ошибки, особенно не когнитивного плана, а опечатки, — они сами в состоянии их исправить. И мы имеем большую статистику по функции замены: сами люди снабжают нас этой бесценной информацией. И, собственно, на базе этой бесценной информации, которая сокрыта в запросах, мы строим эту вероятностную языковую модель. Не буду подробно ее разбирать, просто скажу, из чего она состоит. Главная, базовая единица для нас — это словарь условий. Туда входят однословия и двусловия. Представьте себе, что нам кто-то написал слово «мехавой» в запросе. Вот мы можем прикинуть, что есть рядышком слово «меховой», похожее (я потом объясню, как мы определяем — похожее или непохожее) — и дальше мы смотрим на частоту, видим, что слово «меховой» раз в 5 частотней, чем «мехавой». И, пообучавшись, мы можем определить, что давайте-ка мы слово «мехавой» поправим на «меховой». Почему мы исправили «меховой слон» на «меховой салон»? Если посмотреть на частоты — разница в тысячу раз. Здравый смысл подсказывает, что что-то здесь не то: очень похожее рядом есть слово и в тысячу раз частотнее комбинация «меховой салон», чем «слон». Наверно, разумно подсказать.

Но самих этих частот недостаточно. Решение принимается на основании двух параметров: соотношение частот и, грубо говоря, значение похожести или близости слов, то есть какова вероятность того, что была сделана такого рода ошибка. Допустим, вместо «о» написано «а», или в слове «салон» «а» пропущено. Есть целые классы таких ошибок, и вот примерчик есть на слово «крайсер». Что это — «крейсер» или «Крайслер»? Кто знает? Никто не знает. Скорее всего, «Крайслер», потому что им сейчас больше людей интересовалось, но могло перестать после банкротства, и опять это станет «крейсер». То есть это всё в динамике, всё развивается. Эта информация тоже берется из лога запросов и из исследования функции ошибки. Мы видим, что букву «л» пропускают чаще, чем заменяют «а» на «е». И, соответственно, в этом произведении вероятности — всё построено на вероятности — этот множитель может дать больший вклад. То есть два у нас множителя — частота слов и функция близости.

И в случаях, когда у нас алгоритм обучения не в состоянии, скажем так, обучиться каким-то примерам ошибок, у нас существует словарь замен, построенный просто из пользовательских сессий. То есть мы видим, что «мущина» — это «мужчина» на самом деле. То есть у нас в базовом алгоритме нет, скажем, функции замены «щ» на «жч», но просто по статистике мы можем определить, что такая замена возможна, и тоже можем предложить ее пользователю.

И сбоку к этой языковой модели у нас лежат обычные словари. У нас есть, естественно, информация о словах, являются ли они словарными, но мы эту информацию используем больше как вспомогательную. То есть, еще раз повторяю, у нас нет хороших слов, у нас есть просто вероятности. И слова, которые считаются хорошими, скажем, в каком-то орфографическом словаре, нами ранжируются как более достоверные слова, которые надо, допустим, заменять в крайне редких случаях, скажем так. То есть тогда должна быть разница частот, как в «меховой салон» — «меховой слон»: в тысячу раз — достаточное основание, чтобы сделать замену «слон» на «салон». Вот так работает, собственно, механизм исправления опечаток.

Естественно, поскольку это некоторое упрощение — модель у нас двусловная, как вы видели, — а на самом деле запросы состоят в среднем из трех слов и более, поэтому у нас есть просто ошибка математическая, помимо всяких статистических погрешностей. На нынешний день точность наших исправлений, по нашим данным, составляет 85%. То есть 85% запросов мы исправляем правильно. Полнота у нас при этом чуть пониже — 75%. То есть мы 25% запросов просто не в состоянии исправлять: или боимся, или не знаем, на что, то есть нам еще есть, над чем работать.

И вот несколько примеров тяжелых случаев, когда совершенно непонятно, что делать. Вот появляется в запросе слово «Арфография» — что это, по-вашему? Наверно, вы все решите, что это опечатка. Наверно, вам может быть знакомо слово «Гринландия» — фестиваль авторской песни. На самом деле, это никакие опечатки, а слова, которые получили вторую жизнь и используются компаниями в качестве своего, скажем так, имени. Понятно, что это вносит некий хаос в общественное сознание, потому что, например, «орхидея» — для всех это цветок, а то, что «Архидея» — это какой-то архитектурный фестиваль, знают очень немногие. И когда люди ищут цветок, а попадают на какой-то архитектурный сайт, то многие очень удивляются, думают: что, поисковая система совсем с ума сошла? Потому что люди уверены, что «орхидея» пишется с буквы «а».

И вот еще такой тяжелый случай, когда частоты неправильных написаний оказываются в логе запросов больше, чем частоты правильные. Вот, нам жаловались, что мы даем неправильную подсказку: «трансагенство». У нас слово «трансагенство» набрало большую частоту, чем правильное написание. Такое возможно, это издержки этих языковых моделей.

М.М. ВОЛОВИЧ (Компания «Ашманов и Партнеры»): Я смотрел «трансагентство», там соотношение то же самое: 20 000 неправильных, 60 000 правильных, тем не менее на первом месте то, что вылезает при наборе, — неправильное.

А.В. БАЙТИН: А, это немножко другой случай, это тогда, как говорится, не по делу наезд. Объясню, почему, чуть попозже.

Вот, например, нормальный человек набирает слово «предынсультный» или, еще хуже, «предынфарктный». Неужели он наберет его через букву «ы», как вы считаете? Нет, конечно, все набирают через «и», я бы и сам набрал — набирать через «ы» просто рука не поднимается. Слово «приоретизация» — я не знаю вообще, хорошее ли это слово, наверно, это не очень удачный сам по себе термин, лучше использовать какое-то другое словосочетание. «Контртенор» или «контр-тенор» — как оно пишется? Никто не знает. В этих случаях мы иногда действительно наших пользователей можем дезориентировать, есть за нами такой грех.

Перейду ко второму сервису, сервису проверки орфографии. Это уже более серьезный сервис, он построен не на запросах, здесь вам никогда не дадут в качестве подсказки несловарное слово, здесь будет только конкретно орфографическое слово, признанное профессиональными лингвистами. Наверно, многие из вас пользуются почтовым сервисом Яндекса. Немногие знают, что если там надавить на кнопочку ABC с галочкой, включится функция проверки орфографии, и вы можете проверить свое письмо. Есть такая же функция в «Яндекс.Баре» — это вообще уже встроенная в браузер функциональность, вы уже можете в любой форме: допустим, пост в блог делаете и пишете от всей души там всё, что чувствуете. Иногда получается не очень это всё грамотно. Вот, допустим, вы резюме свое где-то пишете, и ошибочек там много, нехорошо это получается для вашего карьерного роста или поиска работы. Ну и другие случаи есть, когда грамотность имеет некоторое значение. Зачем мы начали делать этот сервис, наверно, понятно: люди должны получать исправленный вариант без опасени, что мы им подскажем какую-то опечаточную форму. Допустим, люди пишут какие-то безумные слова, и мы им предлагаем исправленный вариант. Но предлагать такие слова в деловом письме совершенно невозможно. Я могу сказать, мы еще занимаемся фильтрацией так называемого порно из запросов, у нас порнословарь — 42 000 слов. Там, конечно, много слов, не очень распространенных… И вот количество этого словарного мусора огромно, и иногда мы этот словарный мусор, к сожалению, в качестве подсказки подаем — просто, как говорится, частотно. Поэтому этот сервис, «Я.Спеллер»… Кстати, лингвисты, само слово «спеллер» не коробит ухо профессионального лингвиста? Это рабочее, внутреннее название, мы занимаемся уже 20 лет этой проверкой и привыкли это слово использовать. В основе этого сервиса лежит обычный орфографический словарь — тот же самый, что в Microsoft Word’е, только побольше процентов на 10, и та самая вероятностно-языковая модель, о которой я только что рассказывал. Благодаря их комбинации удалось добиться очень интересных результатов. То, что словарь больше, это не наша заслуга, а поставщика словаря, а второе и третье — это достаточно интересные сдвиги в лучшую сторону в данном направлении. Вот многобуквенные ошибки встречаются в 7-9% случаев. Наш сервис их исправляет, в отличие от вордовского. Word оставит такое слово, в котором больше одной ошибки, без подсказки. И очень важный момент — это точная подсказка. Благодаря знанию о частотах мы можем давать подсказки с невероятной даже по нашим меркам точности. Вот сравнение с Word’ом: человек что-то куда-то постил и наделал ошибок. Допустим, Word не знает слова «супер» и дает подсказки: «сапер», «спер», «супец» и так далее. Наш словарь это слово знает. Вместо «пожайлуста» он выдает «пожалуйста», а Word — нет.

А теперь самый интересный момент — зачем нужна языковая модель? Вот, допустим, опечатка «орода». Word подсказывает «Ирода». Наверняка все

понимают, что «Ирод» употребляется крайне редко в нашем бытовом языке. А слово «города» — гораздо чаще. И Яндекс подобное знает. И яндексовская подсказка «города» стоит на первом месте, а в большинстве случаев вообще подсказка единственная. То есть можно производить фактически такую же автозамену, как и при запросах, что мы, собственно, скоро собираемся делать.

Вот последний сервис — для профессионалов, для любителей русского языка. Вы, допустим, можете набрать слово «правописание», и если вы сделаете ошибку, мы вам навяжем свой вариант и здесь, вы никуда от нас не уйдете. Допустим, вы ищете «праваписание», мы вам заменим на «правописание» и покажем некую информацию. В данном случае вы можете морфемно-орфографический словарь посмотреть, школьный этимологический, словарь синонимов, то есть вся эта информация может быть полезной профессиональному лингвисту в том числе.

В заключение я скажу, что мы планируем в ближайшее время выпустить службу «Яндекс.Правописание», в которой будет возможность проверить орфографию, и должна выйти грамматическая справка для людей, которые сомневаются в наших подсказках, они смогут это проверить, скажем, по словарю Лопатина. И появятся новые словари русского языка, которые будут очень полезны для коррекции грамотности. Все это появится уже в скором времени.

М.А. КРОНГАУЗ: Прежде всего хочу поблагодарить за два необычайно содержательных доклада, я лично узнал чрезвычайно много нового, интересного, и дело не только в содержательности докладов, но и в открытости компании «Яндекс», потому что мы увидели в какой-то мере кухню, нам приоткрыли окно.

Я предлагаю сейчас задавать конкретные вопросы. На вопрос одна минута. После этого будет дискуссия.

А.Д. ШМЕЛЁВ (ИРЯ РАН): Правильно ли я понял, что если человек набрал правильно, допустим, слово «агентство», с буквой «т», то он не получит в своем запросе примерно четверть информации, которую он искал, потому что люди, которые пишут в Интернете, тоже делают ошибки? Или сайты, где слово написано с ошибкой, тоже ищутся?

И.В. СЕГАЛОВИЧ: Вопрос замечательный, мы работаем над тем, чтобы орфоварианты тоже появились в поиске. Пока этого нет.

Е.Н. ЗАРЕЦКАЯ (Академия народного хозяйства при Правительстве РФ): Если человек, который набирает запрос, ошибся в первой части слова, в первых буквах даже, как вообще работает механизм опознания, с чем сравнивать, что подсказывает — там, что это корень? Меня интересует с точки зрения кухни.

А.В. БАЙТИН: Ну кухня очень простая. Ни корней, ни приставок не существует у нас, ни подлежащих, ни сказуемых. Только буковки и слова.

Е.Н. ЗАРЕЦКАЯ: Ну вот он ошибся в первых двух буквах...

А.В. БАЙТИН: Если статистически люди чаще ошибаются в этих двух буквах, например, набирают «адноклассники» или «андоклассники»… Тогда подсказка будет не на первом месте, а на седьмом где-то.

Л.Л. ИОМДИН (ИППИ РАН): Что происходит с фамилиями и вообще с собственными именами, особенно иноязычными? Второе — что возникает, когда какая-то выпендрежная единица типа «Аукцыон» через «ы»? И третий вопрос — собираетесь ли вы делать что-то для английского языка?

А.В. БАЙТИН: Я точно могу сказать, что с фамилиями у нас много проблем, потому что любое слово может быть фамилией. А еще хуже, что сейчас модно стало ники заводить в блогах, там вообще любой набор символов может быть таким ником. То есть нам подсказывать стало очень трудно, потому что кто-то скажет: «Это моя фамилия» или: «Это мой ник, почему меня не показывают?» Если миллионы людей под этим подразумевают собственную опечатку — клавиатуру забыли переключить, а кто-то использует этот набор в качестве ника в своем блоге, — значит, кто-то будет страдать: или миллионы человек или этот человек с этим ником. Что же касается фамилий, это действительно очень тяжелый случай, потому что многих это обижает. У нас есть курьезный пример: Иван Колько. Это герой Яндекса, мы наломали кучу зубов: это либо «кольцо», либо «сколько» — что угодно, только не его фамилия. Вот есть такой человек Иван Колько. Он ищет себя на вебе и не находит, не только в Яндексе, но и в Гугле. Причем в Яндексе он себя еще может как-то найти, в Гугле — вообще нет за счет всяких там автозамен, подмешиваний и всего прочего. Потому что «Колько» очень похоже на опечатку слова «сколько».

И.Б. ЛЕВОНТИНА (ИРЯ РАН): А почему он не может найти в Гугле?

А.В. БАЙТИН: У него нет возможности, вот, например, в Гугле, если Гугл там всё это замешал выдачу, он не может зачастую, даже кавычки не помогают. В Яндексе помогают.

«Аукцыон» — это те самые примеры с «Архидеей», с «Гринландией» — это тоже тяжелый случай, нам приходится с ними каким-то образом бороться, в основном за счет уникальных контекстов отделить эти слова от замен.

А с английским — у нас есть такой же сервис для английского языка, он, собственно, комбинированный: мы исправляем ошибки в русских словах, в английских и украинских. Три языка поддерживаются сейчас. Точность разная, но поддержка есть.

И.Б. ЛЕВОНТИНА: Когда человек пишет нормальное слово, присутствующее в словаре, а вы ему даете подсказку, не так, как правильно, как лингвисты решили, что правильно, а так, как чаще пишут, — то тем самым не усугубляете ли вы ту ситуацию с падением грамотности, о которой так здесь сокрушались?

А.В. БАЙТИН: Ну, естественно, усугубляем.

ВОПРОС ИЗ ЗАЛА: А вы не хотите пользователю позволить сохранить cookies, как он вообще искать хочет? Я понимаю, что я каждый день на новом компьютере, в компьютерном классе, но дома-то у меня компьютер стационарный, уж там-то я, что захочу, наверно, должен иметь.

А.В. БАЙТИН: Ну вот мне Илья благосклонно разрешает доложить о таком противоречии. Вот я, например, считаю, что там должна быть такая вкладочка

типа «я лингвист». Потому что понятно, что вы сидите и набираете как раз вот такие вот ужасы, которые нормальный человек может себе представить в дурном сне. Я считаю, что это было бы полезно. Я сам не лингвист, но иногда приходится быть лингвистом. Приходят жалобы: я пытаюсь это слово вводить, он мне тут же исправляет.

И.В. СЕГАЛОВИЧ: У нас внутренние противоречия, мы до конца не согласовали эту позицию. Мы обсуждаем это все время. Действительно, заманчиво поставить такую галочку, но наша практика и наблюдения за жизнью пользователей показывают, что они совершенно не помнят о том, что они поставили эту галочку. Ничтожное количество людей знает, что эта галочка есть. И эта галочка цели реально не достигнет, и большинство забудет, что это за галочка и где ее ставить. А главное — она будет еще больше запутывать, потому что вы привыкните к этой галочке на одном компьютере, а на другом компьютере, в другом браузере она не стоит. И главное, что поведение перестанет быть единообразным, предсказуемым от появления каких-то дополнительных галочек в этом месте. Я не готов говорить, что эта галочка — абсолютное зло, но надо понимать, что это существенный недостаток.

ВОПРОС ИЗ ЗАЛА: Меня порнословарь заинтересовал. Если человек набирает в строке поиска слово из этого словаря, что ему выдается в качестве выдачи?

А.В. БАЙТИН: Опечатка в порнослове не исправляется. Если человек в длинном порнозапросе сделал ошибку в обычном слове, мы даем такую подсказку.

А.С. НАРИНЬЯНИ (НИИ Искусственного интеллекта): Что вы делаете, если парные кавычки поставить с обеих сторон? Неужели это не средство запроса со стороны лингвиста?

М.А. КРОНГАУЗ: Хорошо, простой ответ.

К.В. АНИСИМОВИЧ (ABBYY Software House): Считаете ли вы полезной фонетическую подсказку в стиле Гугла, и если да, собираетесь ли вы ее делать?

А.В. БАЙТИН: Это функция близости.

М.А. КРОНГАУЗ: Это сложный лингвистический вопрос, по-видимому, через графику это происходит, но опосредованно.

ВОПРОС ИЗ ЗАЛА: А вы не считаете, что грамотность развивало бы отсутствие исправлений? То есть у пользователя было бы представление, что надо написать правильно?

И.В. СЕГАЛОВИЧ: Мы заботимся об экологии и электричестве. Посчитайте время. 80 миллионов запросов в день. 15 миллионов ошибок. И если мы 5 миллионов перестанем исправлять, во-первых, они будут обижаться, и будет плохое настроение, потому что они не сразу нашли то, что надо, или просто не нашли. Электричество будет тратиться на лишнюю работу компьютера.

М.А. КРОНГАУЗ: Ну и тогда к конкурентам перейдут, к Гуглу, который помогает.

А.В. БАЙТИН: У нас не все люди категории до семи лет, которых чему-то можно научить. У нас есть люди за 30, за 40, за 50. Я вот сам, например, пытаюсь чего-то выучить в английском языке и встречаю большое сопротивление со стороны своего мозга, который не воспринимает почему-то. Если мы начнем людей среднего и зрелого возраста учить русскую грамматику, я думаю, мы ничего не добьемся. Уже поздно.

Р.В. ШАРАПОВ (Муромский институт (филиал) Владимирского гос. ун-та): Отмечается ли поисковый спам, в котором неправильный запрос на странице скоординирован с неправильным запросом? Например, можно создать сайт и назваться «агенцтво», через «ц», а потом заспамить систему запросами «агентство», в результате на запрос «агенцтво» я буду получать страницу, отсекая все «агентства», которые написаны правильно?

А.В. БАЙТИН: Ну, это проблемы поискового спама. Это есть, естественно, и этим очень много занимаются.

И.В. СЕГАЛОВИЧ: Это есть, но наша задача — все правильно исправить и выделить большинство. По основному написанию труднее заспамить, потому что оно лучше представлено в Интернете: больше информации, больше характеристик можем найти и насчитать и показать качественный результат. Без Алексея мы мало того что ухудшим настроение пользователю, мы еще и покажем плохую выдачу, потому что с «агенцтвом» плохие сайты.

М.А. КРОНГАУЗ: Ну что ж, завершаем часть «Народ хочет знать» и начинаем часть «Народ учит Яндекс, как надо».

М.М. ВОЛОВИЧ: Я просто зачитаю примеры по исправлению Яндекса. Вот, например, «свази», язык такой, Свазиленд от него, Яндекс меняет на «связи» — не дает подсказку, что: «Может быть, вы ошиблись», а именно меняет и говорит: «Может быть, в вашем запросе была ошибка». «Письмо бамун» меняет на «письмо бонум». «Диссимиляция придыхательных в древнегреческом» — на «диссимилияция при дыхательных в древнегреческом». «Валлийско-латинский словарь» на «валлийской латинский словарь». «Гаплоидная фаза» — дает подсказку: «Может быть, вы ошиблись, имели в виду диплоидную фазу». «Изоструктура» — предлагает «мезоструктура». «Сальное мыло» — предлагает «стальное мыло». Стальное мыло, как выяснилось, таки есть, его больше, чем сального, как ни странно. А недавно, последний пример живой, было набрано «ЛФШ», сокращение от «Летняя физическая школа», было заменено без разговоров на «KAI» — решили, что глупый ребенок набрал, «Казанский авиационный институт» выдается, просто замена регистра.

В.И. БЕЛИКОВ (ИРЯ РАН): Во-первых, я хотел спросить здесь присутствующих: кто на запрос получал такую реакцию в виде письма от Яндекса к вам? Прошу поднять руки. Мне Яндекс вместо ответа на мой запрос прислал: «Вы робот?» [Трое поднимают руки.] Он принял меня за своего! Это компания «Яндекс», они рассказывают, как работает робот Яндекс. Мы здесь можем говорить, что мы близки Яндексу по духу. То, что Яндекс в какой-то момент перестал давать мне «телифон» через «и», даже, если я поставлю его в кавычки, поставлю перед ним восклицательный знак — паук такой есть, телифон, — а Яндекс мне категорически предлагал использовать «телефон».

В.П. СЕЛЕГЕЙ (ABBYY Software House): По этому поводу есть явное недоумение у аудитории, потому что было явно сказано, что если «тилифон» меняется на «телефон», то на странице присутствует информация — в результате «вернуться к исходному запросу».

В.И. БЕЛИКОВ: Я опровергаю это. Сейчас может быть так, но в течение длительного времени было не так.

Два слова об орфографии. Мнение, что для русского языка есть кодифицированная орфография, верное. Что ее следует придерживаться, — это результат правового нигилизма. Про это я говорил на одном из «Диалогов», что никто не утверждал всерьез нашу орфографию. Вот картинка, как оно было, и как оно появляется в словаре [демонстрирует слайд]:

Орфография на практике: «Русский орфографический словарь» (2005) и СМИ России

(количество статей с использованием соответствующей орфографии в 3054 СМИ России, по базе «Интегрум»)

	2000	2001	2002	2003	2004	2005	2006
брэнд	**3168**	**5520**	**7419**	**11856**	13921	17913	16369
бренд	1448	2441	4948	9318	**15165**	**23429**	**32950**
барсетка	**427**	**678**	**833**	**1135**	**1291**	**1457**	**1212**
борсетка	4	20	9	32	45	155	93
блоггер	0	0	1	**23**	**50**	**148**	**2968**
блогер	0	0	0	1	4	14	129

«Бренд» через «е» — он сам по себе, а «барсетку» орфографически правильно писать как «борсетка» — всем наплевать, что в словаре записано. Так что народ победит орфографический словарь.

А.Д. ШМЕЛЁВ: Два утверждения, косвенно связанные с Яндексом. Утверждение первое: меня удивляет утверждение, что желание следовать каким-то нормам может быть результатам правового нигилизма. Также я безусловно против того, чтобы какие бы то ни было государственные инстанции вмешивались в орфографию. Второе: у многих лингвистов есть некоторое заблуждение, которое цитировалось вчера М.А. Кронгаузом, что орфография не допускает вариантов и что лингвистика вообще так считает. Орфография всегда изучала узус — это известно всем, кто изучал древнерусскую орфографию, и норма всегда отличалась от узуса, и вовсе не обязательно, чтобы какие-то правила при этом утверждались.

Теперь я скажу несколько слов уже о Яндексе. Действительно, некоторая проблема поиска возникает в таких случаях, как «блогер» и тому подобных вещах. Яндекс, безусловно, будет искать, и считает правильным, и дает подсказки на более частотное. Проблемы возникают, когда частоты конкурируют и при этом меняются, и можно все время это отслеживать, но, кажется, что в этом смысла нет, а смысл есть в том, что есть возможность видеть и то и другое. И это главное

желание пользователя — иметь возможность найти только что-то одно, только что-то другое ,или и то и другое, или еще разные варианты написания слова. А вторая проблема, которая пока непонятно, как решается, — ясно, что бывают слова довольно равной частнотности, которые пишутся по-разному, например, «компания» и «кампания». И вот мне непонятно, как дружески ведет себя Яндекс, который понимает, что можно сделать ошибку, и ошибка вполне вероятна, и вот даются ли по этому случаю подсказки?

М.А. КРОНГАУЗ: Как между русской «коровой» и белорусской «каровай».

А.Д. ШМЕЛЁВ: Совершенно верно.

ВОПРОС ИЗ ЗАЛА: Есть такая большая область, как медицина. Сейчас врачей заставили переходить на международный стандарт диагнозов или симптомов, и врачи не знают, как эти диагнозы и симптомы пишутся. Значит, человек, который интересуется своим здоровьем, а таких большинство, он использует то, что ему написал врач в карточке, и обращается к поисковой системе. И получается, что вариантов терминов — порядка трех-четырех, и получается, что статьи, которые на все варианты запроса, — все одинаково хороши. Может быть, они разные, дополняют друг друга, но они все честные и правильные. Хорошо было бы эту ситуацию неопределенности, может быть, как-то поддержать. Может быть, давать все четыре варианта написания этого диагноза или симптома? Поскольку человек строит свой запрос только по одному варианту, который у него написан в медицинской карточке.

М.А. КРОНГАУЗ: Я хочу сказать, что мы говорим не только о Яндексе, и я сейчас скажу не очень корректную фразу, но тем не менее. Обращаюсь к народу — если вы такие умные, то пользуйтесь Рамблером, он вообще ничего не подсказывает. (Если вы недовольны подсказками.) Так что здесь все-таки мы предпочитаем пользоваться Яндексом, именно потому что Яндекс предлагает довольно интересную систему подсказок. Пока кроме одного указания В.И. Беликова, которое требует проверки, мы видим, что всегда можно все-таки найти ответ на свой запрос, только проделать несколько действий.

В.П. СЕЛЕГЕЙ: Вчера сотрудники Яндекса не присутствовали на докладе, и это несколько разрушило архитектуру обсуждения. В докладах И.В. Сегаловича и А.В. Байтина прозвучало, что следовало бы хранить кластеры вариантов и использовать их при поиске, это очевидно, и могло бы решить проблемы.

М.А. КРОНГАУЗ: По-моему, мы перешли в формат «Народ делает наказ Яндексу». Мне кажется, что это неправильно, потому что Яндекс ориентируется на разные факторы, прежде всего конкурирует с другими поисковыми системами.

К.В. АНИСИМОВИЧ: Я хотел бы добавить пример, что как раз фонетические соответствия «в стиле Гугл» решает проблему. То есть необязательно хранить все описания, а просто давать все фонетические соответствия, опирающиеся на согласные с сильной заменой гласных, то будет то, что надо.

М.А. КРОНГАУЗ: Но вообще мне кажется, что автозамена — это самый высокий способ обращения с пользователем.

А.С. НАРИНЬЯНИ: Вы сказали, что плохо быть в меньшинстве. Я думаю, что быть с народом значительно хуже, поэтому давайте будем в меньшинстве. Другое дело, что у нас есть возможность сделать тот запрос, который мы хотели. Не будем обижаться за то, что мы в меньшинстве.

М.А. КРОНГАУЗ: Просто придется сделать несколько лишних действий.

А.С. НАРИНЬЯНИ: Это нормально, мы этого заслуживаем, в конце концов.

М.А. КРОНГАУЗ: Я согласен с этим, но это просто плата за некоторую изысканность, если хотите.

Е.Я. ШМЕЛЁВА (ИРЯ РАН): Я просто хотела вас развлечь, рассказать, как я как лингвист разговариваю с Яндексом, когда он мне подсказывает. Моя семья застала меня за тем, что я искала там нечто, что находится в моем родном районе Ново-Переделкино. Набрала, естественно, «в Ново-Переделкине». Яндекс мне подсказывает: «Наверно, вы имели в виду “в Ново-Переделкино”». Что я ненавижу, когда говорят «в Переделкино». Причем это очень известная история, что Ахматова говорила, что она очень терпеливо относится ко всем ошибкам, но когда она слышит, что когда кто-то говорит «в Переделкино», ей хочется человека потрясти там, поправить. Я, значит, сижу и говорю компьютеру: «Идиот! Ты будешь меня учить! Я всей стране объясняю по радио, как правильно говорить!» Понятно, что я сама сразу понимаю, откуда это вылезает. Значит, частотность не только орфографическая, сочетаемостная, но и вообще разная — управление неправильное, которое сейчас сплошь и рядом. И вот как с этим быть? Я прекрасно понимаю вашу задачу. Но все-таки, как я понимаю, вам тоже приятно, чтоб люди были грамотные, у вас есть такой посыл.

А.А. КОТОВ (РГГУ): А у меня такое общее соображение, не про Яндекс совсем. Да, много было нужных слов сказано про то, что нужно экономить электроэнергию и нужно сделать общение пользователя с комьютером максимально продуктивным и максимально интерактивным, и поэтому мы будем отдельные опечатки будем заменять на слова по частотности, и мы начало поисковых запросов будем заменять на полные поисковые запросы тоже по частотности, встречаемости в поисковых запросах в Интернете, но вроде тогда получается, что поисковые системы нам навязывают не только орфографию на основе частотного запроса, но они и еще и навязывают нам объекты интереса, который появляется в сети. Вот маленькая история про моего друга. Он ко мне приходит буквально на прошлой неделе и говорит: «Ты знаешь, Медведев — еврей. Я вот в поисковом движке искал “медицинские анализы”, набрал “мед” — показывает: “Медведев еврей”. Я проверил: на “мед”, он подсказывает: “Медведев”, а на “Медведев” он подсказывает: “Медведев еврей”. Ты знаешь, я пошел на этот сайт, представляешь, там все это описано». Получается, что за счет такого обеспечения интерактивности к сайту человек, который начал искать «медицинские анализы», получает некоторое такое, что интересно и что обсуждается в сети. Понятно, что нам сказали, что всякая порнография попадает в специальный словарь, значит, «Медведев еврей» не попал в соответствующий словарь, пока эта выдача не прекратилась. Но получается, что поисковая машина становится средством трансляции мнения большинства, как средство массовой информации, как телевизор, как блог, куда заходишь, и там написано на первой странице: «самая обсуждаемая тема сегодня», и там какие-то гнусняцкие темы, а если зайти в свою ленту, то там какие-то умные люди чего-то пишут.

Общая проблема, которая мне кажется важной, — не происходит ли за счет развития этой интерактивности навязывание искажения взаимодействия пользователя с сетью и навязывание пользователю не только орфографии каких-то интересов, которые навязываются пользователями то ли спамерами, то ли какими-то сообществами, не очень понятно.

И.С. АШМАНОВ (Компания «Ашманов и Партнеры»): Я хотел еще два вопроса задать. Первый, он такой, что… Я не в качестве наезда говорю, что Алексей делал «Орфо» в свое время. На самом деле, это очень прекрасная преемственность, он когда-то мне написал: «Я снова работаю в Яндексе, снова исправляю чужие ошибки», опять к этому пришло через 20 лет, условно говоря. Но вот смотрите, еще покойный Старостин говорил, как письменность фиксирует язык, что скорость появления языка еще с появлением письменности резко падает. Вот затем появляется спеллинг-чекер. Он, по идее, должен еще притормозить сильно, потому что вот стоит Word, одну и ту же грамотность в кавычках навяывают сорока миллионам пользователей в России, или больше. Вот и Яндекс, собственно, предлагает некий усредненный варинат. С другой стороны, язык меняется. Понятно, что Яндекс — это такое скользящее окно, понятно, что он учитывается частотность. Почему, собственно, есть возможность сравнивать «Орфо» и Word с Яндексом — потому что никто спеллинг-чекер в последние 10 лет не развивал, он сам зафиксировался. Вот вся команда ушла, человек, которые делал там, «Орфо» выпускал, — в Яндексе, человек, который делал подсказки, — у меня, ну и все, а там никого нет. И Microsoft просто воспроизводит, собирает под 64 бита и так далее. Нет, он не стал хуже. Там была история, когда напали эти долбанутые индусы с толерантностью, которые управляют этим проектом, потребовали убрать слово «голубой», «негр» и так далее, это известная история. Я про другое. Яндекс все-таки скользящим окном движется в некотором смысле по языку. То есть какая у него есть инерция? Слово, как В.И. Беликов показал, уже добралось до словаря, — что делает Яндекс? Оно у него где в этот момент? И есть еще вторая проблема, как я ее называю — проблема сингулярности. Тот же «Орфо» можно взять или проверку спама. По какой-то причине не доходят письма. А на самом деле, человек, который разрабатывает этот фильтр, он либо один, либо их пятеро, и они определяют, что произойдет с письмами 25 миллионов человек на Mail.Ru, например. Здесь точно такая же ситуация, что один технократ, или, может быть, их там пятеро, определяет то, что случится дл 30 миллионов человек. Понятно, что оператор в Чернобыле тоже обладал такой властью. Здесь происходит сингулярная концентрация власти у технократов. Когда претензии к Word’у предъявляют, это же претензии ко мне лично. Я один занимался всей лингвистикой там, по сути. Все сомнительные решения приниимал только я. А проблема «слитно или раздельно» до сих пор не решена.

М.А. КРОНГАУЗ: Но вы их принимали.

И.С. АШМАНОВ: А мне что было делать? Любой технократ принимает решения, точно так же, как Яндекс. Они должны принимать, чтобы эта штука работала. Вот, значит, два вопроса: первый — это инерция и фиксация языка, а второй — вот эта самая сингулярность: вы там втроем или впятером решаете за страну.

М.А. КРОНГАУЗ: Вопрос очень хороший, но возвращаемся к самому началу: есть поисковые системы, которые не принимают этих решений. Что лучше — принимать и подвергаться критике или не принимать?

Е.Н. ЗАРЕЦКАЯ: Я хочу сказать, что компьютерщики отличаются завидной прямолинейностью. Но выбранный принцип частотности имеет огромное количество недостатков. Главный его недостаток заключается в том, что большинство, которое не собирается его менять никоим образом, оно ищется в нижних слоях населения, где очень много неправильного правописания, неправильной стилистики и так далее. Как только вы начинаете транслировать эту норму большинства, вы делаете ее нормой еще большего количества людей, и это колоссальная проблема. А выход из нее, наверно, тоже может быть. Даже если есть принцип частотности… вот, например, тут говорили об ошибках, где неправильное правописание встречается чаще, чем правильное. Но таких случаев не так много — и «Переделкино», и «Трансаэро», и «агентство»… Вот можно выделить класс таких ситуаций и что-нибудь отдельное про них придумать, и огромное количество возражений будет сразу снято. Это называется исключение из общего правила. Вот, что я хочу порекомендовать Яндексу.

А.В. БАЙТИН: По поводу ошибок — у нас известна точность, у нас в 15% запросов мы делаем неправильную подсказку, мы об этом знаем. Где-то в проценте от исправлений мы тоже делаем неправильную автозамену. Вы можете это умножить на поток и посчитать, какие это цифры. Мы об этом знаем, но, к сожалению, мы не боги: что можем, то делаем. По нашим данным, мы находимся на одном уровне с Гуглом сейчас.

И.С. АШМАНОВ: По точности лучше.

А.В. БАЙТИН: По точности лучше, да. По полноте также. А по поводу частотности, вот этот тезис, что в большинстве случаев люди пишут грамотно, то есть грамотных написаний больше, это факт. Тяжелые случаи, где частнотность складывается неудачно, обрабатываются специальным образом. У нас целый набор так называемых стоп-листов, куда мы вручную заносим тяжелые случаи.

И.В. СЕГАЛОВИЧ: Я несколько важных вещей запомнил и сразу хочу на них ответить. На самом деле, две вещи, над которыми мы сейчас работаем, и которые помогут ответить на какие-то важные вопросы. Ну вот, например, поиск орфовариантов, приводились медицинские примеры — у меня есть в голове замечательное слово «таэквондо» с четырьмя равночастотными вариантами, причем два из них — это две разных федерации, которые между собой еще воюют, с разными написаниями (или даже три). Дальше еще важная мысль о том, что хорошо бы показать несколько вариантов ответов ясным образом, это у нас тоже разрабатывается, то есть мы хотим сделать одновременно выдачу, состоящую из ответов для двух вариантов. Но вот тема «в Переделкине» — она действительно болезненная, и я не знаю, как… то есть я сам пишу «в Переделкине», но недавно по «Эху Москвы» показывали, что победило «в Переделкино», или где-то я читал, то есть это такой бюрократический, канцелярский вариант — он сейчас доминирует, к сожалению. У нас в «Яндексе» была чудовищная дискуссия на эту тему, у нас тоже много географических названий, тоже была война, но победили вроде бы те, которые хотят склонять.

И.С. Ашманов замечательно сказал, что у нас замечательная сингулярность и концентрация власти в нескольких руках, чтобы ее не было, что мы делаем — мы приходим сюда, рассказываем, как это всё работает, то есть мы понимаем опасность и стремимся опрозрачить свою работу. Чем понятнее будут механизмы и наше поведение, тем ниже уровень опасности вот этой страшной концентрации власти в руках А.В. Байтина, который одним небрежным движением руки может поменять грамотность у миллионов.

А по поводу нормы и скользящего окна Алексей ответил. Я могу сказать следующее: я не думаю, что поисковая система так уж прям действительно формирует норму. Все-таки есть огромное количество разных вариантов, есть разные слои письма, есть блоги, где люди пишут, не спрашивая поисковую систему, и они сами создают некие нормы. Вот этот «блоггер» или «блогер» — никто не ищет это слово в Яндексе, это слово в запросах встречается ничтожное количество раз. Но оно встречается в текстах, и люди, когда пишут, они консультируются в Яндексе. Кто там победил сейчас — «блоггер» с двумя «г»? Ну, значит, так тому и быть.

М.А. КРОНГАУЗ: «Блоггер» занял ведущее место в Интернете, но проиграл словарю.

Пора завершать наш разговор. Я хочу поблагодарить коллег из «Яндекса», которые показали нам свою кухню и что на этой кухне происходит много интересного. Кроме того, я хочу поблагодарить всю аудиторию, которая была чрезвычайно активна. Было много интересных высказываний, но вы сами видели, что мы все время уходили на глубокие философские темы. И одна из них состоит в том, трудно ли быть меньшинством. На мой взгляд, результат этой дискуссии состоит в том, что чтобы оставаться меньшинством, надо постоянно прилагать некоторые усилия, например, три раза жать на кнопку вместо одного. То есть это, по-видимому, требует от нас усилий настоять на своем, оставаясь в меньшинстве.

Яндекс еще обвиняли в том, что он направляет мнение большинства или, что хуже, спамеры, имитирующие большинство, таким образом подавляют меньшинство. Но ведь это не Яндекс, так устроен наш мир, и мы это прекрасно знаем. Но для нас важно, что так устроен язык, ведь язык тоже подавляет меньшинство, и мы тоже это знаем, уступаем мнению большинства и изменяем норму.