

Урок 59. Распознавание документа

Приводить в порядок документ, распознанный в ABBYY FineReader – неблагоприятное занятие. Я дам несколько советов, которые помогут облегчить распознавание документа и дальнейшую работу с ним.

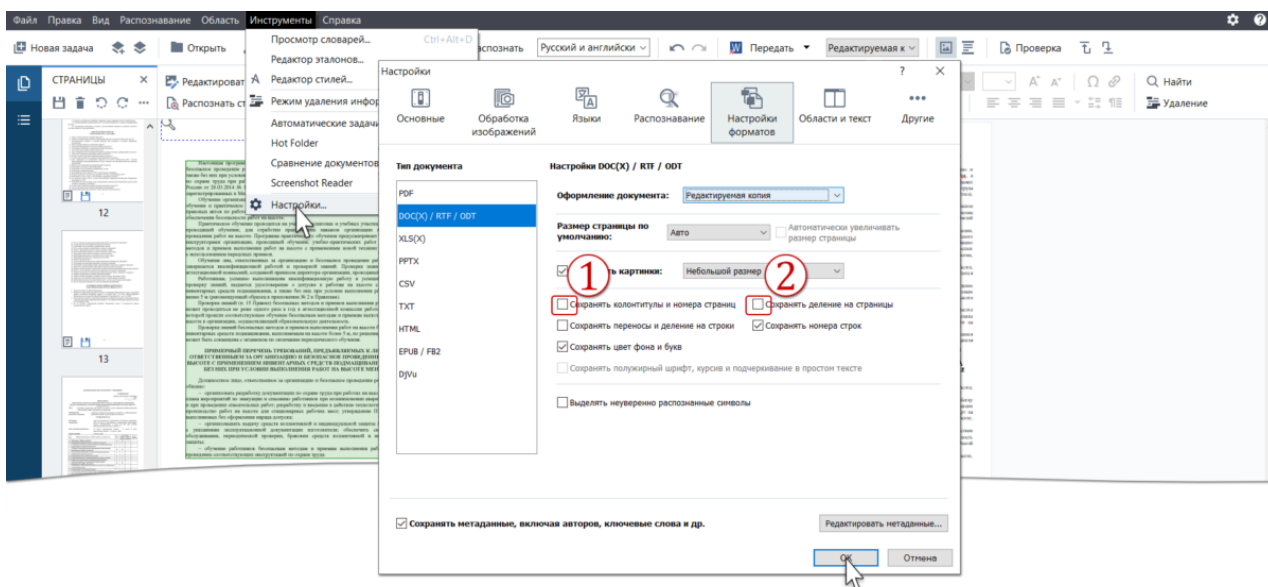
По окончании урока вы сможете:

1. Настроить ABBYY FineReader
2. Повторить алгоритм работы с распознанным документом

Откройте программу ABBYY FineReader. Подготовьте какой-нибудь документ *.pdf для распознавания.

1. Настройка ABBYY FineReader

Шаг 1. Настраиваем формат распознавания (лента Инструменты → команда Настройка → закладка Настройка форматов в диалоговом окне Настройка):

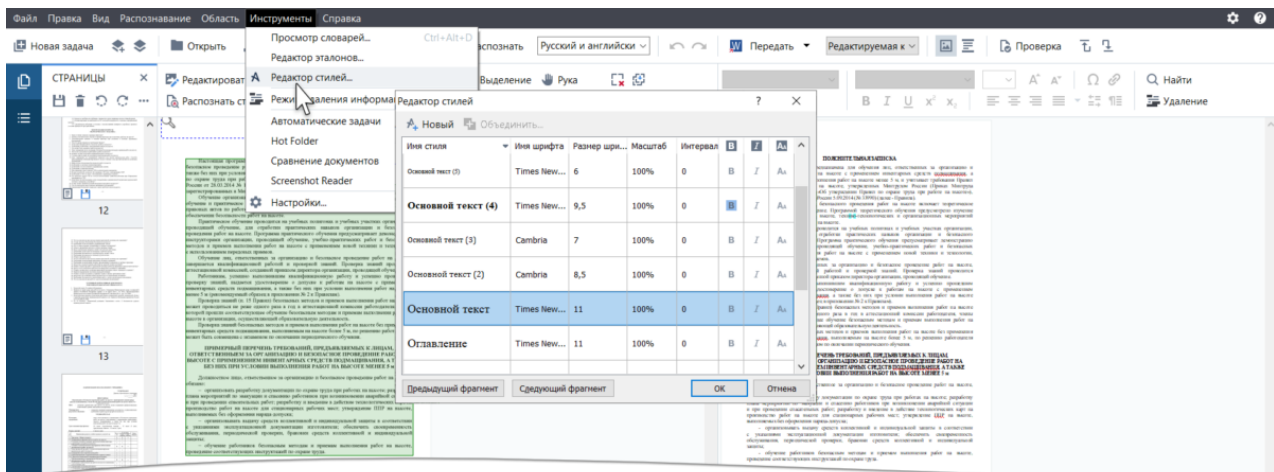


Снимаем галочки с команд:

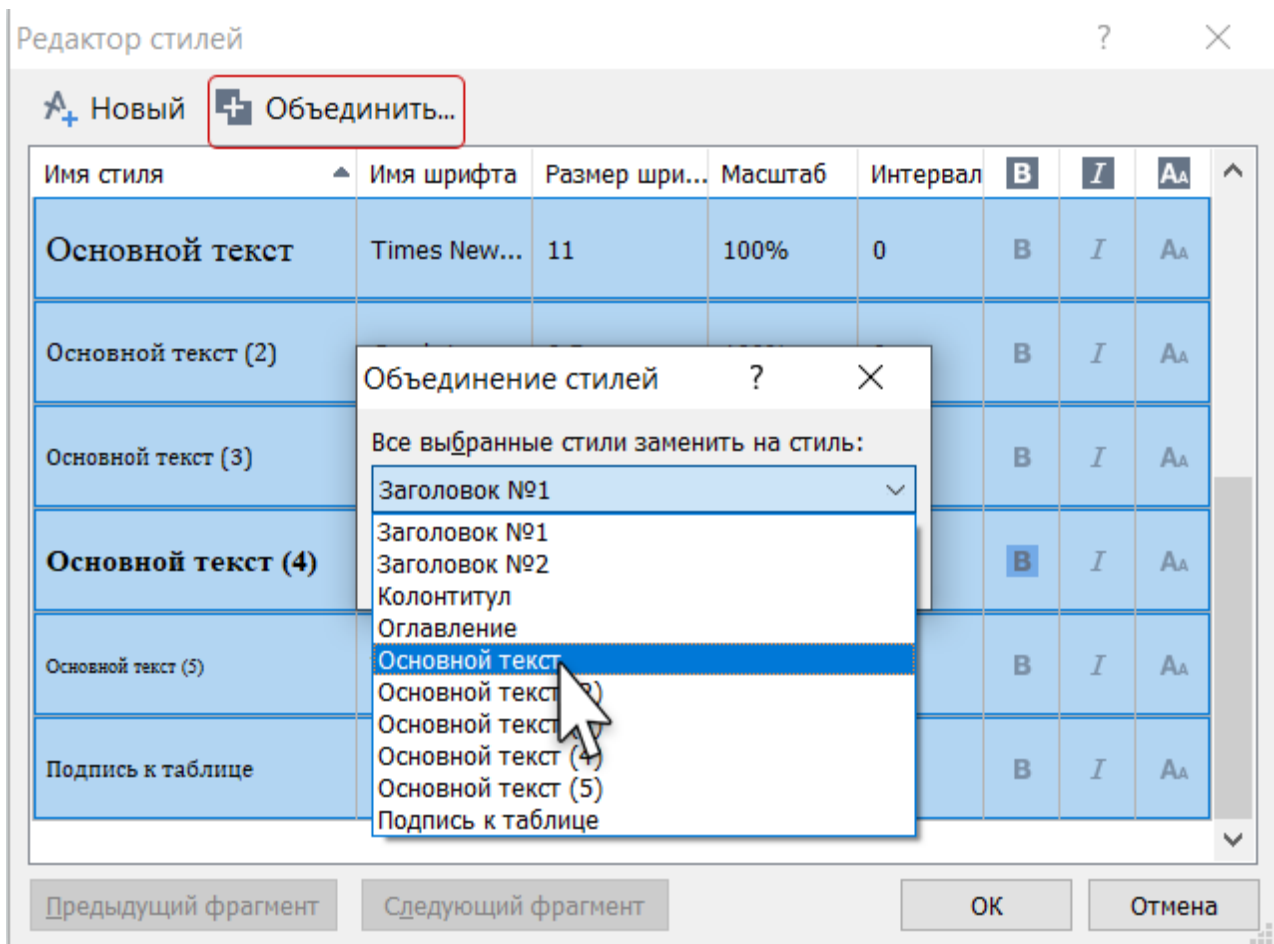
1. Сохранять колонтитулы и номера страниц
2. Сохранять деление на страницы

В самом деле: зачем нам номера страниц из исходного документа? Ещё неизвестно, в каком виде они распознаются. Бывает, что и виде графических объектов. А номера страниц расставит правильно. И от колонтитулов дополнительно избавляться не надо будет.

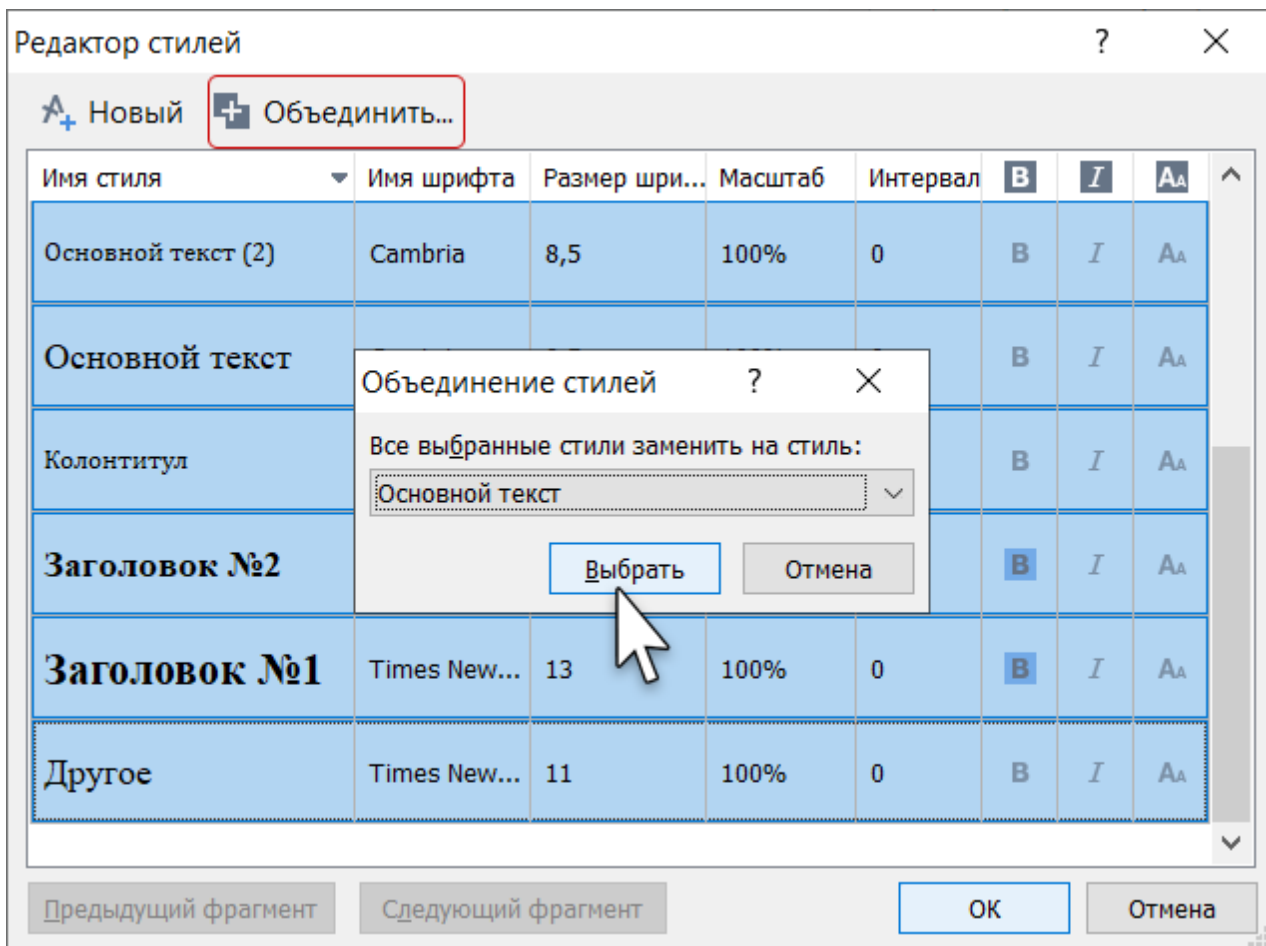
Шаг 2. Открываем окно «Редактор стилей» (лента Инструменты → команда Редактор стилей):



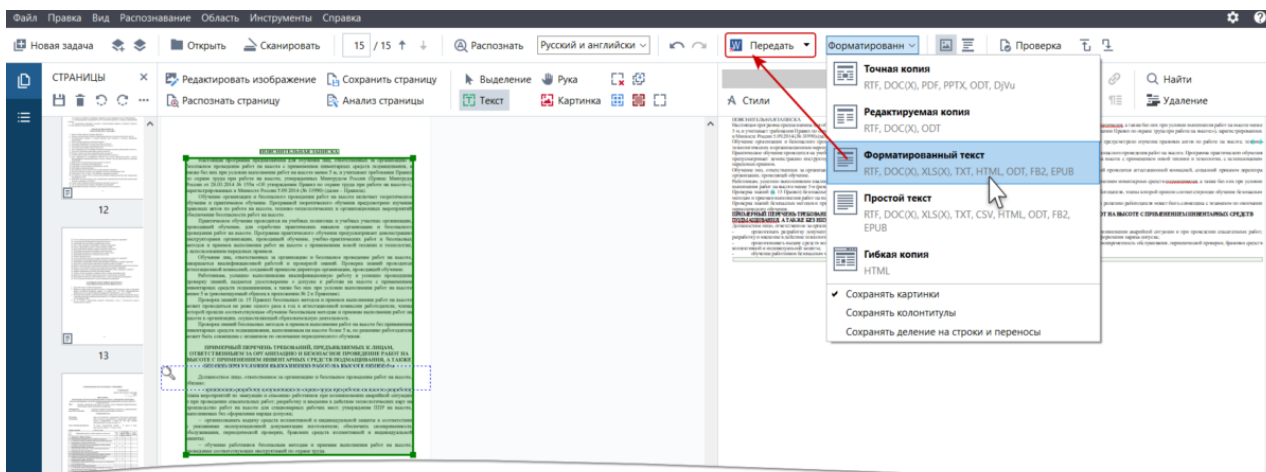
Шаг 3. Уничтожаем все стили (команда Объединить → кнопка выпадающего меню с выбор имени стиля):



Шаг 4. Выбираем имя стиля (любое на ваше усмотрение) → команда Выбрать:



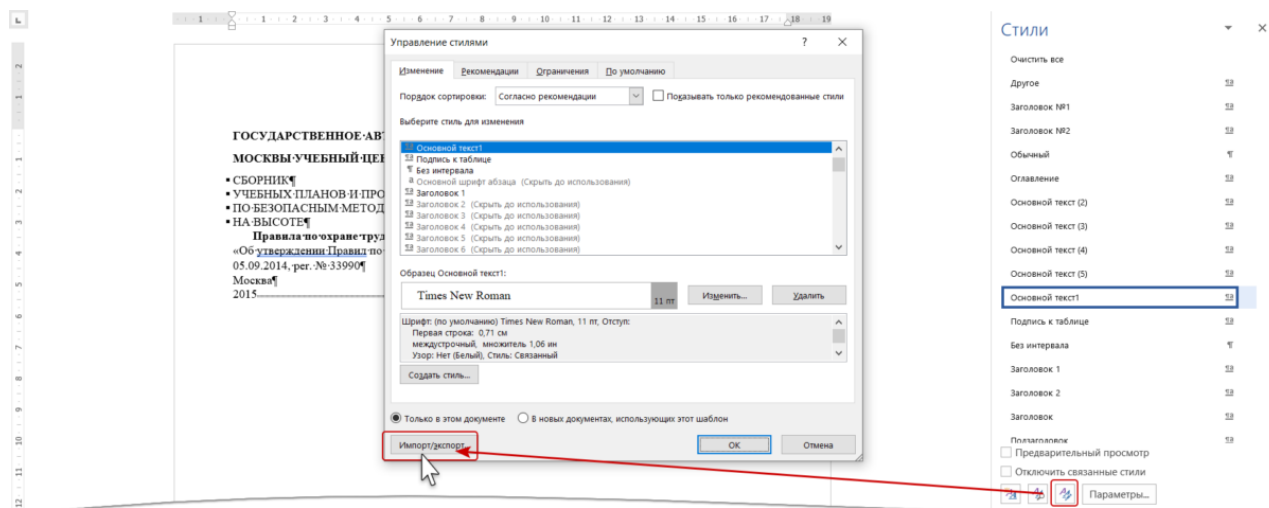
Шаг 5. Передаём текст в Word (выбрать вид Форматированный текст → команда Передать):



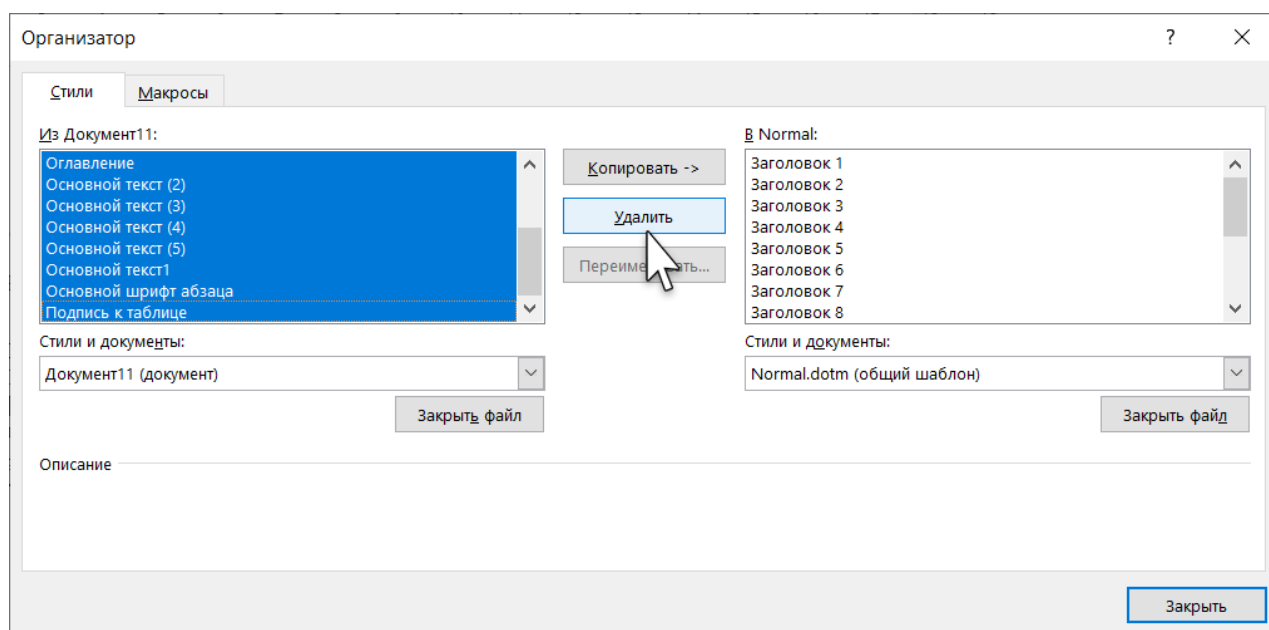
2. Алгоритм работы с распознанным документом

Смотрим на документ и видим кучу стилей. Из моего опыта: если не удалить стили в программе ABBYY FineReader, то удалить ненужные стили в программе Word будет проблематично. Но я работаю, далеко не в последней версии ABBYY FineReader. Вполне возможно, что в последних версиях такого казуса нет. Но не пожалейте время на 5-секундную операцию. А теперь по порядку.

Шаг 1. Открываем окно Импорт-экспорт (кнопка Инспектор стилей в рабочей области Стили → кнопка Импорт-экспорт):

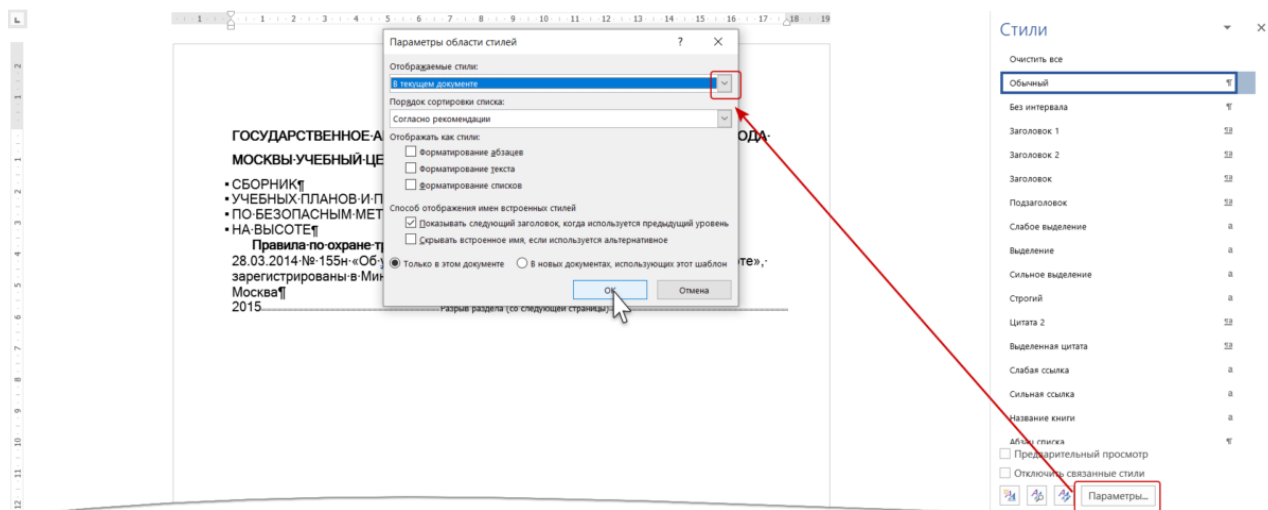


Шаг 2. Удаляем все стили (выделяем стили в окне Из Документа → кнопка Удалить → кнопка Заккрыть):

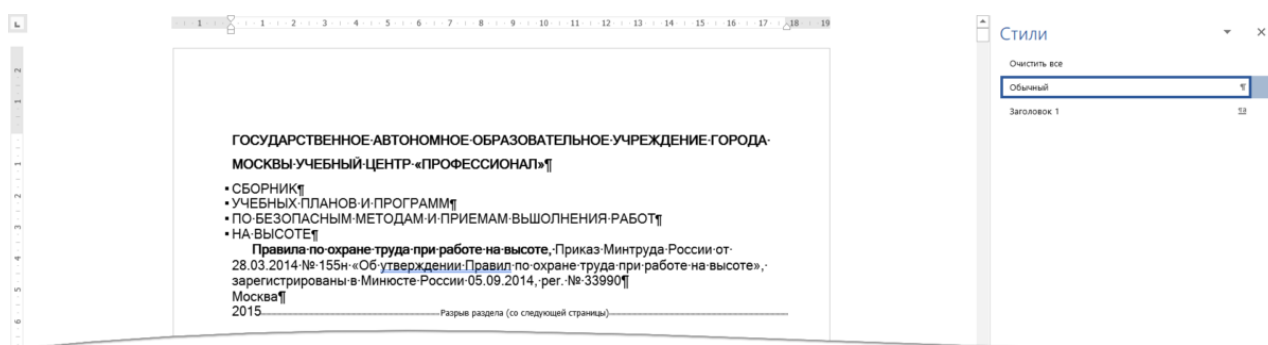


Если в рабочей области «Стили» перечень солидный, то скорее всего документ открыт в режиме «Рекомендованные стили» (спасибо разработчикам Word'a за заботу).

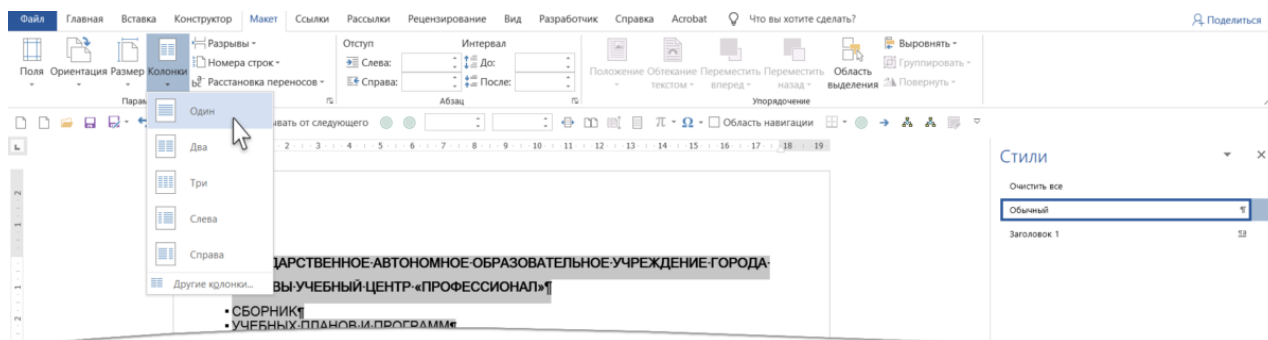
Шаг 3. Устанавливаем стили только для текущего документа (кнопка Параметры в рабочей области Стили → кнопка выпадающего меню в поле Отображаемые стили → режим В текущем документе):



Вот уже легче:



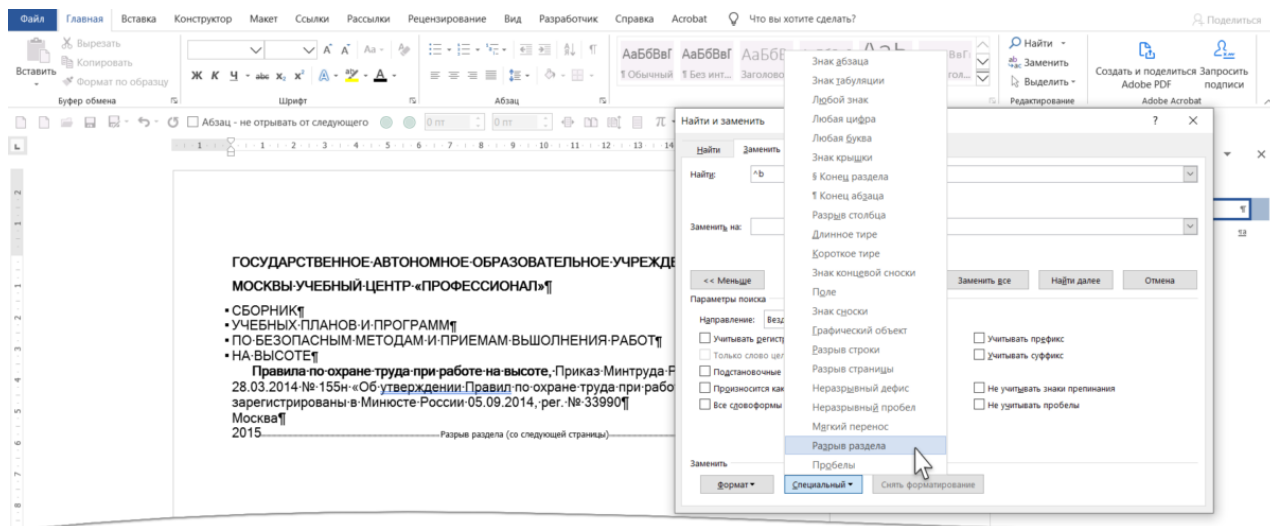
Шаг 4. Устанавливаем одноколоночный текст (выделяем весь текст **Ctrl+A** → лента Макет → команда Колонки → команда Одна колонка из выпадающего меню):



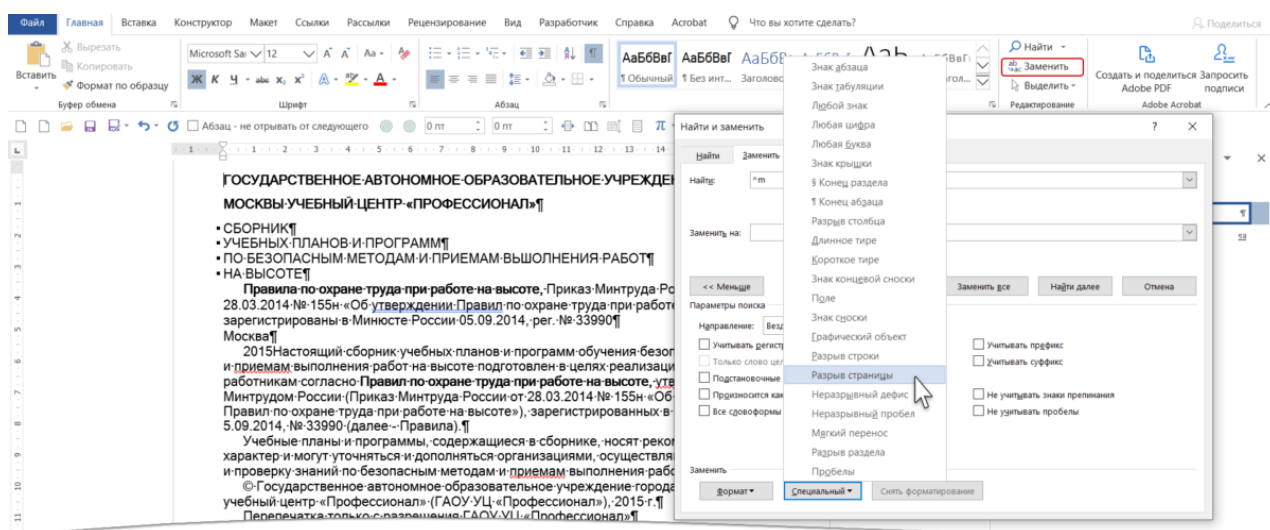
Конечно, может повезти, и многоколоночный текст не образуется при распознавании, но на всякий случай.

Шаг 5. Удаляем лишние непечатаемые символы.

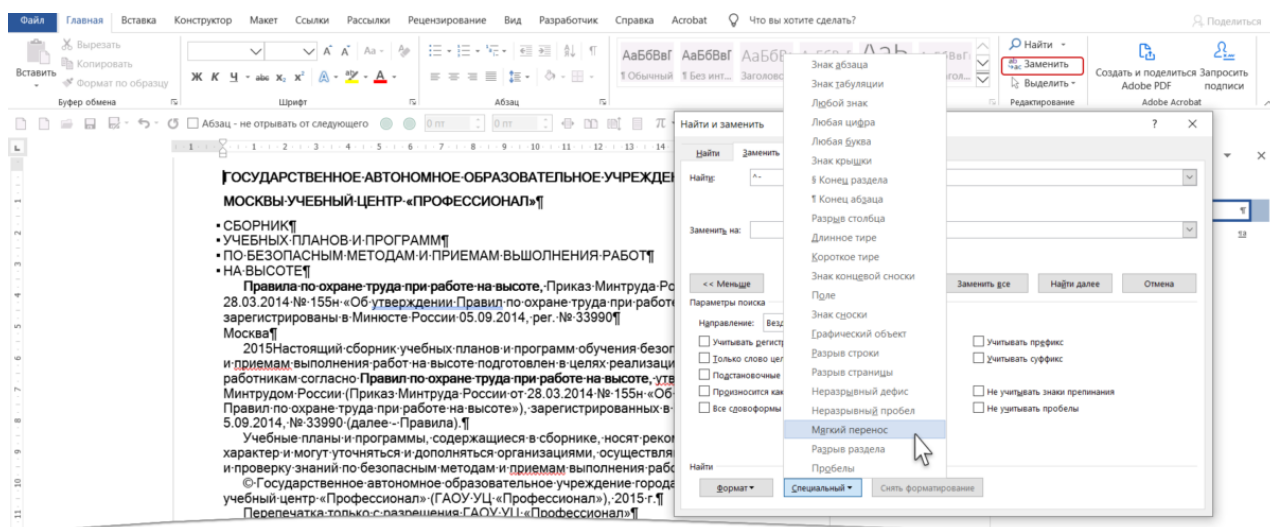
1. Разрывы разделов:



2. Разрывы страниц:



3. Мягкий перенос:

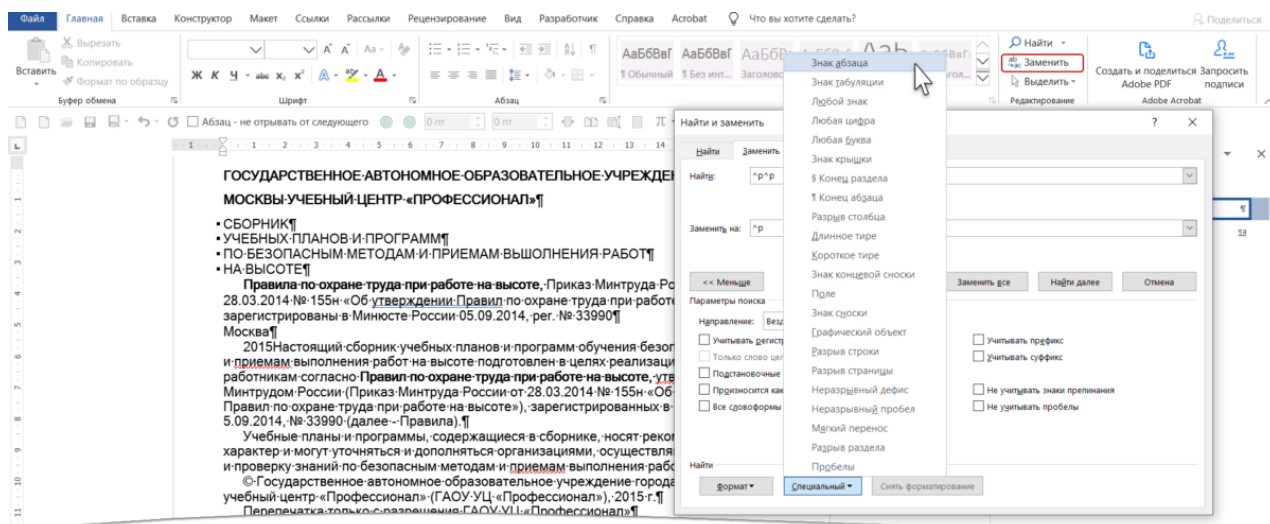


Непечатаемый символ «Мягкий перенос» образуется, если в распознаваемом документе были переносы

The screenshot shows the Microsoft Word 2010 interface. The 'Find and Replace' dialog box is open, with the 'Format' tab selected. The 'Find what' field contains the text 'ГОСУДАРСТВЕННО-АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ МОСКВЫ: УЧЕБНЫЙ ЦЕНТР «ПРОФЕССИОНАЛ»'. The 'Replace with' field is empty. The 'Format' dropdown is set to 'Специальный' (Special). The 'Find all' button is highlighted. The background document text is partially visible, showing a list of documents and a section titled 'Правила по охране труда при работе на высоте'.

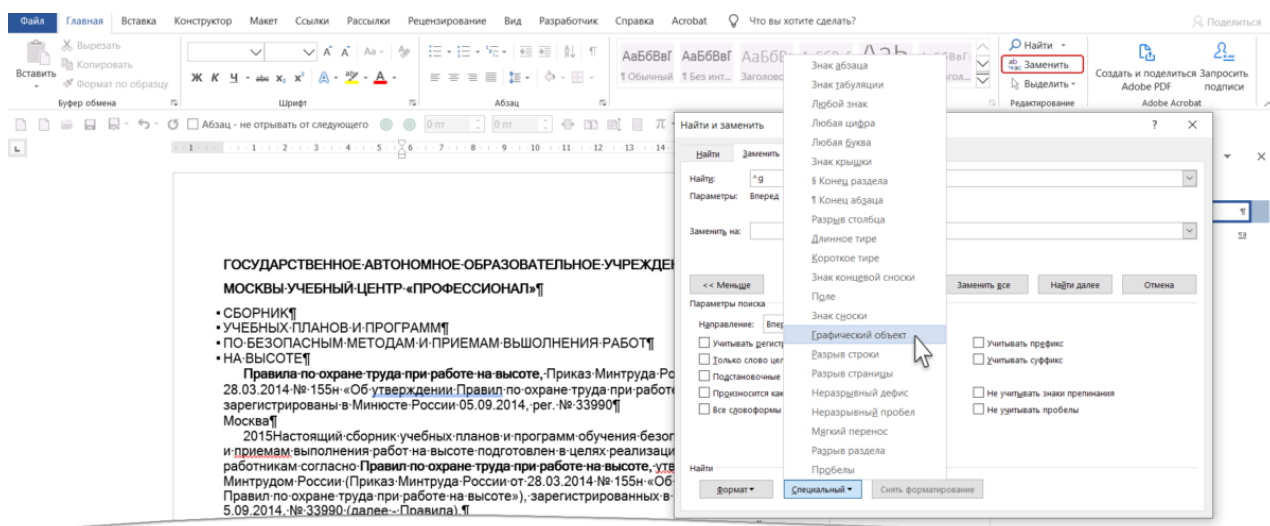
The image shows the 'Find and Replace' dialog box in Microsoft Word. The title bar reads 'Найти и заменить' (Find and Replace). There are three tabs: 'Найти' (Find), 'Заменить' (Replace), and 'Перейти' (Go to), with 'Найти' currently selected. The 'Найти:' (Find what) field contains the text 'Два пробела' (Two spaces) in red. The 'Заменить на:' (Replace with) field contains 'Один пробел' (One space) in red. Below these fields are four buttons: '<< Меньше' (Previous), 'Заменить' (Replace), 'Заменить все' (Replace all), and 'Найти далее' (Find next), with the last button highlighted by a blue border. A section titled 'Параметры поиска' (Search parameters) includes a 'Направление:' (Direction) dropdown set to 'Везде' (Anywhere). Below this are two columns of checkboxes: 'Учитывать регистр' (Match case), 'Только слово целиком' (Match whole word), 'Подстановочные знаки' (Match wildcards), 'Произносится как' (Match pronunciation), 'Все словоформы' (Match all word forms), 'Учитывать префикс' (Match prefixes), 'Учитывать суффикс' (Match suffixes), 'Не учитывать знаки препинания' (Ignore punctuation), and 'Не учитывать пробелы' (Ignore spaces). At the bottom, under the 'Заменить' (Replace) section, there are three buttons: 'Формат' (Format), 'Специальный' (Special), and 'Снять форматирование' (Remove formatting).

6. Два символа конца абзаца на один символ конца абзаца:



Вполне возможно, что вам придётся повторить эту операцию несколько раз аналогично предыдущей операции.

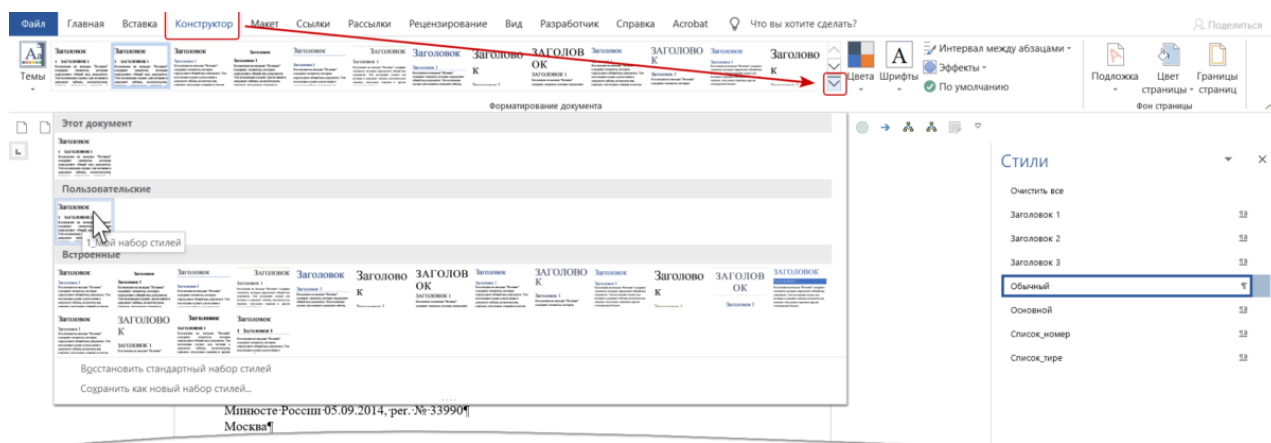
7. Удаление графических объектов (хорошо, что разработчики Word предусмотрели эту возможность):



В таблице я показывала, что на что надо менять:

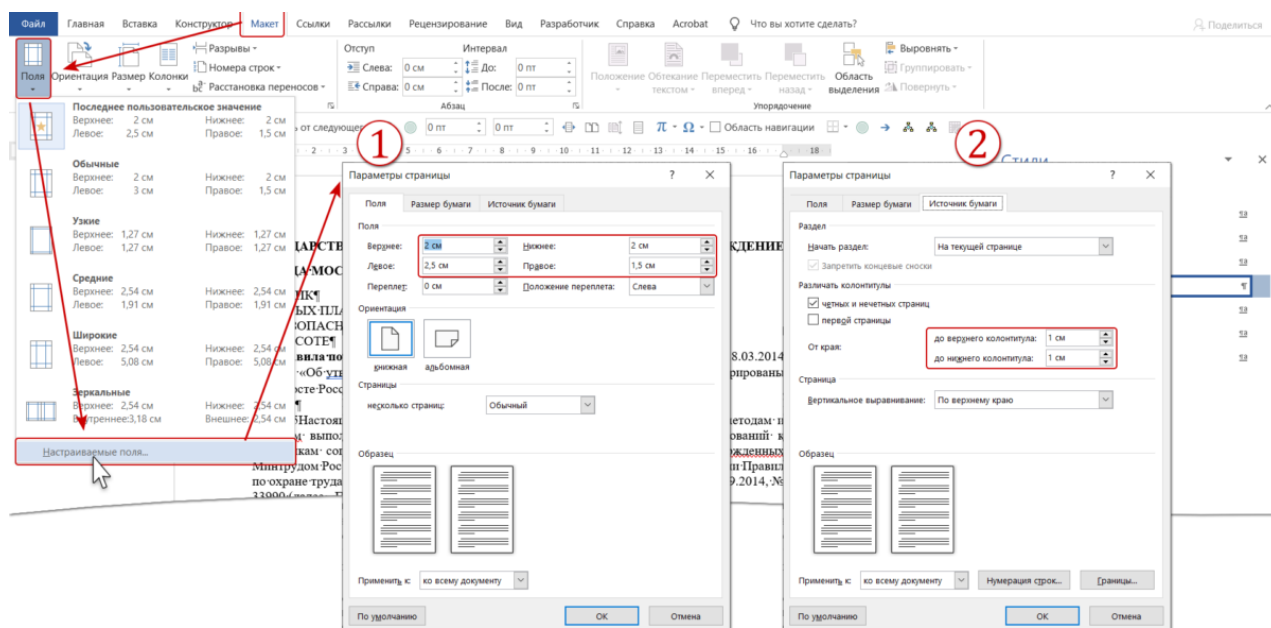
Поле «Найти»	Поле «Заменить на...»
1 Разрыв раздела	Пусто
2 Разрыв страницы	Пусто
3 Мягкий перенос	Пусто
4 Знак табулятора	Пробел
5 Два пробела	Пробел
6 Два символа конца абзаца	Один символ конца абзаца
7. Графический объект	Пусто

Шаг 6. Применение набора стилей (лента Конструктор → группа команд Форматирование документа → кнопка выпадающего меню → пользовательский набор стилей):



Шаг 7. Установка параметров границы печатного поля (лента Макет → группа команд Параметры страницы → команда Поля → команда Настраиваемые поля из выпадающего меню):

1. Границы печатного поля для страницы
2. Положение колонтитулов



Всё, документ готов к форматированию.

Теперь вы сможете:

1. Настроить ABBYY FineReader
2. Повторить алгоритм работы с распознанным документом

Ещё остались таблицы. Если в документе не больше трёх таблиц, то нетрудно отформатировать их вручную. А если 50 таблиц? Вот об этом будет следующий урок. Заодно состоится первое знакомство макросами.

