Сбор текстовых ошибок в отдельный файл

Это решение — не для всех. Оно лишь для тех, кому не пофиг, насколько безупречен в плане отсутствия ошибок текст вёрстки, с которым ты работаешь. Очевидно, что над задачей доведения текста до такого результата должны вместе работать верстальщик и редактор. Проблема в том, что пока не было подходящего инструмента для решения такой задачи. Теперь он есть, это скрипт, запускаемый в программе индизайн.

Данная программа позволяет собрать в отдельный текстовый файл все слова, которые спеллер отмечает, как ошибочные. Такой файл можно сделать сразу после помещения файла в макет книги. А дальше редактор поработает уже именно с этими словами. Что-то изменит, что-то оставит как есть. И потом верстальщик уже внесёт правку в текст. Такая адресная работа с требующими внимания словами должна повысить качество подготовки текста.

А вам всё равно или нет, что в вашей вёрстке есть грамматические ошибки?

Меня это всегда расстраивало, когда было много редакторской правки именно ошибок. Можно, конечно, утешать себя, что де редактор вычитает, а я исправлю. Но это больше похоже на самообман. Время-то твоё, личное на внесение правки уходит.

И ещё такой момент. В идеале текст должен вычитываться в Word, и на правку приходить без ошибок. Но реальность такова, что сейчас редактор отправляемый на вёрстку текст только просматривает. А настоящая читка выполняется уже, когда редактор получает вёрстку. Так сложилось. Нельзя сказать, что это плохо, просто надо учесть такой современный подход в организации работы с текстом.

Если верстальщик имеет знания и амбиции исправить текст

Да, в индизайне есть динамическая проверка правописания. Но на большом отрезке времени отличный результат можно получить только работая в режиме однозадачности. И мы и заняты в первую голову вёрсткой. А ошибки, ну иногда правим, но это в большинстве случаев действия по остаточному принципу.

И дело не в нас, что мы такие балбесы, а просто в индизайне нет удобной возможности включить просмотр проблемных с точки зрения спеллера слов. Включить так, чтобы это было на какое-то время основной задачей.

То, как есть сейчас — окно **Проверка орфографии (Редактирование > Орфография > Проверка орфографии)** — нельзя считать удобным. Ну в самом деле — индизайн показывает слова одно за другим, которые он считает ошибочными, ты нажимаешь кнопку **Пропустить** для показа следующего «подозрительного» слова, и этот процесс очень скоро станет казаться бессмысленным: неизвестно, сколько проблемных слов найдено, и сам подход щелкать мышкой на каждом слове затратен по времени.

Если текст правит редактор, читая вёрстку

Хорошо бы сначала собрать вместе все проблемные слова, чтобы просмотреть именно их, а не вылавливать слова в тексте вёрстки.

И вариант «собрать вместе все проблемные слова» — один из возможных. А если организовать эту выборку таким образом,

например, отдельно только русские слова, отдельно только английские, отдельно такие, в которых есть смесь русских и латинских букв, это точно упростит работу редактору в приведении текста в порядок.
И конечно, полезным было бы указание, на каких страницах встречается такое слово в тексте, это может понадобиться на этапе правки.

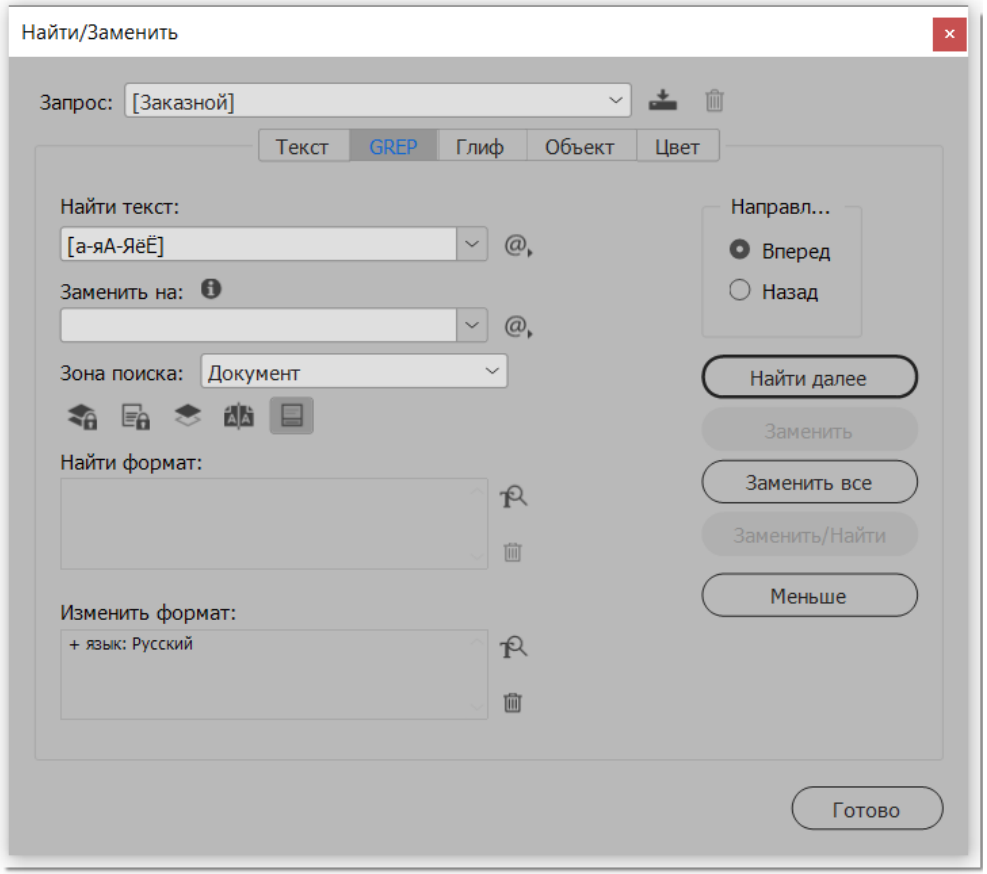
Наверняка такие мечты «хорошо бы…» приходили в головы многим. Да не было удобного инструмента, чтобы решить эту задачу. Но теперь такая программа есть. Верстальщик может собрать в текстовый файл все слова документа, которые спеллер считает ошибочными.

Тогда работу с файлом целесообразно организовать так:
1) первая вёрстка текста, только приведение текста в порядок и назначение нужных стилей; получение списка слов, которые надо проверить.
2) проверка редактором этого списка и внесение верстальщиком исправлений в сделанную вёрстку.
3) полноценная верстка, которая будет внимательно вычитываться редактором/корректором.
При таком подходе вероятность пропуска ошибок в тексте вёрстки будет несравнимо меньше, чем при обычном варианте.

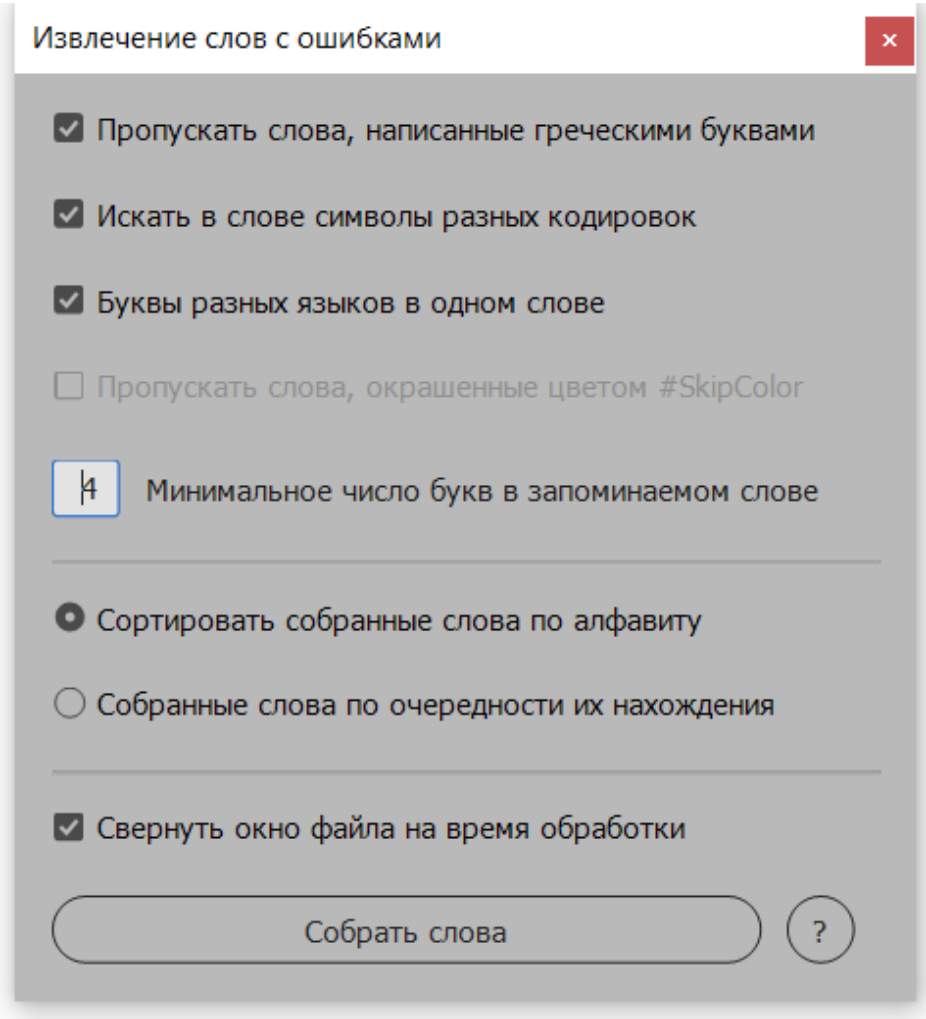
Про небрежные ошибки в тексте

Прежде чем смотреть ошибки, надо убедиться в том, что текст имеет правильное указание выбранного языка. Бывает так, что на экране русские буквы, но весь текст отмечен, что он английский. Или английский текст имеет языковую отметку, что он русский.

Это задача верстальщика — назначение правильной кодировки. И для русских и английских букв это делается при помощи grep-поиска.
Назначение русским буквам языка «Русский»:
• в строке grep-поиска указать **[а-яА-ЯёЁ]**;
• в установках изменения формата определить, что для этих букв будет русский язык;
• Зона поиска — Документ;
• Нажать «Заменить всё».
Для присвоения английским буквам кода нужного языка в строке поиска указать **[a-zA-Z]** и определить язык.



Эти запросы, кстати, можно сохранить для использования в других работах.
После этих важных изменений можно запускать программу **CollectSpellWords.jsx**. Справа её окно.

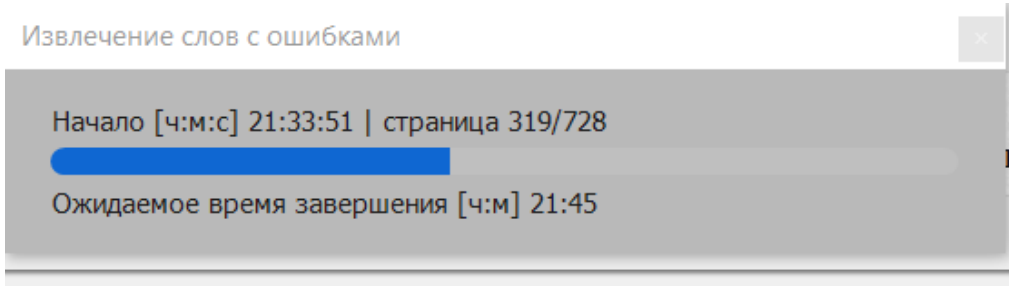


Пропускать слова, написанные греческими буквами — не включать в файл условные обозначения, хотя спеллер их подчеркивает как ошибки;
Искать в слове символы разных кодировок — это иногда встречается в русских текстах: похожие по написанию буквы оказываются английскими. При этом слово отмечено, что его язык русский. Это может быть как для строчных, так и для прописных. Вот русские буквы, визуально совпадающие с английскими: суроеха МАРТЕХНОВСК.
В этой программе такие слова собираются вместе.
Буквы разных языков в одном слове. Это как тот случай, о котором говорится в пункте «Про небрежные ошибки в тексте».
Пропускать слова, окрашенные цветом #SkipColor — всего не предусмотришь, и может быть так, что слова какой-то части вёрстки не имеет смысла отправлять в файл с ошибками. Надо сделать служебный цвет **#SkipColor** и отметить им эту область. Если в вёрстке такого цвета нет, то флажок недоступен.
Минимальное число букв в запоминаемом слове — вряд ли имеет смысл помещать в файл сокращения и артикли, поэтому можно определить порог длины слов. Всё, что меньше, будет пропускаться. Если там единица, то будут собраны все найденные спеллером слова.
Сортировать... — две радиокнопки, определяющие, как будут представлены собранные слова.
Свернуть окно файла на время обработки — когда этот флажок снят, то мы видим прохождение программы по тексту. Это движение обусловливает постоянную перерисовку экрана, и теоретически обработка должна идти быстрее, если свернуть окно вёрстки.

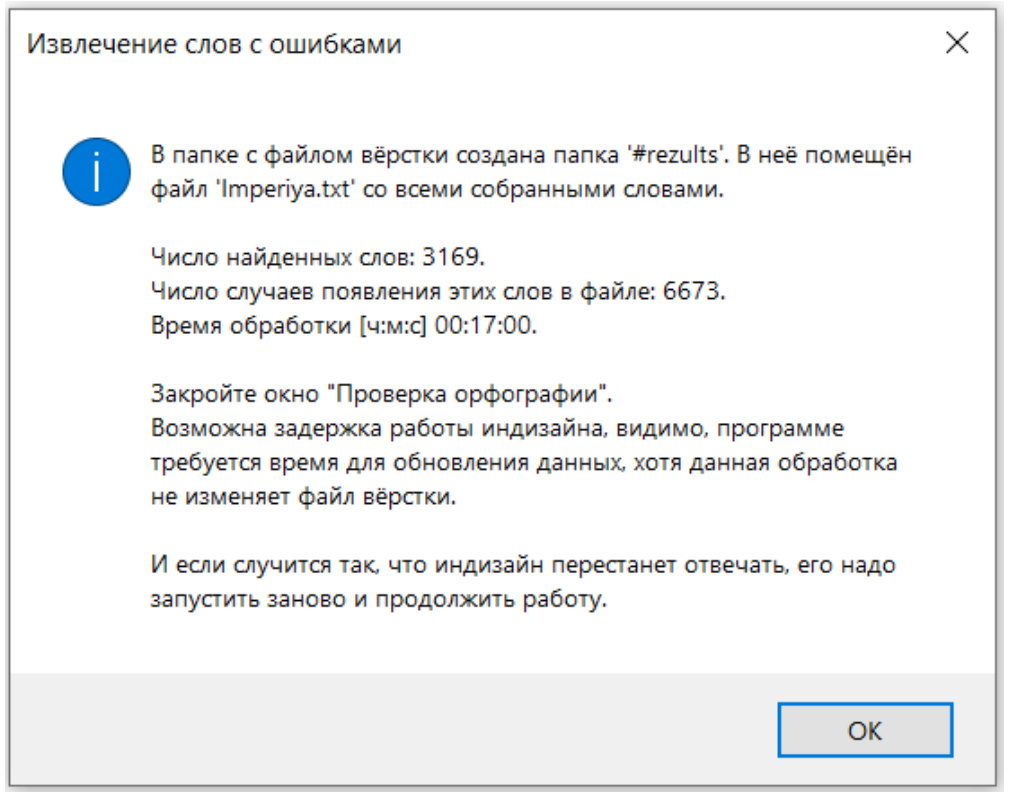
При запуске скрипта кроме этого окна появится окно «Проверка орфографии». Его надо расположить около края экрана, чтобы оно не перекрывало окно прогрессбара, которое будет размещено в центре экрана.

Идея прогрессбара

Неизвестно, сколько слов спеллер считает ошибочными. А обработка может длиться минуты, а то и десятки минут. Всё зависит от числа слов, которые отметил спеллер. И нельзя, чтобы всё это время на экране ничего не происходило. Мы должны быть уверены, что программа не зависла и обработка продолжается.
Вот как это сделано в отсутствие информации о числе проблемных слов. Известно, сколько страниц в документе; для каждого выделенного слова можно узнать, на какой оно странице; и можно сделать прогресс-бар прохождения страниц документа, а не выборки найденных слов.



Отображается время начала обработки, число обработанных страниц, в него включена и та, с которой сейчас извлекаются слова, и общее число страниц. Тут «общее число страниц» это не только страницы документа, но и все мастер-страницы, т.к. на них тоже могут быть ошибки.
После обработки 10-й страницы вычисляется ожидаемое время завершения обработки и выводится в окне прогрессбара. Поскольку реальность такова, что скорость обработки замедляется, то это время уточняется после обработки каждой страницы.



Результат работы

В папке обрабатываемого файла создаётся папка **#rezults**, в ней будет txt-файл, название которого совпадает с названием indd-файла (кодировка UTF-8) со всеми найденными спеллером словами. Об этом будет сообщение, как показано справа. Этот 728 страничный документ обработан за 17 мин., и в нём собрано 3169 слов.
Кто-нибудь готов столько раз щелкать на кнопке **Пропустить**, чтобы просмотреть все слова? Вряд ли. Да и смотреть проблемные слова и принимать решение должен редактор, а для этого эти слова должны быть отдельно от вёрстки. Теперь такая возможность есть.

На следующей странице показан простой тестовый файл **Test(rus).idml**, он в папке **Info**, и результат работы данной программы. Обработка была с установками, показанными на с. 3.

Пример файла

Ошибка в слове на мастер-странице: самалет

Похожие буквы из разных кодировок
(вместо русской х оказалась английская x): вездеход

Слова написаны греческими буквами ΑΛΡΗΑ Οδίλ

Ошибка в русском слове: конпьютер

Ошибка в английском слове: komputer

Русское слово, но язык остался английский: книга

Слово набрано русскими буквами, но часть букв имеет атрибут другого языка: алфавит

В тексте на рабочем столе есть слово с ошибкой. Оно попадёт в отчёт, но вместо номера страницы будет прочерк.

Сокращение отмечено как ошибочное, но в файл не попадёт, т.к. длина слова с ошибкой меньше установленного минимума: сделано в 2024 г.

Окно «Проверка орфографии» надо закрывать самим. В инструментарии скриптинга есть возможность открыть окно, но нет способа его закрыть.

Результат обработки

Test(rus).indd

Число найденных слов: 7.
Число случаев появления этих слов в файле: 7.
Время обработки [ч:м:с] 00:00:02

== Слова с ошибками ==
komputer : 1
книга : 1
конпьютер : 1
пррограмма : –
самалет : A

== Разные кодировки в одном слове ==
вездеход : A

== В одном слове символы разных языков ==
алфавит : 1

После слова двоеточие, и дальше страница, где оно встречается. Если страниц больше одной, они будут через запятую. В данном примере два слова вместо номера имеют букву. Это слова с ошибками размещены на мастер-странице.

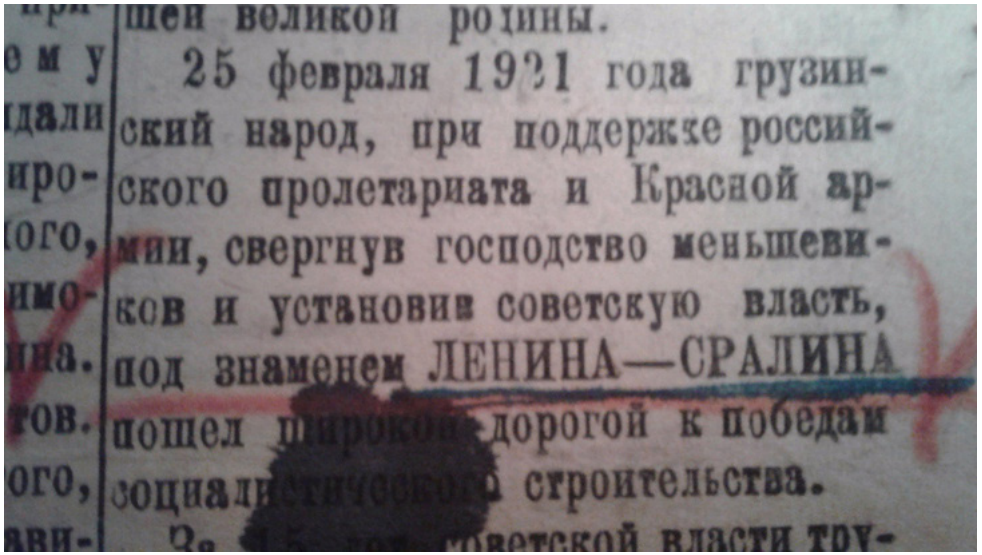
Если фрейм с ошибочным словом на рабочем столе, то вместо номера страницы будет прочерк.

Как использовать этот файл

Текст надо поместить в редактор Word. Может быть, разделить английский и русский тексты на отдельные файлы.

А дальше при чтении отмечать действительно проблемные слова цветом. Может быть, появится понимание, что хорошо бы что-то из этих слов поместить в пользовательский словарь. Эти слова надо отметить другим цветом.

Маркировка цветом — это самый простой способ передать верстальщику информацию, какие слова надо исправить.



шей великой родины.

25 февраля 1921 года грузинский народ, при поддержке российского пролетариата и Красной армии, свергнув господство меньшевиков и установив советскую власть, под знаменем ЛЕНИНА—СТАЛИНА пошел широкой дорогой к победам социалистического строительства.

А оно мне это надо?

Я понимаю эти сомнения. Ну были же опечатки. И будут. Интересно узнать, что когда-то в Англии готовили к печати энциклопедию, над ней работало много специалистов. Все статьи перечитывали по нескольку раз. После получения тиража увидели ошибку: на обложке написано Энциклопудия.

В СССР был случай, когда на титуле вместо Политиздат оказалось Полипиздат.

Хорошо веселиться, читая о разных очепятках:

1910 год, газета «Киевская мысль». Скандал из-за ошибки в названии публикации об императрице. Предполагалось, что она будет называться так «Пребывание вдовствующей императрицы Марии Федоровны в Финляндии». Но почему-то в первом слове вместо буквы р оказалась буква о.

1972 год. В газете «Киевский комсомолец» вышла статья с заголовком «Советы задоводу-любителю». Говорят, главред поседел после этого.

и бомбардировщиков. Несколько часов ряд они наносили удары по укреплениям ивника и его прифронтовым коммуни-

ветская пехота. Стремительным броско наши бойцы преодолели многочисленны заграждения и три сплошных линии тра

ВЕРХОВНОМУ ГАВНОКОМАНДУЮЩЕМУ КРАСНОЙ АРМИИ
МАРШАЛУ СОВЕТСКОГО СОЮЗА
СТАЛИНУ ИОСИФУ ВИССАРИОНОВИЧУ

связи с освобождением столицы Юго-ии Белграда позвольте мне от имени одно-Освободительной армии и народов славии выразить благодарность Вам, Верховному Главнокомандующему ей освободительной армии, которая со-стно с Народно-Освободительной арми-с изумительным героизмом освободила град. Освобождение Белграда для на-народов имеет историческое значе-особенно потому, что страна этих ученных народов является той ареной, совместно проливали кровь сыны ве-

запечатлено кровное братство народ Югославии с народами Советского Сою

Да здравствует героическая Крас Армия!

Да здравствует великий Советс Союз!

Да здравствует Верховный Гла командующий Красной Армии вели Сталин!

Смерть фашизму, свободу народу!

Верховный Главнокомандую Народно-Освободительной а Югославии

Маршал Иосип БРОЗ-ТИТ

Мы будем беспечно хохотать до колик, читая про *чужие* ошибки. Но представьте себя, дикий ужас и холодный пот, если бы в сфере *ваших* обязанностей оказались напечатанными вот такие ошибки, как показаны на предыдущей странице.
Редколлегия центральной махачкалинской газеты была расстреляна за букву Р вместо Т в фамилии Сталина.
Сотрудники туркменской газеты «Коммунар» были уволены без права заниматься печатной деятельностью за ошибку в слове «Главнокомандующему».
И это просто пропущенные ошибки. Тогда технических средств поиска их не было, только глаза.

К нам в типографию КГБ принесли срочный заказ — напечатать книгу о штурме Белого дома. Спрашиваем «Текст читать надо?» — «Нет, всё в порядке, печатайте скорей».
Отогнали тираж. Похоже, текст никто толком не читал. Всякого мусора было много, и самой «сильной» ошибкой было «...защитники еблого дома...»
Я не знаю, как тот посредник потом объяснялся с заказчиками.

И я абсолютно уверен, что лучше глубоко вздохнуть, а потом спокойно внимательно прочитать каждое из этих отмеченных спеллером слов. Благо они все в одном файле.
Не просмотреть текст по диагонали, а осознанно прочесть каждое слово. Это не займёт много времени. А ведь даже если ты среди нескольких тысяч слов найдешь только одну такую ошибку, «Полипиздат», «гавнокомандующего» или «еблого», это уже повод махнуть 50 грамм коньяка за свой профессионализм.
Ошибка всё равно будет найдена, и если это уже в тираже, то возможен громкий разбор полётов, на котором я никому не желаю быть главным обвиняемым.
Надо беречь свою репутацию. Она долго создаётся, но может быть быстро и навсегда испорчена. Оно вам это надо?
Подумайте над этим.

Обновление 17.08.2024

В пункте «Про небрежные ошибки в тексте» была высказана идея, что нужно готовить grep-запросы для обработки текстов на разных языках. Эта мысль мне уже не нравится, поскольку эту непростую задачу надо решать не запросами, а отдельной программой.
Сделан скрипт **SetLanguageByCharCode.jsx**. С его помощью подготовка текстов к вёрстке должна стать более качественной.

Обновление 23.08.2024

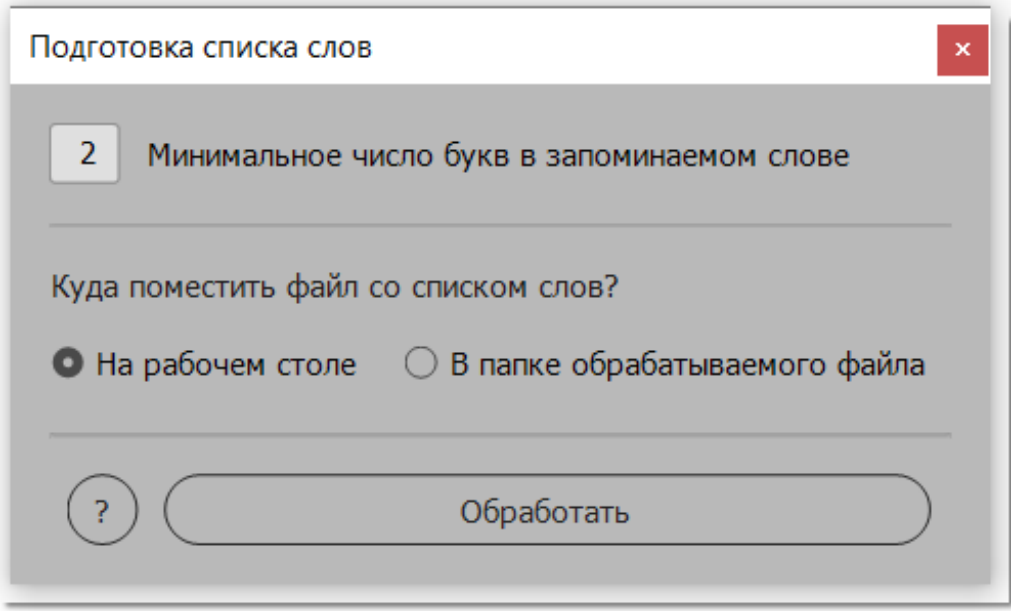
Убрана проверка, есть ли в работе вытесненный текст. Для слов, находящихся вне пределов фрейма, вместо номера страницы будет указываться прочерк. Так уже делается для слов, находящихся в фреймах на рабочем столе.

Михаил Иванюшин
https://dotextok.ru
dotextok@gmail.com

Подключение пользовательских словарей

Наверное, с этой темы надо было начать это руководство. Но эта программа появилась позже: тема необходимости иметь возможность исключить из области внимания спеллера какие-то слова, например, фамилии, возникла в переписке.

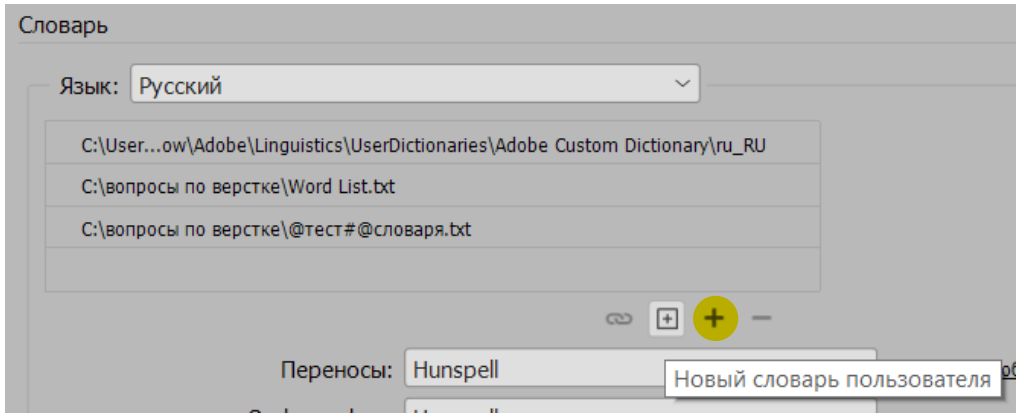
И раз такой запрос появился, почему бы не сделать программу. Собрать слова из файла — каждое слово в своей строке — это не сложно. Вот скрипт **WordsForDictionary.jsx** для решения этой задачи.



Как вариант, именно он и возник в переписке, можно собрать фамилии, чтобы спеллер не отмечал их как слова с ошибками. И поскольку нужны именно фамилии, то в этом скрипте есть возможность исключить из рассмотрения инициалы, достаточно в поле **Минимальное число букв в запоминаемом слове** ввести число 2.

И два места на выбор, где будет размещён txt-файл со словами: в папке обрабатываемого файла или на столе.

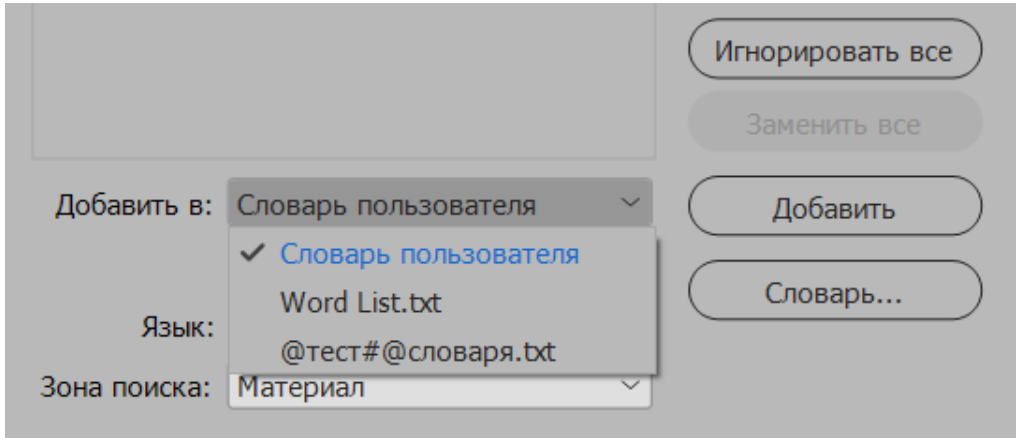
Итак, у вас есть один, а может быть, даже несколько файлов, о которых надо сообщить индизайну, что это словарные файлы. Это делается в панели Установок (Редактирование > Установки > Словарь). Вот этот знак плюс (отмечен жёлтым кружком) —



работает как кнопка выбора словарного файла. Вот тут уже выбрано два файла из папки C:\Вопросы по верстке.

Ну и соответственно кнопкой «минус» можно убирать из верстки подключенные ранее словари.

Подключенные словари отображаются в панели **Проверка орфографии** (Редактирование > Орфография > Проверка орфографии).



Первая строка в этом выпадающем меню **Словарь пользователя**. Совершенно топорное название, никак не отражающее назначение этой опции. Если пользователь загрузил только один словарь, то такое меню вроде как говорит, что будет работа со словарём пользователя. Но когда нажимаешь кнопку **Добавить**, красное подчёркивание спеллера на экране исчезает, то есть это слово уже не считается ошибочным, но в загруженном словаре вы его не увидите.

Вспоминаешь фразу Семёна Фарады из к/ф «Чародеи»: «Кто так строит?» Индизайн запоминает это слово, но *в своём файле для добавляемых пользователем слов*, если в том выпадающем списке конкретный пользовательский словарь не выбран. Если бы эта опция называлась, например, **Непомещённые в словарь слова**, это было бы ближе к тому, что реализуется данной опцией.

Что же это за файл, и где его искать? Его название **added.txt**, и на моей машине он размещается тут: C:\Users\Михаил\AppData\LocalLow\Adobe\Linguistics\UserDictionaries\Adobe Custom Dictionary\ru_RU, то есть в пользовательской области установок индизайна.

Итак, помещение выделенных спеллером слов выполняется в этот файл **added.txt**, если в выпадающем меню **Добавить в**: конкретный словарь не выбран. Когда выбран, то слова помещаются в указанный файл.

И есть ещё совершенно сбивающая с толку ситуация, когда содержимое словаря пользователя добавляется в этот **added.txt**. Меню Редактирование > Орфография > Словарь пользователя…, откроется окно панели **Словарь пользователя**, и там есть кнопка **Импортировать**. Вы импортируете какой-нибудь новый файл, его слова появляются в этой панели, и вы их увидите в файле **added.txt**. Но этот импортированный файл в списке словарей не появится!

Когда вы поймёте, что слова в файл **added.txt** были добавлены по ошибке, и удалите их, то в работе спеллера ничего не изменится. Похоже, эта информация ещё и кешируется, поэтому для возврата к исходному состоянию надо и перезагрузить индизайн.

Итог: если планируете работать со словарями, то важно ясно понимать, как их правильно подгружать. Индизайн сохраняет для обрабатываемого файла информацию, какие словари используются. С этим проблем не будет.

Полезно иметь ярлык, ведущий к этому служебному файлу **added.txt**. Надо быть уверенным, что оказавшиеся в нём слова действительно должны быть только там, а не в имеющихся словарях. Или переносить в словари оказавшиеся там слова.