Урок 59. Распознавание документа

prooffice24.ru/document-recognition

5 мая 2020 г.

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

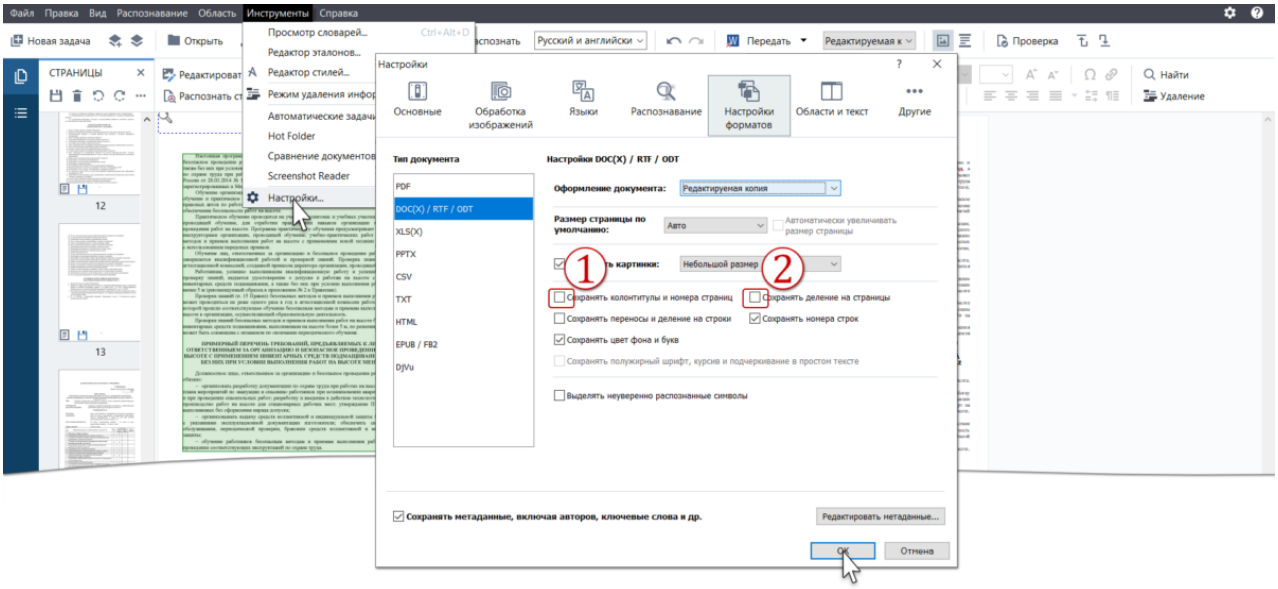
По окончании урока вы сможете:

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

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

1. Настройка ABBYY FineReade

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

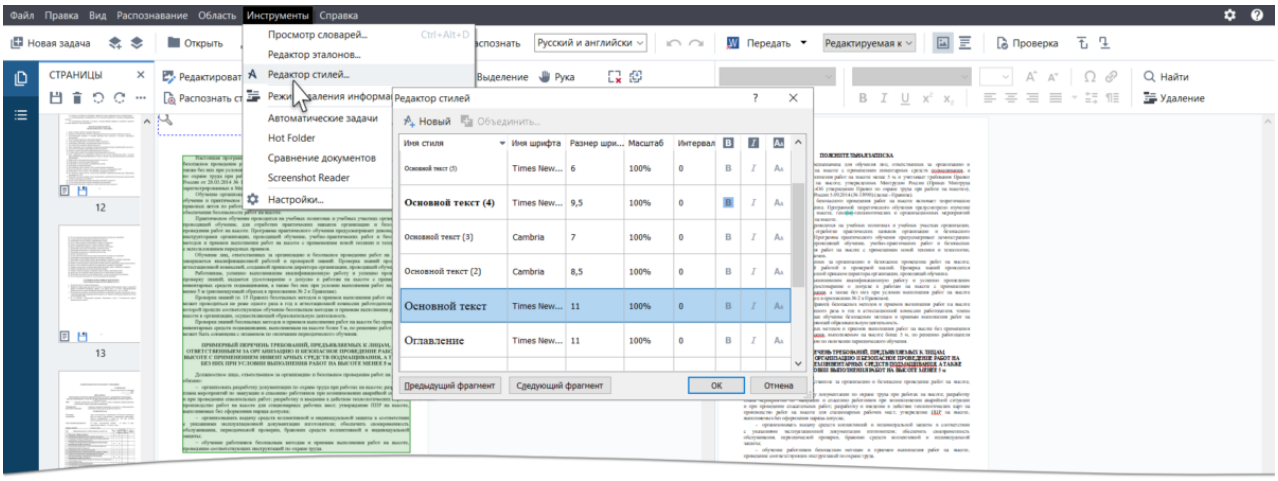


Снимаем галочки с команд:

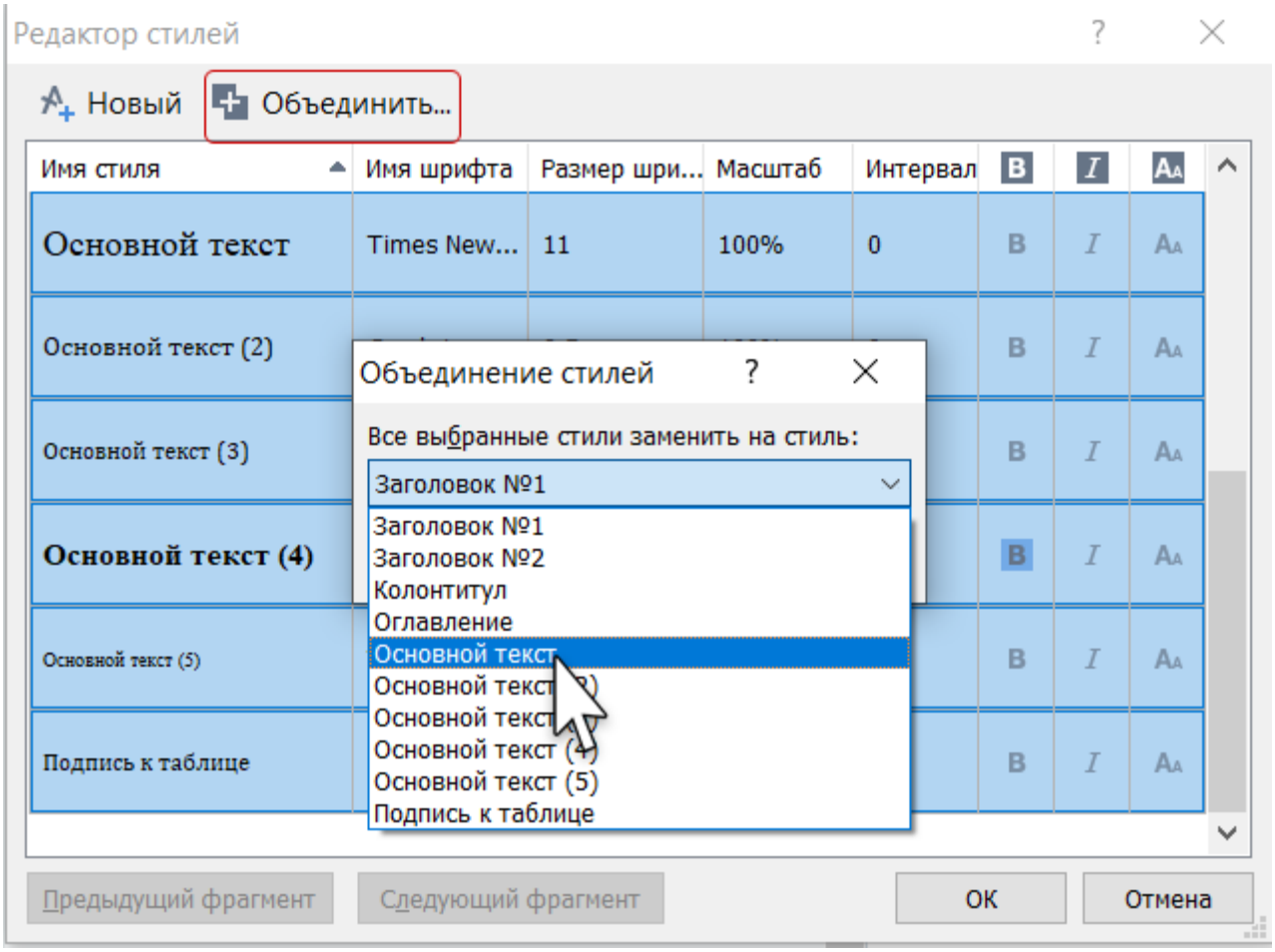
1. Сохранять колонтитулы и номера страниц
2. Сохранять деление на страницы

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.

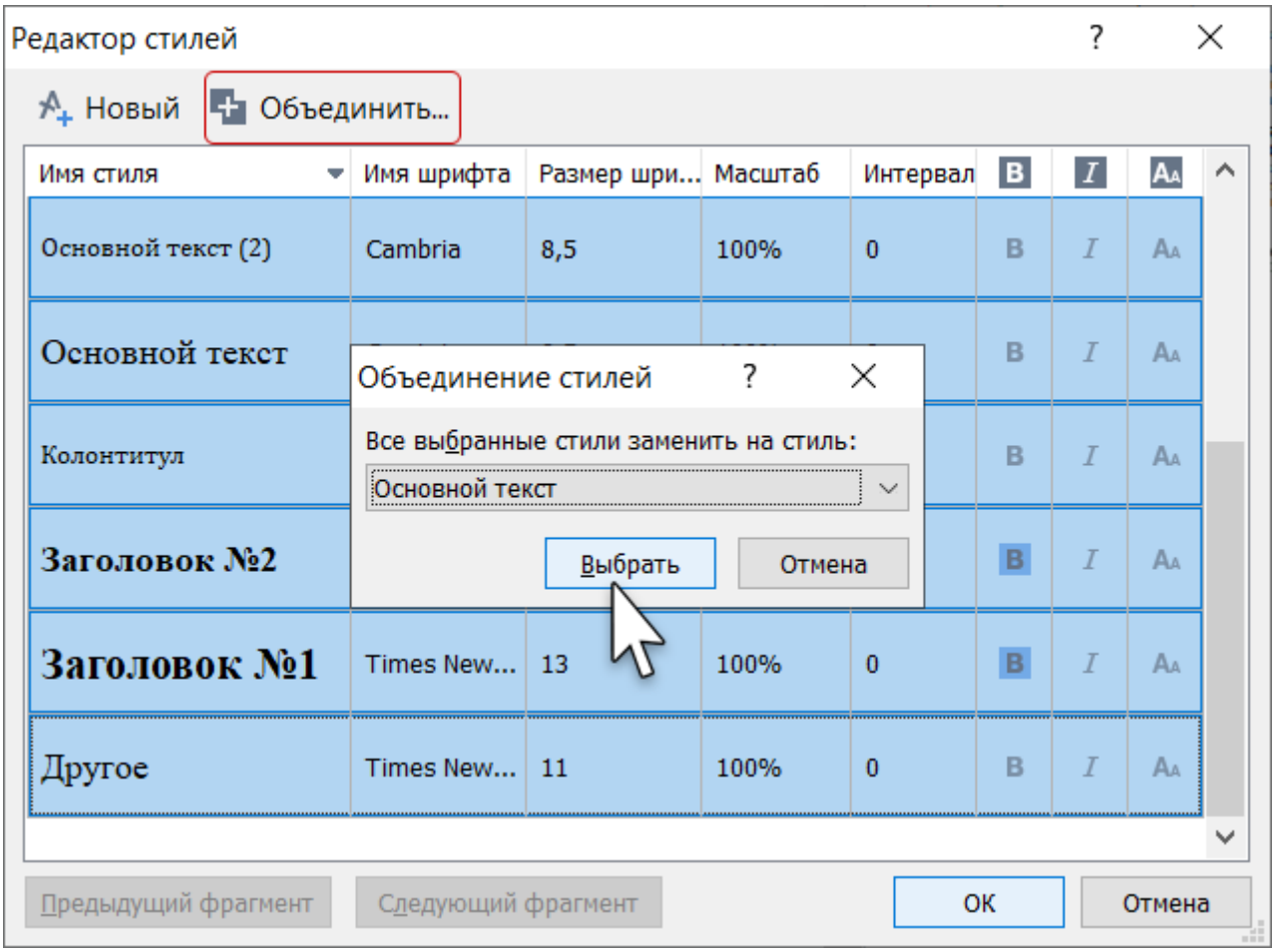
Шаг 2. Открываем окно «Редактор стилей» (лента Инструменты → команда Редактор стилей»):



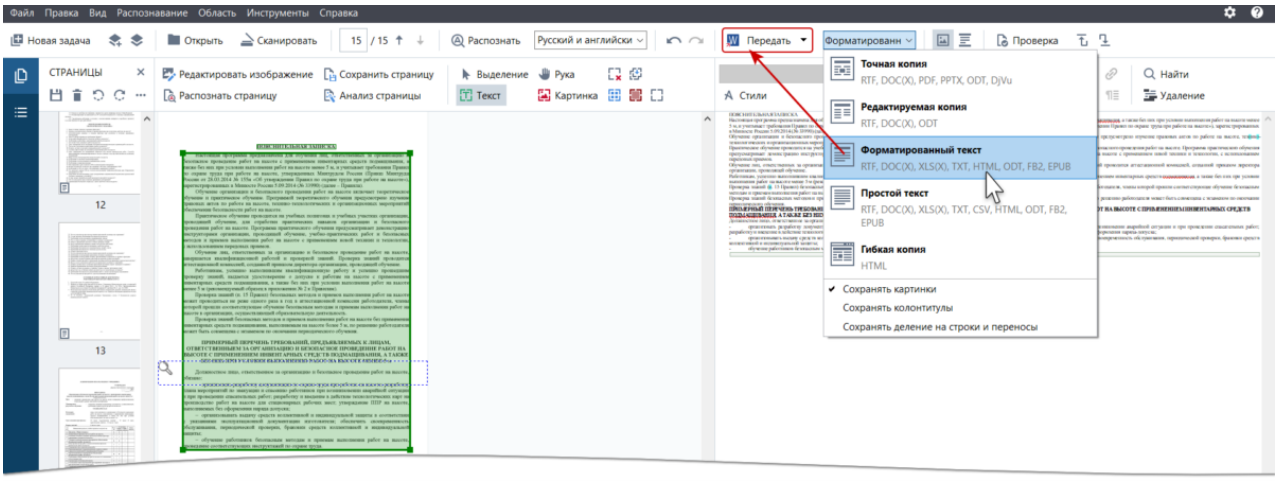
Шаг 3. Уничтожаем все стили (команда Объединить → кнопка выпадающего меню с выбор имени стиля):



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



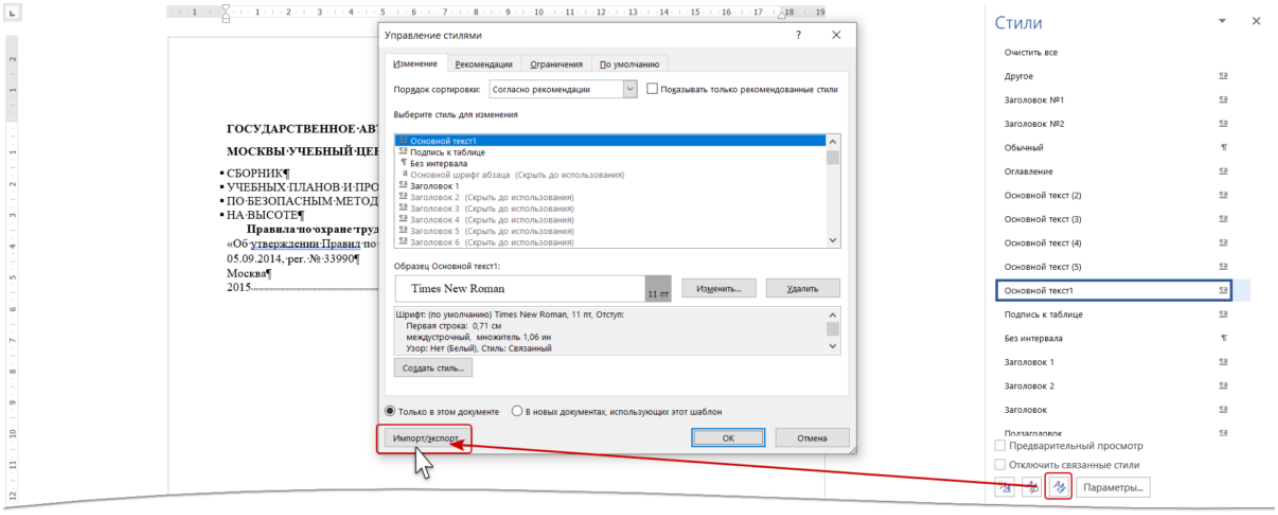
Шаг 5. Передаём текст в Word (выбрать вид Форматированный текст → команда Передать):



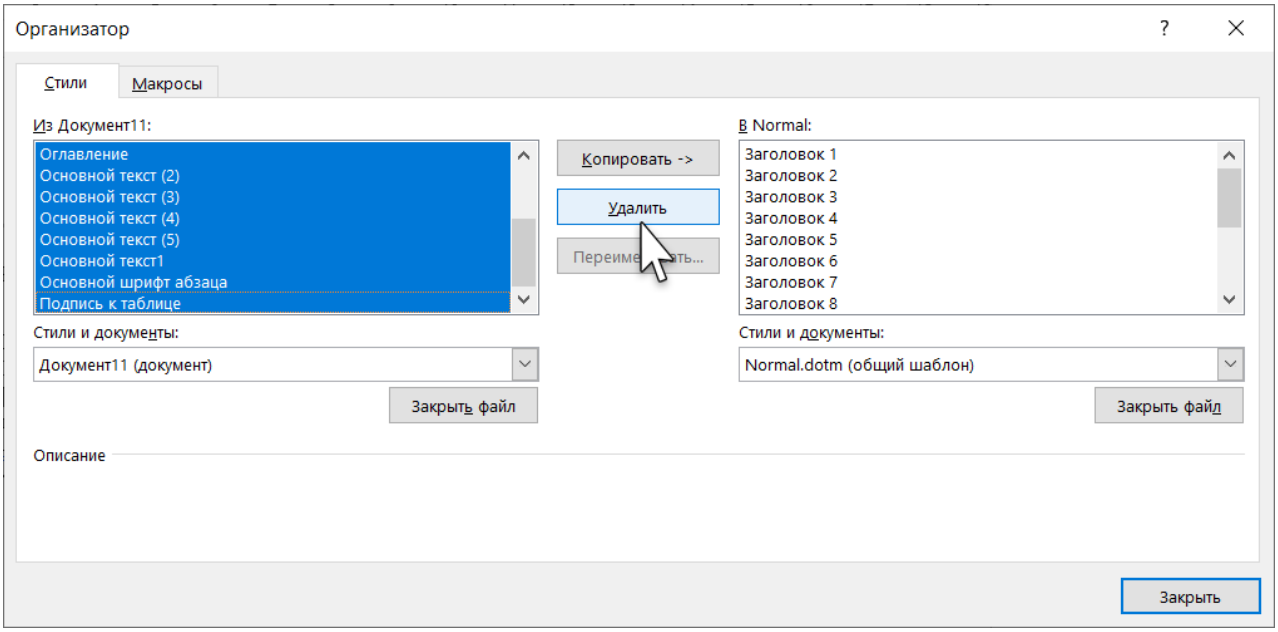
2. Алгоритм работы с распознанным документом

Смотрим на документ и видим кучу стилей. Из моего опыта: если не удалить стили в программе ABBYY FineReade, то удалить ненужные стили в программе Word будет проблематично. Но я работаю, далеко не в последней версии ABBYY FineReade. 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):

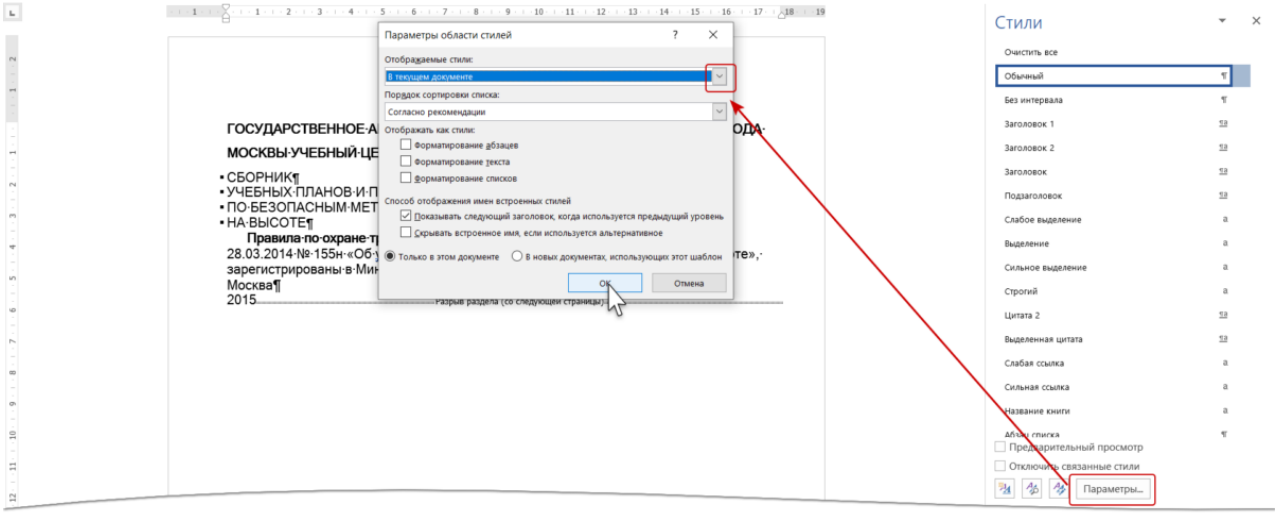


Шаг 2. Удаляем все стили (выделяем стили в окне Из Документа → кнопка Удалить → кнопка Закрыть):

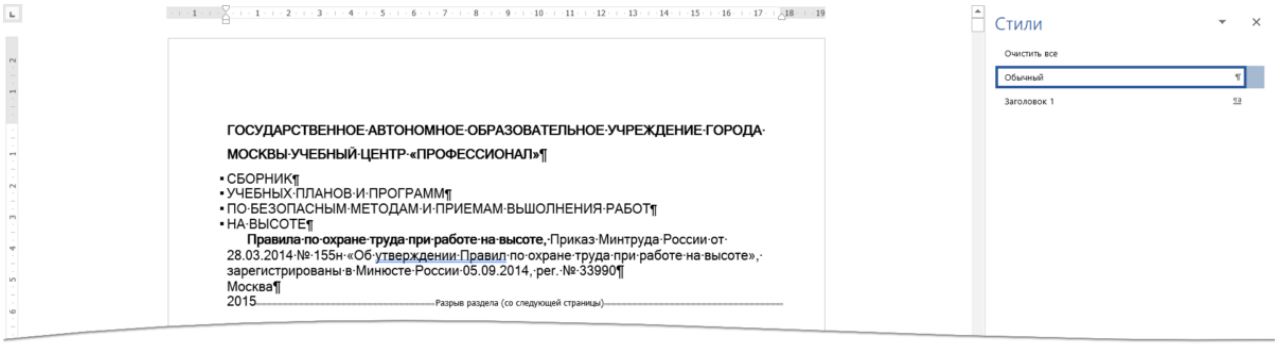


Если в рабочей области «Стили» перечень солидный, то скорее всего документ открыт в режиме «Рекомендованные стили» (спасибо разработчикам Word’а за заботу).

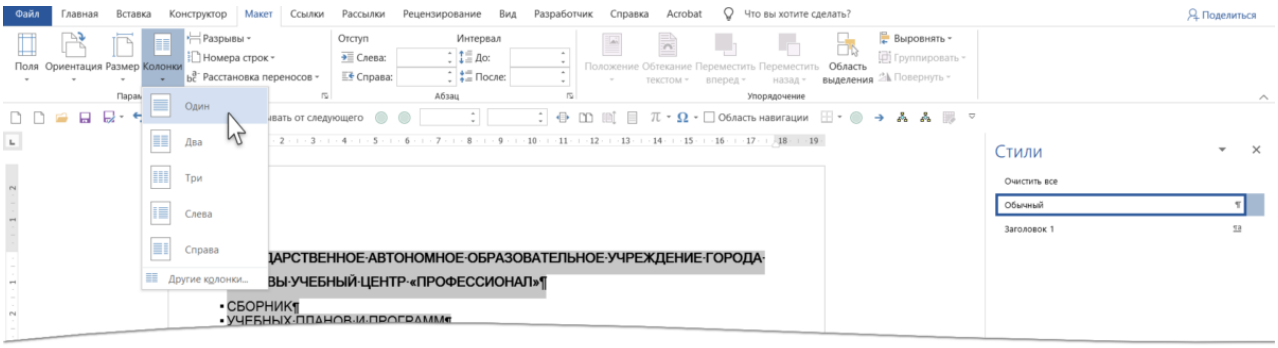
Шаг 3. Устанавливаем стили только для текущего документа (кнопка Параметры в рабочей области Стили → кнопка выпадающего меню в поле Отображаемые стили → режим В текущем документе):



Вот уже легче:



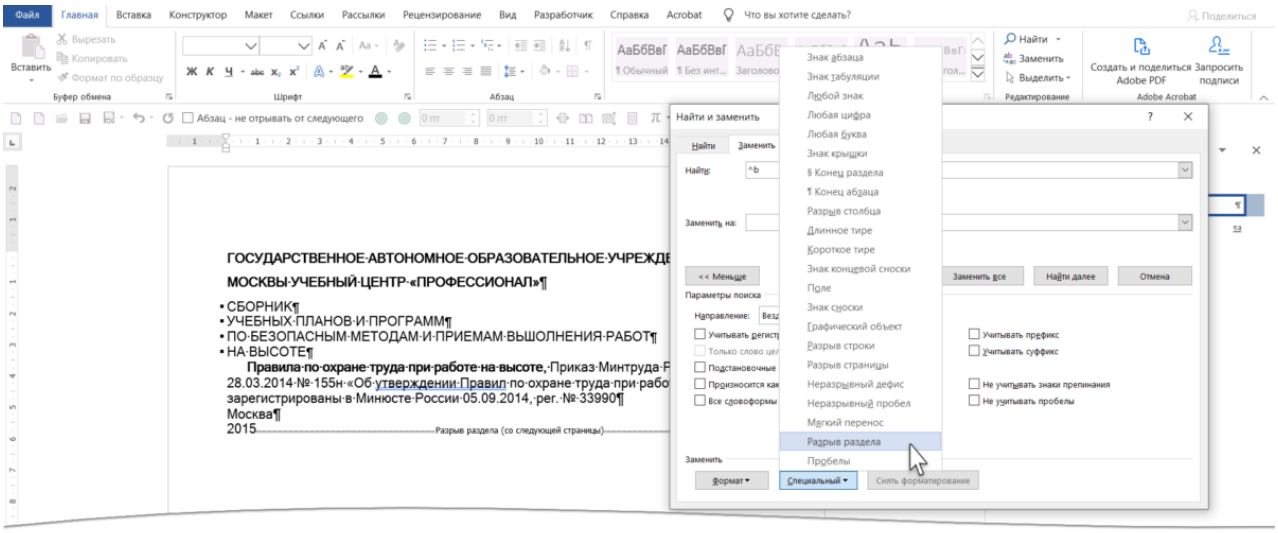
Шаг 4. Устанавливаем одноколоночный текст (выделяем весь текст **Ctrl+A** → лента Макет → команда Колонки → команда Одна колонка из выпадающего меню):



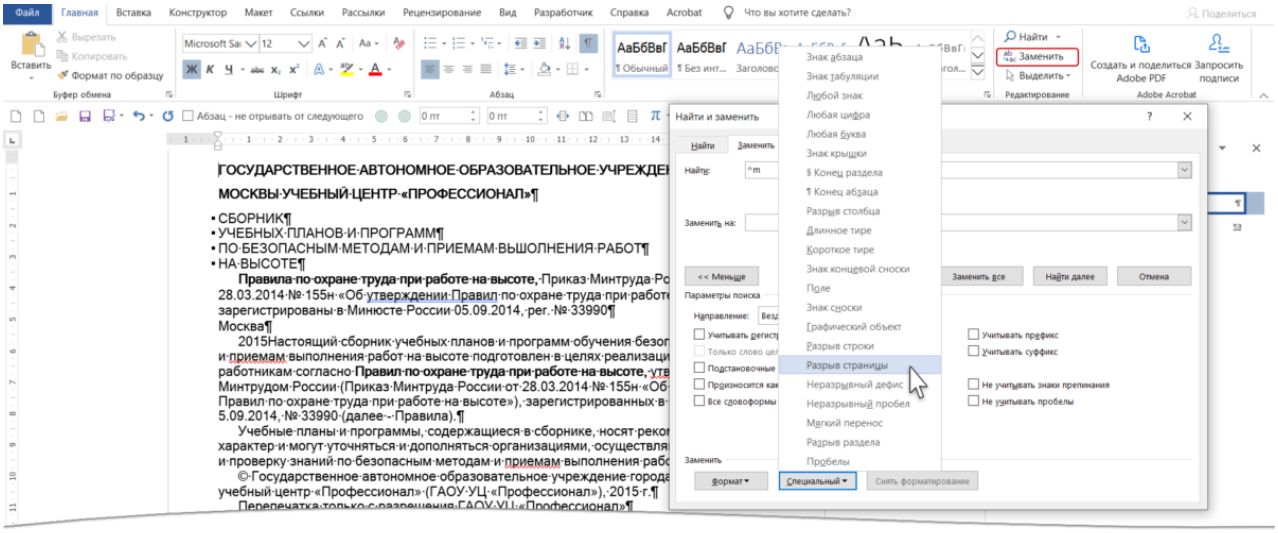
Конечно, может повезти, и многоколоночный текст не образуется при распознавании, но на всякий случай.

Шаг 5. Удаляем лишние непечатаемые символы.

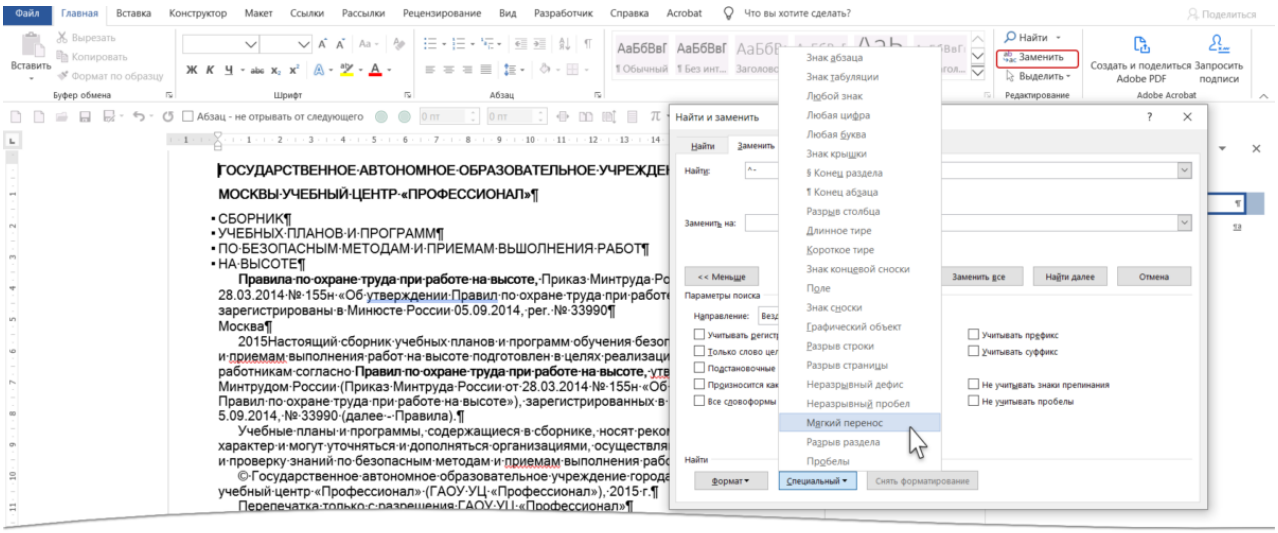
1. Разрывы разделов:



2. Разрывы страниц:

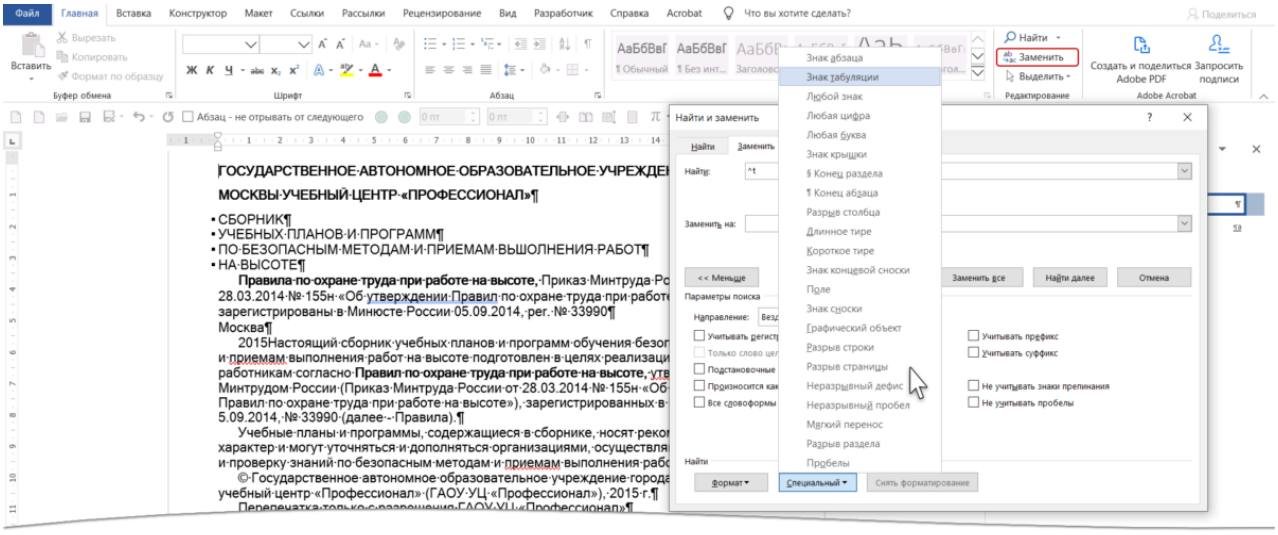


3. Мягкий перенос:

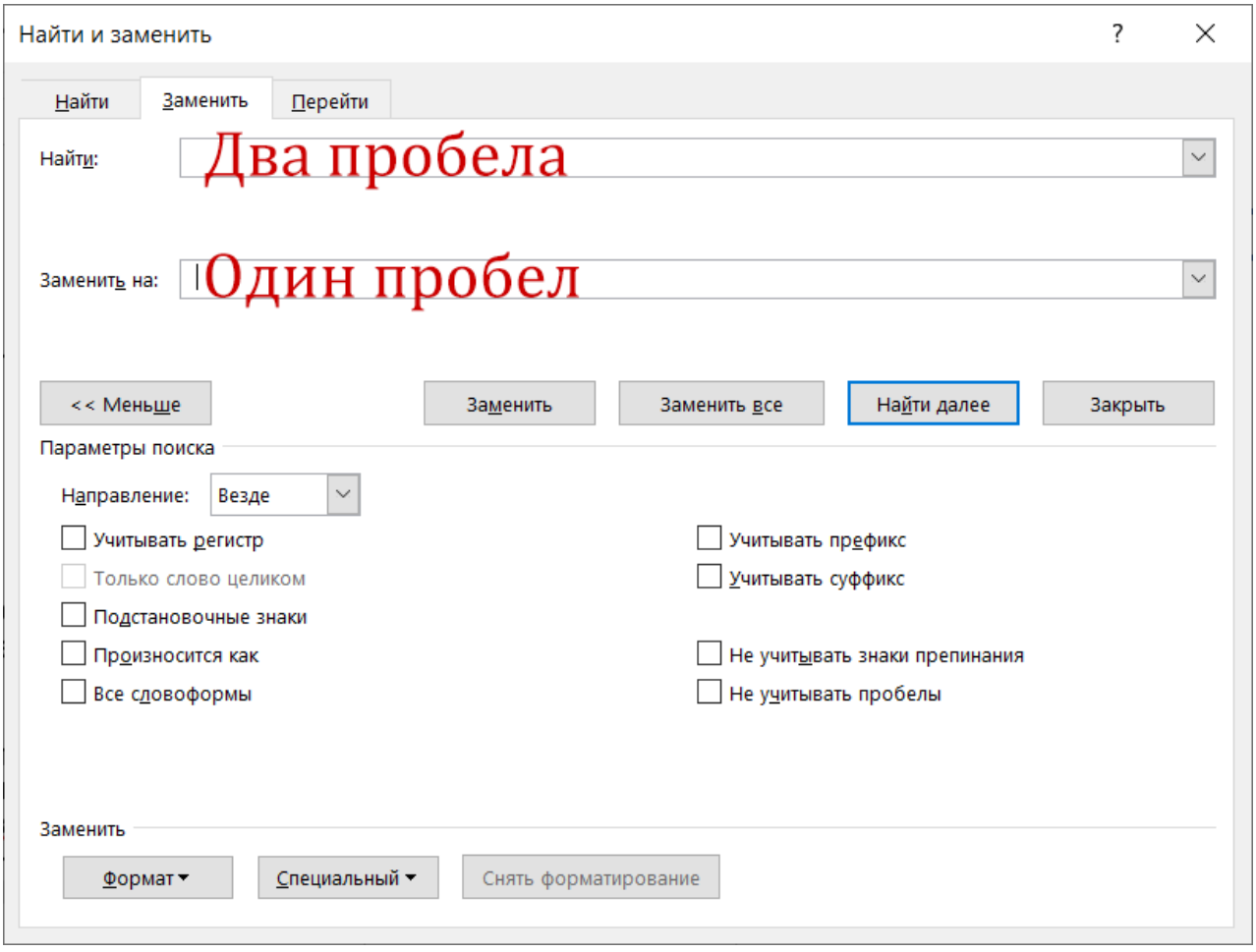


Непечатаемый символ «Мягкий перенос» образуется, если в распознаваемом документе были переносы

4. Знак табулятора на пробел:

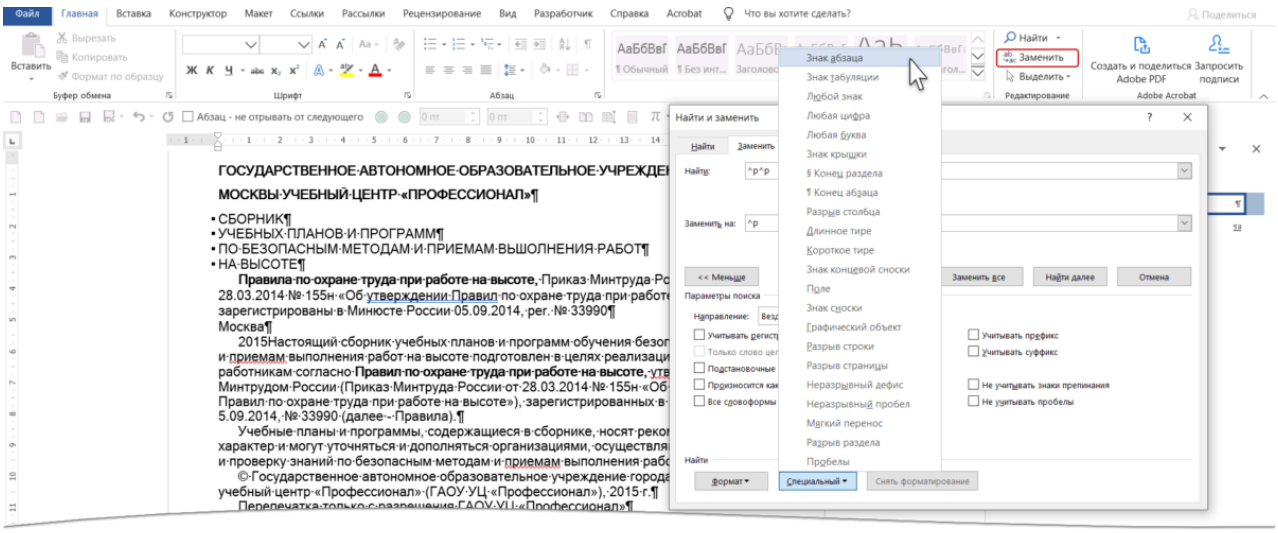


5. Два пробела на пробел



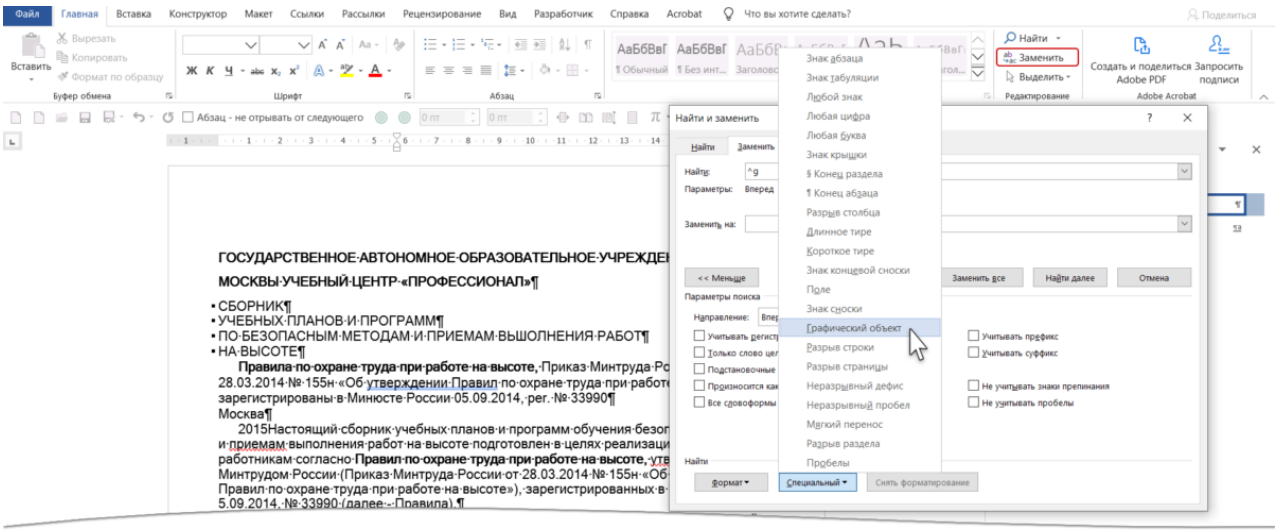
Вполне возможно, что вам придётся повторить эту операцию несколько раз, так как при выравнивании текста по ширине в распознаваемом документе, могут быть большие промежутки между словами. ABBYY FineReade заботливо заполняет эти промежутки большим количеством пробелов.

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



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

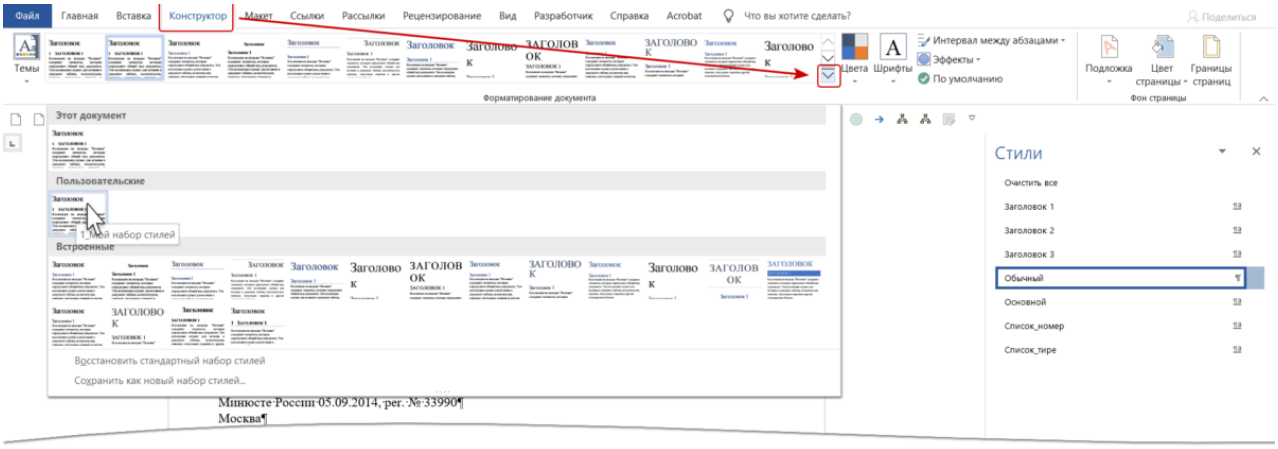
7. Удаление графических объектов (хорошо, что разработчики Word предусмотрели эту возможность):



В таблице я показывала, что на что надо менять:

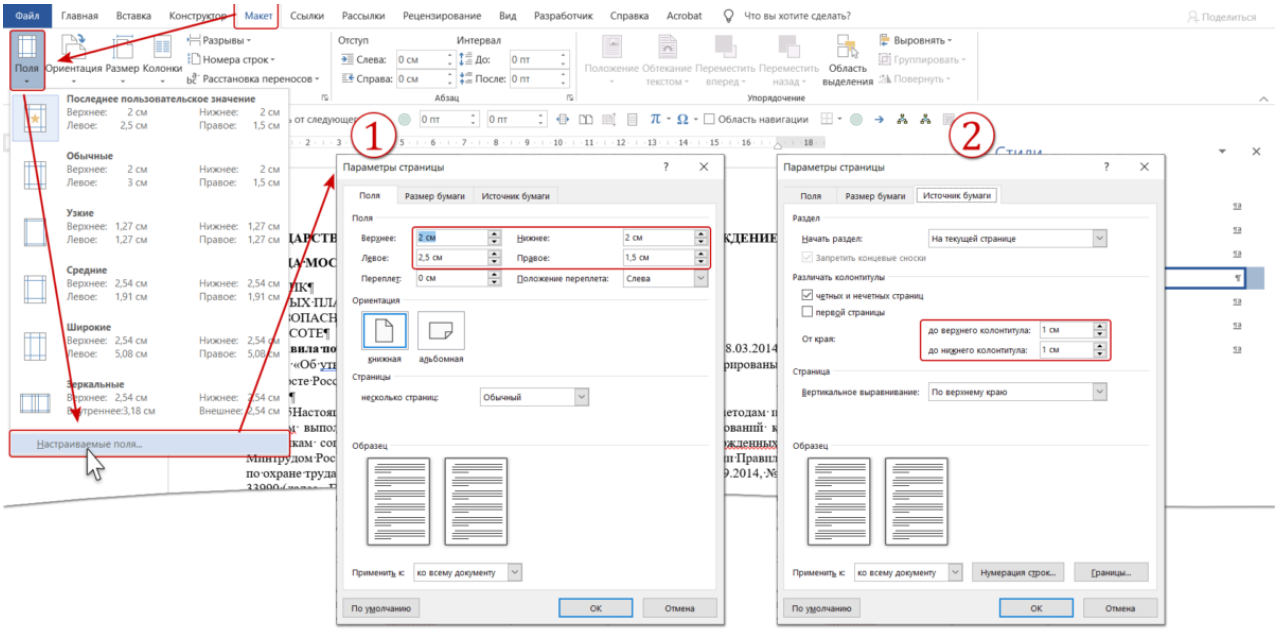
	Поле «Найти»	**Поле «Заменить на…»**
1	Разрыв раздела	Пусто
2	Разрыв страницы	Пусто
3	Мягкий перенос	Пусто
4	Знак табулятора	Пробел
5	Два пробела	Пробел
6	Два символа конца абзаца	Один символ конца абзаца
7.	Графический объект	Пусто

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 FineReade
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.