Жадаев А. Г.

Сканирование и распознавание текстов

Коллекция программ ABBYY® в подарок*

САМОУЧИТЕЛЬ
по работе с программой

ABBYY® FineReader 10

ДМК ПРЕСС
ИЗДАТЕЛЬСТВО

ABBYY® PRESS

Жадаев А. Г.

Сканирование и распознавание текстов

Самоучитель по работе с ABBYY® FineReader 10

ДМК Пресс
издательство
Москва, 2010

УДК 32.973.26-018.2
ББК 004.4
Ж15

Ж15 **Жадаев А. Г.**

Сканирование и распознавание текстов. Самоучитель по работе с ABBYY® FineReader 10. – М.: ДМК Пресс, 2010. – 248 с.: ил.

ISBN 978-5-94074-595-2

Работать с электронными документами во многом удобнее и проще, чем с их бумажными аналогами. Электронный документ можно редактировать, использовать при создании собственных работ, его легко копировать и пересылать по электронной почте. Вместе с тем, многие материалы изначально доступны нам в нередактируемом виде (бумажные или отсканированные документы, цифровые фотографии). Программа ABBYY® FineReader – лучший инструмент для создания электронных копий любых печатных материалов: книг, справочников, журналов, договоров, бланков.

Книга включает описание приемов сканирования и распознавания разных оригиналов – от простых книжных страниц до сложно оформленных документов. А приведенные скриншоты программы позволят читателю быстро освоить интерфейс ABBYY® FineReader и получить практические навыки по работе с программой.

Изложение материала сопровождается практическими примерами. Читатели, которые еще не пробовали самостоятельно переводить печатные материалы в электронный вид, найдут в этой книге простое пошаговое руководство. Для тех же, кто хочет в совершенстве освоить работу с программой, книга откроет многочисленные тонкости настройки для эффективного использования ABBYY® FineReader.

УДК 32.973.26-018.2
ББК 004.4

ISBN 978-5-94074-595-2

Содержание

Глава 1 Текст и графика в компьютере

Любая информация в компьютере хранится и обрабатывается в цифровом виде. Хотя в этой книге повсюду упоминаются слова «текст» и «графика», на самом деле компьютер работает исключительно с цифровой информацией. Любая информация в компьютере хранится в виде файлов, а файл – всего лишь последовательность двоичных чисел (единиц и нулей). С помощью двоичных чисел можно закодировать все, что угодно: от команд, которые должен выполнять сам компьютер, до текста, изображений, музыки и фильмов.

Нужно лишь договориться, что в каждом случае будут означать группы двоичных чисел, из которых состоит файл, как их должны воспринимать и обрабатывать компьютерные программы. Алгоритм (правило), в соответствии с которым данные превращаются в цифры и помещаются в файл, называют *форматом файла*. Разумеется, для разного рода информации требуются разные способы кодирования. Более того, одну и ту же информацию можно представить в цифровом виде различными способами, поэтому форматов было изобретено много.

Например, Объединенная группа экспертов в области фотографии (Joint Photographic Experts Group) разработала набор спецификаций, позволяющих значительно сократить размер файла, в котором хранится изображение. Как вы уже догадались по первым буквам английского названия группы, в результате появился формат JPEG. С файлами этого формата работают и компьютерные программы, и цифровые фотоаппараты, и некоторые бытовые устройства.

В этой книге речь идет о файлах, содержащих данные двух видов. Это *текст* и *графика*.

Для компьютера **текст** – последовательность символов. Символами являются буквы разных алфавитов, цифры, знаки препинания. Пробел, разделяющий слова, – тоже символ. Каждому символу соответствует определенный числовой код.

С таким представлением текста компьютеру легко выполнять самые обычные математические операции. Например, он может найти в тексте все символы с определенным кодом, вставить в указанное место коды введенных с клавиатуры символов или заменить определенную последовательность на другую. На этом основана работа всех программ – текстовых редакторов. Простой пример – программа Блокнот (Notepad), поставляемая в составе ОС Windows. В текстовом редакторе можно выделить часть текста, скопировать ее в буфер обмена, а затем вставить в другое место того же документа или в другой документ; найти в тесте определенные последовательности символов и т. д.

Графика, или рисунок, в компьютерном представлении – набор отдельных точек, о каждой из которых известны ее положение и цвет.

ПРИМЕЧАНИЕ

В компьютерах представление цвета зависит от выбранной цветовой схемы.

Точки, из которых состоит изображение, называют пикселами (pixel, от англ. *PICture'S ELement или PICture Cell* – элемент или клетка изображения). Часто их зовут и просто точками (Dots). Таким образом, изображение является «картой точек», по-английски *Bitmap*. Графический файл содержит информацию обо всех точках изображения.

На рис. 1.1 показан пример рисунка «с точки зрения компьютера». Каждая клеточка изображает одну точку (пиксел). Мы видим, что ширина рисунка со-

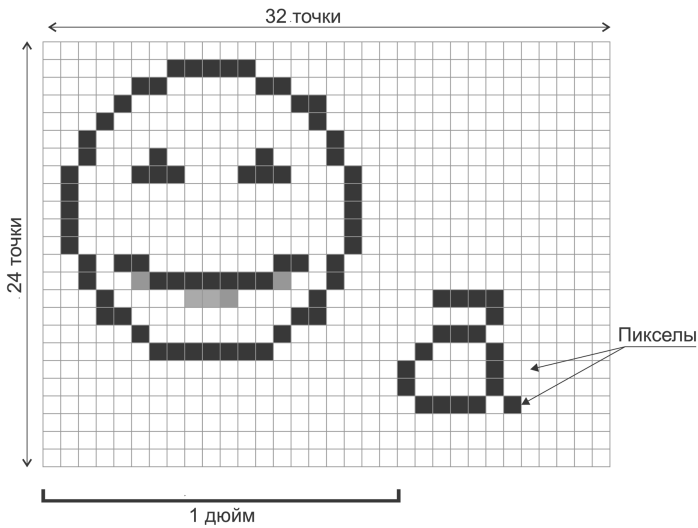


Рис. 1.1 ▾ Пример изображения

ставляет 32 точки, а высота – 24 точки. Белые, или «пустые», точки тоже считаются! В таком случае говорят, что размер этого рисунка – 32r24 точки (пиксела). Вместе с распространением цифровых камер размеры изображения чаще стали выражать общим числом точек. Например, размер этого рисунка – 768 пикселов, или примерно 0,0008 мегапиксела (Мпикс).

На том же примере продемонстрируем еще одну характеристику компьютерного изображения – *разрешение*. Для сравнения под рисунком помещена линейка длиной 1 дюйм (2,54 см). На рисунке ей соответствуют 20 точек. Таким образом, разрешение этого изображения составляет 20 точек на дюйм, или 20 DPI (Dots Per Inch – точек на дюйм).

Возникает вопрос: как же связаны размер изображения и его разрешение? У цифрового изображения есть лишь «истинный размер в пикселах». О разрешении можно говорить, только если мы одновременно уточним, с оригинала какого размера, в сантиметрах или дюймах, было получено (отсканировано, сфотографировано) это изображение, или каков должен быть размер изображения при его выводе на экран. На практике любой графический файл содержит служебную информацию (заголовок), где именно это и указано: размер в пикселах и разрешение в DPI одновременно.

Для работы в программе FineReader изображение обычно получают со сканера или цифровой камеры. При этом в настройках сканера задают разрешение получаемого изображения, например 150 или 300 DPI. Размер изображения будет зависеть от размера сканируемого оригинала. При сканировании с разрешением 300 DPI стандартной книжной страницы получится изображение размером примерно 2400×1600 точек, или, по «фотографической» терминологии, около 4 Мпикс.

Цифровой фотоаппарат выдает снимки определенного размера. Часто цифровые камеры позволяют регулировать размер снимка. Максимально возможный размер является одной из самых важных ее характеристик. В обиходе такой параметр называют «разрешением», хотя на самом деле это именно *размер* выходного изображения в точках.

Нужно различать *размер изображения* и *размер файла*. Первый выражается в количестве точек (пикселов) и характеризует размер самого изображения. Размер (объем) файла показывает, сколько места занимает файл на диске, и измеряется в байтах (килобайтах, мегабайтах). Размер файла зависит не только от размера изображения, но и от формата файла, а также использованных алгоритмов сжатия: если при сохранении изображения применяется сжатие информации, то файл получается меньше.

Форматы файлов

Сложилось так, что разработчики программ регулярно создавали и предлагали новые форматы файлов. В настоящее время существуют сотни различных форматов, каждый из которых обладает своими особенностями, достоинствами и, возможно, недостатками по сравнению с другими. Тип файла (документа) – бо-

лее общая характеристика содержимого файла, например «текст», «изображение», «звук», «архивы» и т. д. Так, к графическим форматам, в которых компьютер работает с изображениями, относятся форматы **BMP**, **TIFF**, **GIF**, **JPEG** и многие другие.

Любая программа способна открывать, обрабатывать и сохранять файлы только определенных типов и форматов. Некоторые программы поддерживают лишь один-два формата, но чаще приложения работают с целым рядом форматов. Рассмотрим некоторые типы и форматы файлов (документов), которые могут нам встретиться при работе с программой FineReader. Заметим, что от версии к версии набор поддерживаемых форматов, как графических, так и текстовых, может изменяться.

Текстовые форматы

Самым старым и простым типом файлов является текстовый. По сути, текстовый файл представляет собой только последовательность символов (букв, цифр, знаков препинания и математических символов), отображаемых на экране или выводимых на печать. Помимо них, в текстовом файле могут содержаться «символ перевода строки», «символ разрыва строки», «символ табуляции» и «символ конца страницы», которые не отображаются на экране и не распечатываются на бумаге. Это так называемые «непечатаемые символы».

Текстовый файл не содержит информации ни о размере или начертании шрифта, ни о выравнивании строк или отступе первой строки абзаца. То, как отображается текст на экране, зависит только от настроек программы, в которой открыт этот файл.

Текстовые файлы имеют расширение **TXT**. Работу с такими файлами должны поддерживать все программы, которые способны так или иначе обрабатывать текстовую информацию. На основе текстового файла было создано еще несколько форматов для более узких применений.

Файлы формата **CSV** (*Comma Separated Values* – значения, разделенные запятыми) являются такими же текстовыми файлами, как файлы формата txt, но специально предназначены для хранения данных в виде списков или таблиц. Каждая строка считается строкой таблицы, а запятые разделяют текст на части, которые должны быть помещены в отдельные ячейки. Когда вы открываете такой файл обычным текстовым редактором, например Блокнотом, то видите обычный текст. Если же открыть файл **CSV** программой для работы с электронными таблицами, например Microsoft Excel, то такая программа воспримет все запятые как разделители ячеек и автоматически разместит данные в ячейки таблицы. Иногда для разделения значений используется точка с запятой. При открытии или сохранении файла программы предлагают уточнить, какой символ следует считать разделителем.

В формате **HTML** (*HyperText Markup Language* – язык разметки гипертекста) создаются файлы, которые служат основой веб-страниц и предназначены для просмотра с помощью программ-браузеров. Эти файлы имеют расширение **HTML** или **HTM**. По существу, это тоже текстовые файлы, но в тексте есть спе-

циальные выражения (теги), обрамленные символами < и >, дающие указания браузеру, что, где и как следует отображать.

Графические файлы

Одним из самых старых и простых является формат **BMP**. Иначе его называют просто **Bitmap**. В файлах **BMP** информация об изображении практически не сжимается. Это и хорошо, и не очень. С одной стороны, сжатие часто приводит к ухудшению качества рисунка, утрате или смазыванию мелких деталей. С другой – файлы **BMP** получаются довольно большими.

Формат **TIFF** (**TIF**) по свойствам близок к формату **BMP**. Этот формат тоже позволяет сохранять любые изображения без потери их качества. При сохранении может применяться один из нескольких алгоритмов сжатия: **packbits**, **LZW**, **ZIP**.

В форматах **GIF** и **PNG** принято размещать изображения в Интернете. В эти форматы заложена поддержка анимированных рисунков и «прозрачного» фона – как раз то, чем часто пользуются веб-дизайнеры.

Формат **JPEG** (**JPG**) предназначен для сохранения цветных полутоновых изображений (фотографий) и интересен эффективным способом сжатия «картинки». Алгоритм основан на особенностях человеческого зрения: если убрать из рисунка самые мелкие цветные детали, или усреднить цвета близко расположенных точек, зритель почти не заметит такого «упрощения». За счет снижения качества изображения размер файла резко уменьшается – в 10 и более раз. Сжатие полезно, например, когда на флэш-карте фотоаппарата надо сохранить максимальное количество снимков, или файлы приходится передавать через Интернет. Однако для снимков страниц с текстом сжатие **JPEG** подходит хуже – ведь в изображениях такого рода важны именно детали. Если сгладить эти «мелочи», то изображение букв может стать неразборчивым.

DjVu (**DJV**) – относительно новый формат. Он был изобретен специально для сохранения сканов или снимков всевозможных печатных изданий.

Здесь тоже применяется сжатие с потерей качества, но по особому «интеллектуальному» алгоритму. Изображение текста и других элементов с четкими контурами почти не теряет в качестве, тогда как изображение фона и полутоновых иллюстраций «упрощается» и сжимается очень сильно. Еще одна особенность формата – в одном файле хранятся изображения многих страниц. Хотя формат **DjVu** и принято считать графическим, в нем вместе с изображением может храниться и текст – так называемый *текстовый слой*. Файлы **DjVu** создаются в несколько этапов. Сначала страницы книги или журнала сканируются или фотографируются, а изображения сохраняются в файлы формата **BMP** или **TIFF**. Затем с помощью специализированных программ эти изображения «собираются» и сохраняются в единый файл, уже в формате **DjVu**. Среди таких программ можно назвать пакет Document Express и целый ряд приложений, распространяемых бесплатно. Открывать файлы **DjVu** способны многие современные программы-просмотрщики графики. Существуют программы, предназначенные исключительно для просмотра таких файлов, например WinDjView

(URL: windjview.sourceforge.net). Программа FineReader, начиная с версии 9, «умеет» открывать файлы **DjVu** и распознавать их.

Составные (сложные) документы

В жизни мы обычно встречаемся с документами, в которых одновременно присутствуют и текст, и разного рода графические элементы: иллюстрации, рамки, линии. При этом часть текста может быть оформлена в виде таблиц. Для создания и редактирования таких документов служат программы, которые обычно называют «офисными приложениями». Самый известный инструмент подобного рода – пакет Microsoft Office. Приложения Microsoft Word и Microsoft Excel стали фактическим стандартом в этой области. Файлы, создаваемые такими приложениями, сохраняются в особых форматах, которые нельзя отнести ни к текстовым, ни к графическим.

Текст в таких документах обычно *отформатирован*. Под форматированием текста понимают оформление его определенными шрифтами, размер и цвет шрифта, начертание (подчеркивание, *курсив* или **жирный**). Формат абзаца – выравнивание строк на странице: по левому или правому краю, по центру, по всей ширине страницы, а также интервал между строками внутри абзаца и между абзацами, отступ первой строки и т. д. Форматирование применимо и к тексту в ячейках таблиц.

В составных документах хранятся и текст, и изображения, и сведения о разметке страницы: форматировании текста и расположении рисунков относительно текста. Кроме того, такие файлы могут содержать гиперссылки (ссылки на какие-то другие файлы или ресурсы Интернета), макрокоманды и другие элементы, которые используют соответствующие приложения.

Документы Microsoft Office – целый набор форматов, разработанных корпорацией Microsoft. Полная спецификация форматов является коммерческой тайной компании, и другие разработчики программного обеспечения имеют доступ лишь к отдельным техническим подробностям. Поэтому, за редкими исключениями, документы Microsoft Office могут быть открыты только в приложениях того же пакета. Одним из исключений являются программы бесплатно распространяемого пакета OpenOffice.org, позволяющие открывать, редактировать и сохранять документы Word и Excel, хотя и с некоторыми ограничениями.

Впрочем, пакет Microsoft Office устанавливают практически на всех компьютерах с ОС Windows. Файлы в форматах Microsoft Office давно стали основным и самым массовым видом документов.

- Документ Microsoft Word (файл с расширением **DOC**) может включать в себя и редактируемый текст, и таблицы, и изображения, и элементы графического оформления, хотя в основе такого документа лежит текстовое содержимое.
- Документ Microsoft Excel – электронная таблица, в которую могут включаться и графические элементы. Эти файлы имеют расширение **XLS**.

- Документы, или презентации, Microsoft PowerPoint (расширение **PPT**) позволяют объединять в одном файле текст и графику с элементами мультимедиа: видео- и звукозаписями.

С появлением очередной версии пакета, Microsoft Office 2007, был предложен и новый набор форматов, основанных на спецификации **XML**. В этом проявилась одна из стратегических идей компании Microsoft – создание некоторого «универсального» представления данных. Такое представление ориентировано прежде всего на Интернет, где широко используются ссылки, а данные могут находиться в самых разных местах. Пока еще эти форматы не стали популярными. Многие продолжают пользоваться прежними версиями Microsoft Office.

Формат **PDF** (Portable Document Format – переносимый формат документов) разработан компанией Adobe Systems. Это полностью открытый стандарт, который широко используют для публикации различных документов в электронном виде. Для открытия и просмотра файлов **PDF** достаточно установить бесплатную программу Adobe Reader (URL: www.adobe.com). Поддержка этого формата заложена и во многих браузерах.

В файлах **PDF** может храниться как текстовое, так и графическое содержимое. По принятой терминологии, они образуют отдельные *слои* (layers) – текстовый и графический.

Текстовый слой – текст, который можно выделять, копировать и затем вставлять в другие документы. Однако при создании файла **PDF** автор мог установить защиту от копирования. В таком случае выделить текст указателем мыши удается, а скопировать выделенный текст в буфер обмена не получится. Графический слой – это изображения, причем их размер и положение относительно краев страницы зафиксированы.

Часто можно встретить документы **PDF**, в которых текстовый слой пуст, а на графический слой помещены отсканированные изображения каких-либо печатных материалов. В таком документе выделить и скопировать текст тоже невозможно – текста там попросту нет. В документах **PDF**, которые содержат оба слоя, текстовый и графический слои могут быть расположены по-разному, и в них можно выполнять поиск по тексту.

- Документы, в которых текст расположен поверх графики. Этот вариант PDF-документов поддерживает поиск по тексту, выделение и копирование текста и изображения. Однако при отображении документа на экране и при печати используются стандартные шрифты, размер и начертание которых могут несколько отличаться от изображенных на графическом слое. Поэтому внешний вид такого документа в целом соответствует оригиналу, но может в большей или меньшей мере отличаться от него. В интерфейсе FineReader такой тип документа обозначается как **Редактируемый PDF-документ**.
- Документы, в которых графический слой расположен поверх текстового. Благодаря наличию текстового слоя возможен поиск по тексту документа. Текст документа можно выделить (целиком или частично) и скопировать, изображение, содержащееся в документе, также можно

выделить и скопировать. Внешний вид документа повторяет оригинал. В интерфейсе FineReader такой тип документа обозначается как **PDF (изображение с поиском)**.

Программа FineReader работает с PDF-документами. С одной стороны, они являются одним из источников изображений, которые надо распознать. С другой – в формат **PDF** программа способна сохранять уже распознанные документы. Вы можете выбрать любой из двух типов сохранения: с расположением текста поверх изображения или с расположением изображения поверх текста.

При работе с файлами **PDF** следует учитывать, что этот формат имеет различные версии. Новые версии программ Adobe для работы с файлами **PDF** (Adobe Acrobat, Adobe Reader) поддерживают работу с форматами **PDF** прежних версий, но не наоборот – новые форматы **PDF** для устаревших версий этих программ недоступны. Разновидностью формата **PDF** является формат **PDF/A, специально предназначенный для долгосрочного хранения архивов**.

Оптическое распознавание символов (OCR)

Основная задача, которую решает программа FineReader, – оптическое распознавание символов. Другими словами, по изображению текста нужно воссоздать сам текст как последовательность символов. Это одна из самых интересных задач кибернетики, причем корни проблемы уходят глубоко в область философии, математики и психологии.

Почти каждому из нас легко прочитать текст, независимо от того, как и на чем он изображен. Пусть буквы напечатаны или написаны на бумаге, высечены в граните, согнуты из неоновой трубки на рекламной вывеске или криво нацарапаны на заборе – почти всегда мы понимаем, что написано. Все распознавание занимает доли секунды, и результат его оказывается точен почти на 100%. Более того, мы узнаем слова, даже если в них пропущены какие-то буквы. Есть лишь одна оговорка: текст должен быть написан на знакомом языке. Так что дело не только в узнавании отдельных символов, но и в анализе слов и предложений в целом. Разработкой алгоритмов оптического распознавания символов (Optical Character Recognition, OCR) уже второй десяток лет занимается целый ряд коллективов. Принцип, казалось бы, «лежит на поверхности». Компьютер должен сравнивать то, что изображено на рисунке, с образцами изображений букв и других символов, хранящимися в его памяти. Когда очертания совпадут, как «ключ с замком», символ можно считать распознанным. Однако в действительности все не так просто. Во-первых, существует великое множество шрифтов, и все они отличаются именно очертаниями знаков. К тому же, высота и ширина знаков может быть любой. Во-вторых, на реальном снимке или отсканированном изображении всегда присутствуют искажения: поворот, перекос, искривления и шумы различного происхождения. Четкость изображения тоже бывает далека от идеала. Распознавание путем простого перебора и сравнения «точек с точками» требует очень большого набора образцов и не гарантирует

успеха. На «контурное распознавание» опирались программы OCR, создававшиеся в начале 90-х годов.

В современных системах OCR задействованы более интеллектуальные механизмы. В основе их работы лежит принцип «проверки гипотез». Компьютерные программы анализируют изображение в несколько этапов. На каждом шаге берутся некоторые предположения, сначала наиболее вероятные. Если все гипотезы оказались верны, в результате должен получиться распознанный текст, состоящий из слов и предложений. В каждом языке существуют свои нормы правописания. Когда получившийся текст отвечает этим нормам, скорее всего, он был распознан правильно. Если нет – программа возвращается к предыдущим шагам и проверяет другие гипотезы. Возможно, это слово не на русском, а на английском языке? Возможно, то, что было принято за букву «ф», на самом деле слившиеся на изображении буквы «о» и «р»?

Сначала все изображение подвергается предварительной обработке. В нем выделяются области (блоки), где, вероятно, изображен текст. Программа делает такое предположение, поскольку область с текстом обычно выглядит как строки, состоящие из отдельных знаков примерно одинаковой высоты. Там, где находятся иллюстрации, такая закономерность не прослеживается, и программа решает, что в этих областях искать текст не надо.

В изображениях символов программа пытается выделить характерные элементы – линии, углы, закругления. Строится как бы структурный «скелет» знака – в отличие от контуров, это постоянная его характеристика. Эталон знака тоже является описанием структурных признаков, отличающих этот знак от других. Сравнение не контуров, а структуры позволяет уверенно различать похожие символы и распознавать изображения символов с частично искаженными контурами. Отличительной особенностью программы ABBYY FineReader являются малая чувствительность к дефектам печати и способность распознавать тексты, набранные практически любыми шрифтами.

После распознавания знаков наступает очередь анализа на уровне алфавита и языка. При этом учитываются статистические закономерности. Определив язык, на котором написан документ, программа может проверять распознаваемый текст в соответствии с грамматическими правилами и словарями для этого языка. Для языков со словарной поддержкой программа может предложить вероятные слова из словаря, если слово распознано неуверенно. В реальных текстах порой встречаются отдельные слова или фразы на другом языке – такая возможность тоже предусмотрена.

Российская компания ABBYY занимает среди разработчиков систем OCR достойное место. В распознавании кириллических текстов (русский, украинский, белорусский и другие языки) у программы FineReader практически нет конкурентов. Поскольку компания ABBYY также известна созданием электронного словаря Lingvo, технологии языкового анализа являются «сильной стороной» всех ее продуктов.

Благодаря передовой технологии адаптивного распознавания документов ADRT® (Adaptive Document Recognition Technology), разработанной компани-

ей ABBYY, программа ABBYY FineReader позволяет анализировать и обрабатывать документ целиком, а не постранично. В результате восстанавливается исходная структура документа, включая форматирование, гиперссылки, адреса электронной почты, а также колонтитулы, подписи к картинкам и диаграммам, номера страниц и сноски.

ABBYY FineReader 10 распознает документы, написанные на одном или нескольких из 186 языков, включая корейский, китайский, японский, тайский и иврит. В программу встроена также функция автоматического определения языка документа.

От чего зависит качество распознавания?

Попробуем разобраться, в каких случаях тексты распознаются легко, а в каких – не очень.

Прежде всего для успешного распознавания нужно качественное изображение на оригинале. Примеры хороших оригиналов – текст, недавно распечатанный на исправном лазерном принтере, журнальные и книжные страницы. Хорошо распознается текст и из газет, отпечатанных на бумаге хорошего качества. Если вы используете заранее сделанные фотографии или сканы таких оригиналов, необходимо, чтобы размер изображения книжного разворота или листа формата A4 составлял не менее 2 Мпикс. Другое условие – точная фокусировка и правильно выбранная выдержка: границы букв должны быть четкими, шрифт – практически черным, а фон – близок к белому.

Большую роль играет и сам характер документа. Когда иллюстрации и таблицы четко отделены от окружающего текста, меньше вероятность ошибок при определении областей.

В подобных случаях программа легко справляется с распознаванием «на полном автомате» – такой режим мы рассмотрим в следующей главе. Как правило, точность распознавания простых текстов с хороших оригиналов превышает 98%, то есть одна ошибка на 50 слов и реже.

Если через тонкую бумагу просвечивает напечатанное на обороте, на отсканированном изображении проявляется много артефактов. На плохих ксерокопиях присутствуют точки, штрихи, светлые и темные полосы: такие следы оставляет отработавший свой срок фотобарабан. Трудности при распознавании создают мятая и надорванная бумага, затертые сгибы, пятна и пометки поверх текста. Контрастность изображения ухудшают пожелтевший фон, бледная или выцветшая краска. Работа с оригиналами такого вида считается более сложным случаем. Точно так же неудобны для распознавания фотографии или сканы страниц, сделанные с малым разрешением или сильно сжатые при сохранении в формате JPEG.

Программы распознавания обычно «не любят» документы, в которых текст расположен на фоне иллюстраций или где много врезок и выносок. Типичный пример – страницы глянцевых журналов, рекламные материалы. В них шрифт иногда расположен на фоне рисунков, а отдельные фрагменты и врезки могут быть повернуты под прямым углом по отношению к остальному тексту. При

работе с такими документами программе приходится «помогать» в выделении областей, содержащих текст и иллюстрации. Для успешного распознавания таких документов в программе FineReader 10 появилась возможность сделать картинку фоном и выделить на ней текст.

Качество распознавания зависит и от содержания обрабатываемых документов. В научной и технической литературе регулярно встречаются специальные термины, имена собственные, аббревиатуры, слова на других языках и формулы. Кроме того, в подобных изданиях попадаются сложные таблицы и схемы. Затруднения бывают связаны и с выделением областей, и со словарной проверкой. Для достижения наилучших результатов при распознавании таких документов может потребоваться дополнительная настройка программы FineReader.

Системные требования

Программа ABBYY FineReader, как и другие подобные инструменты, интенсивно использует вычислительные ресурсы компьютера. Это и неудивительно – распознавание символов связано с перебором огромного числа вариантов, а достаточно объемные промежуточные результаты компьютер должен держать в оперативной памяти.

Согласно документации, для нормальной работы программы требуется компьютер на процессоре с тактовой частотой не менее 1 ГГц и объемом оперативной памяти от 512 Мб. Однако это лишь минимум – на таком компьютере распознавание каждой страницы занимает несколько минут. Во время работы остальные приложения будут заметно «тормозить». Для комфортной работы желательны двухъ- или четырехъядерный процессор с частотой 2,4 ГГц и более и оперативная память объемом 2–4 Гб.

На жестком диске файлы установленной программы FineReader 10 занимают до 650 Мб. При обработке документов создаются временные файлы. После того как документ обработан и вы сохранили результаты распознавания, эти файлы удаляются автоматически. При стандартных настройках операционной системы Windows временные файлы создаются на том же диске, на котором установлена сама система (обычно это диск **C:**). Считается, что для создания временных файлов на этом диске должно быть не менее 1 Гб свободного места. Однако при работе с большими документами, например с книгой из 500 страниц, объем временных файлов может достигать нескольких гигабайт. В таком случае на диске следует заранее освободить около 5 Гб или даже больше.

Резюме

Для компьютера текст – последовательность символов, а изображение – набор точек. И то, и другое существует в виде файлов. Текст создается и обрабатывается с помощью программ – текстовых редакторов, а для работы с изображениями предназначены графические редакторы. В первом случае идет работа

с символами, из которых складываются слова и предложения, а во втором – только с точками разного цвета.

Главное отличие собственно текста от картинки, на которой такой текст изображен, – то, что его можно редактировать: вставлять, удалять, копировать отдельные символы, слова и предложения. На изображении же можно только закрашивать отдельные участки, менять яркость и контрастность и т. п.

В результате **сканирования**, или **фотографирования**, текста получается лишь изображение, то есть картинка, на которой изображен текст. С «точки зрения» компьютера, текста в изображении нет и быть не может, пусть даже на картинке изображены буквы и цифры.

Для превращения «нарисованного текста» в текст редактируемый нужна программа оптического распознавания символов, например ABBYY FineReader. В качестве источника эта программа берет изображение, получаемое со сканера или хранящееся в файле одного из графических форматов, или цифровую фотографию. На выходе программа выдает редактируемый текст. Этот текст может быть передан в различные приложения для дальнейшей обработки или сохранен в разных форматах, отправлен по электронной почте, скопирован в буфер обмена.

Глава 2 Быстрый старт

В программу FineReader 10 встроен набор сценариев. Каждый из них – готовая последовательность действий для автоматического выполнения одной из типичных задач. Выполняя встроенные сценарии, программа предложит вам только отсканировать бумажный оригинал или указать файл(ы) с изображениями, который нужно открыть, – все остальное будет сделано без вашего участия. При работе с простыми для распознавания документами проще всего воспользоваться этой функцией.

При запуске программы в ее главном окне по умолчанию сразу открывается окно **Новое задание** (рис. 2.1). Оно содержит кнопки для вызова встроенных сценариев.

Сценарии в окне **Новое задание** для удобства сгруппированы в четыре закладки. На закладке **Основные**, которая показана на рис. 2.1, собраны шесть наиболее часто используемых сценариев. При конвертировании в разные форматы с помощью встроенных сценариев создается новый документ в указанном формате, и после завершения распознавания он сразу же открывается в целевой программе.

- **Сканировать в Microsoft Word**. На практике к этому сценарию обращаются чаще всего. Бумажный оригинал сканируется, программа FineReader обрабатывает и распознает полученное изображение, сразу же вызывает программу Microsoft Word и передает в нее распознанный документ.
- **Сканировать в Microsoft Excel**. После сканирования и распознавания документ передается в программу Microsoft Excel. Этот сценарий подходит для работы с документами, в основном состоящими из таблиц, с которыми в дальнейшем вы планируете работать в программе Microsoft Excel.
- **Сканировать в PDF (изображение с поиском)**. После сканирования и распознавания программа сохраняет результат в файл PDF с возможностью поиска по тексту.

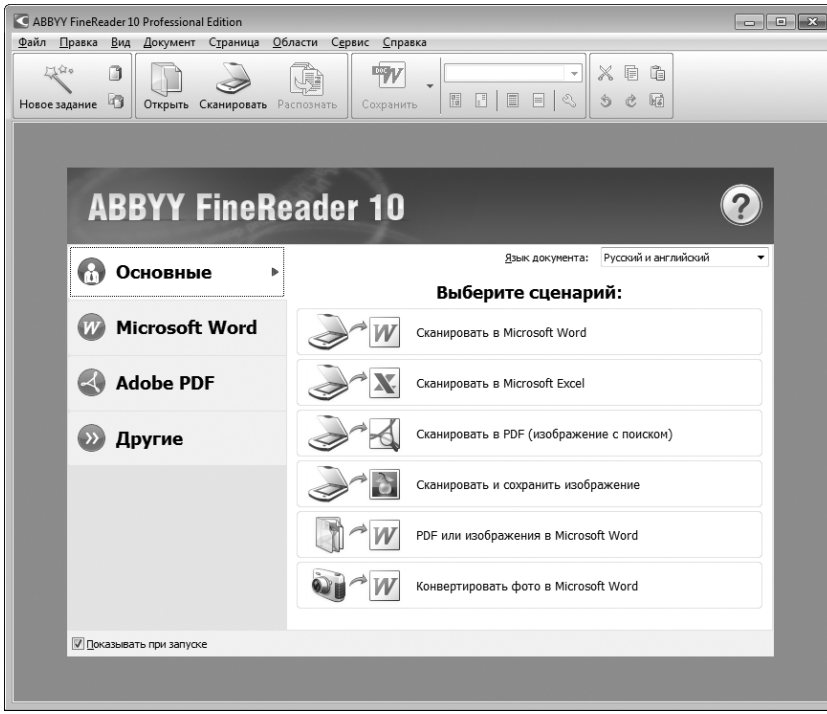


Рис. 2.1 ▾ Главное окно программы FineReader с окном **Новое задание**

- **Сканировать и сохранить изображение**. По этому сценарию изображение, полученное со сканера, сохраняется в файл одного из графических форматов без распознавания.
- **PDF или изображения в Microsoft Word**. По этому сценарию программа открывает указанные файлы PDF или файлы изображений, распознает их и передает распознанный документ в Microsoft Word.
- **Конвертировать фото в Microsoft Word**. Это практически тот же сценарий, что и **PDF или изображения в Microsoft Word**, но он предназначен для работы с цифровыми фотоснимками.

ПРИМЕЧАНИЕ

*Доступность встроенных сценариев зависит от того, установлены ли на вашем компьютере «целевые» программы. Сценарии **Сканировать в Microsoft Word**, **PDF или изображения в Microsoft Word** и **Конвертировать фотографию в Microsoft Word** отображаются в окне **Основные сценарии** только в случае, если на компьютере установлена программа Microsoft Word. Сценарий **Сканировать в Microsoft Excel** доступен при наличии на компьютере программы Microsoft Excel. Сценарий **Сканировать в PDF** доступен, если на компьютере установлена хотя бы одна из программ: Adobe Reader или Adobe Acrobat.*

Закладка **Microsoft Word** содержит три из уже названных сценариев: **Сканировать в Microsoft Word**, **PDF или изображения в Microsoft Word** и **Конвертировать фото в Microsoft Word**. Когда выбрана эта закладка, в нижней части окна **Новое задание** отображаются элементы управления дополнительными настройками, предназначенными именно для конвертирования в Microsoft Word (рис. 2.2).

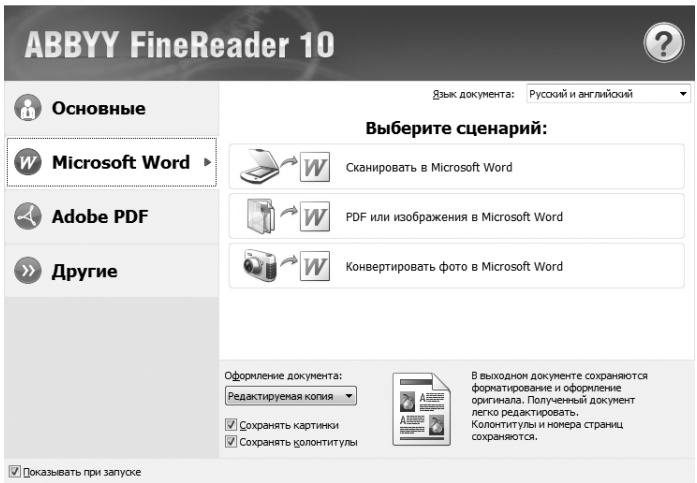


Рис. 2.2 ▾ Закладка Microsoft Word

В раскрывающемся списке **Оформление документа** доступны четыре режима сохранения оформления распознанного документа:

- **Точная копия** – оформление исходного документа передается в документ Word максимально точно. Однако данный режим не предполагает внесения значительных правок в текст и оформление;
- **Редактируемая копия** – оформление распознанного документа достаточно близко к оригиналу, но может немного отличаться от него. Документ, полученный в данном режиме, легко редактировать;
- **Форматированный текст** – при передаче документа в Microsoft Word сохраняются начертание и размер шрифта, разбиение на абзацы, но расположение объектов на странице может не соответствовать оригинальному. Например, все абзацы выравниваются по левому краю, а текст, который в исходном документе был ориентирован вертикально, в распознанном документе будет расположен горизонтально, как и все абзацы;
- **Простой текст** – передача текста без форматирования. В этом случае в документе Microsoft Word ко всему распознанному тексту будут применены шрифт и формат абзацев, принятые в этой программе по умолчанию.

Кроме того, при установленном флажке **Сохранять картинки** рисунки из исходного документа будут вставлены в документ Microsoft Word. Если этот флажок снят, будет передан только текст без рисунков.

Флажок **Сохранять колонтитулы** определяет, будут ли переданы в документ Word колонтитулы – особые строки у верхней и нижней границ страницы, в которых обычно указываются номер страницы, название книги или главы. Когда флажок установлен, программа Fine Reader, найдя на странице подобные элементы, постарается поместить их в документ Microsoft Word именно как колонтитулы. Если флажок снят, строки, распознанные как колонтитулы, не будут передаваться в выходной документ.

На закладке **Adobe PDF** предлагаются четыре сценария для конвертирования документов в формат PDF. Для их выполнения необходимо, чтобы на компьютере было установлено одно из приложений, работающих с файлами формата PDF: программа Adobe Reader или Adobe Acrobat.

- **Сканировать в PDF (изображение с поиском)** – оригинал сканируется, а результат распознавания сохраняется в документ Adobe PDF в режиме Текст под изображением страницы.
- **Сканировать в редактируемый PDF-документ** – сценарий делает то же, что предыдущий, но в выходном документе Adobe PDF текст помещается поверх изображения страницы.
- **Конвертировать в PDF (изображение с поиском)** – сценарий преобразовывает документы PDF и файлы изображений в документ Adobe PDF в режиме Текст под изображением страницы.
- **Конвертировать в редактируемый PDF-документ** – позволяет конвертировать PDF-документы и файлы изображений в документ Adobe PDF в режиме Текст поверх изображения страницы.

Кроме того, в нижней части этой закладки есть опция **Сжать в черно-белый PDF-документ**. Включив эту опцию, можно уменьшить размер выходного PDF-файла.

Закладка **Другие** содержит ссылки для вызова остальных встроенных сценариев программы ABBYY FineReader. Эти шесть сценариев позволяют автоматически решить дополнительные задачи, с которыми пользователь может встретиться при распознавании различных оригиналов и сохранении результатов их распознавания.

- **Сканировать в документ HTML** – бумажный документ сканируется, а результат распознавания сохраняется в виде документа HTML. Получившуюся веб-страницу можно, например, разместить впоследствии на сайте в Интернете.
- **Сканировать в другие форматы** – сценарий позволяет отсканировать бумажный документ и сохранить его в файл любого из форматов, поддерживаемых программой FineReader 10.
- **PDF или изображения в Microsoft Excel** – сценарий нужен для конвертирования документов PDF или файлов изображений в документ Microsoft Excel.

- **PDF или изображения в другие форматы** – сценарий служит для преобразования документов PDF и файлов изображений в любой из форматов, поддерживаемых программой FineReader 10. При выполнении этого сценария вам предлагается выбрать формат файла, в который будет сохранен результат распознавания.
- **Сканировать** – сценарий сканирования бумажных документов. В дальнейшем вы продолжите работу с отсканированным документом в пошаговом режиме.
- **Открыть** – сценарий открывает PDF-документы и файлы изображений. В дальнейшем вы продолжите работу с открытым документом в пошаговом режиме.

В верхней части окна **Новое задание** расположен раскрывающийся список **Язык документа**. Перед запуском любого из сценариев выберите в этом списке нужный язык или языки. По умолчанию предлагаются два языка: русский и английский. В этом случае программа будет распознавать в тексте слова и на русском, и на английском языке. Для большинства документов на русском языке лучше прямо указать один язык – русский. Если же в русском тексте присутствуют отдельные слова на английском языке (типичный пример – компьютерная литература), оставьте предлагаемый по умолчанию вариант: **Русский и английский**.

Указав язык, выберите нужный сценарий. Поскольку некоторые сценарии похожи, рассмотрим всего три случая.

Сканирование в MS Word, MS Excel, PDF

Эти сценарии отличаются лишь тем, в какую программу будет передан результат. Сканировать в приложения Microsoft Office удобно, чтобы продолжить работу с распознанным документом, отредактировать его. Например, вы собираетесь составить реферат. У вас есть несколько учебников и журналов. Из каждого нужно взять отдельные фрагменты, «состыковать» их, а потом добавить немного собственных фраз и излюбленных выражений преподавателя. В таком случае стоит передать результаты распознавания в документы Microsoft Word.

Если оригинал – напечатанная таблица и вы хотите впоследствии работать с данными именно как с таблицей, воспользуйтесь сценарием **Сканировать в Microsoft Excel**. В электронной таблице легко отсортировать строки, добавить новые колонки или поменять их местами, внести формулы и выполнить операции с числами. Например, вы получили от кого-то распечатанный прайслист, отсканировали его и передали в программу Microsoft Excel. В электронной таблице легко добавить новую колонку, чтобы в ней пересчитать все цены с учетом скидки или изменившегося курса валюты.

В формат PDF с возможностью поиска по тексту обычно сохраняют отсканированные документы, чтобы просто получить «электронную копию» для чтения, отправки по электронной почте или публикации в Сети. Впоследствии

документ PDF всегда можно просмотреть программой Adobe Reader, найти и при необходимости скопировать из него любую часть текста.

1. В окне **Новое задание** выберите одну из ссылок, начинающихся со слова «сканировать». Откроется диалог сканирования. По умолчанию программа FineReader использует собственный интерфейс управления сканером, и окно выглядит так, как на рис. 2.3.
2. Положите документ в сканер. Нажмите кнопку **Просмотр**. Сканер быстро получит предварительное, «грубое» изображение листа, и оно появится в левой части окна (рис. 2.3).

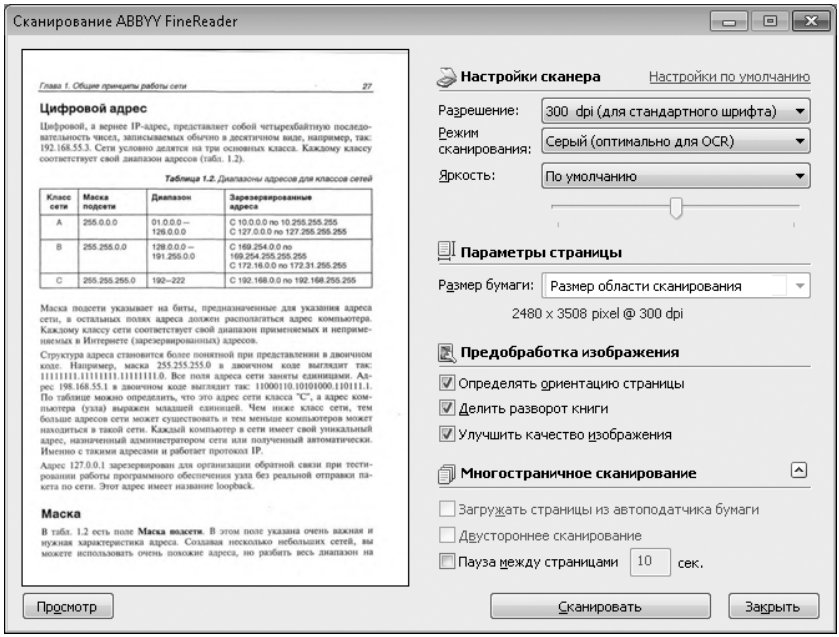


Рис. 2.3 ▾ Окно сканирования

3. Убедитесь, что в группе **Настройки сканера** выбраны подходящие значения в раскрывающихся списках:
   - **Разрешение** – 300 dpi (для стандартного шрифта);
   - **Режим сканирования** – серый (оптимально для OCR);
   - **Яркость** – по умолчанию.

   В большинстве случаев эти настройки менять не нужно. Если вы сканируете оригинал с цветными иллюстрациями и хотите увидеть такие же рисунки в распознанном документе, либо вам важно сохранить цвета текста, в раскрывающемся списке **Режим сканирования** выберите значение **Цветной**.

В группе **Предобработка изображения** по умолчанию установлены все три флажка: **Определять ориентацию страницы**, **Делить разворот книги** и **Улучшать качество изображения**. Выполнение предобработки в любом случае не ухудшит качества распознавания документа, поэтому целесообразно оставить эти флажки по умолчанию. Если ваш документ содержит текст на иероглифическом языке в сочетании с каким-то из европейских языков, то рекомендуется отключить опцию автоматического определения ориентации страниц, а опцию разбиения сдвоенных страниц использовать только в том случае, если все изображения страниц имеют правильную ориентацию.

4. Нажмите кнопку **Сканировать**. В зависимости от модели сканера получение и передача изображения занимают от нескольких секунд до минуты. После того как первая страница отсканирована, а ее изображение передано в компьютер, название кнопки изменится на **Сканировать следующую страницу**.
5. Положите в сканер следующий лист. Нажмите кнопку **Сканировать следующую страницу**. Таким образом отсканируйте все листы.
6. Нажмите кнопку **Закрыть**. Окно сканирования закроется, а программа приступит к обработке изображений и распознаванию документа. При этом в главном окне программы FineReader виден индикатор выполнения задания.

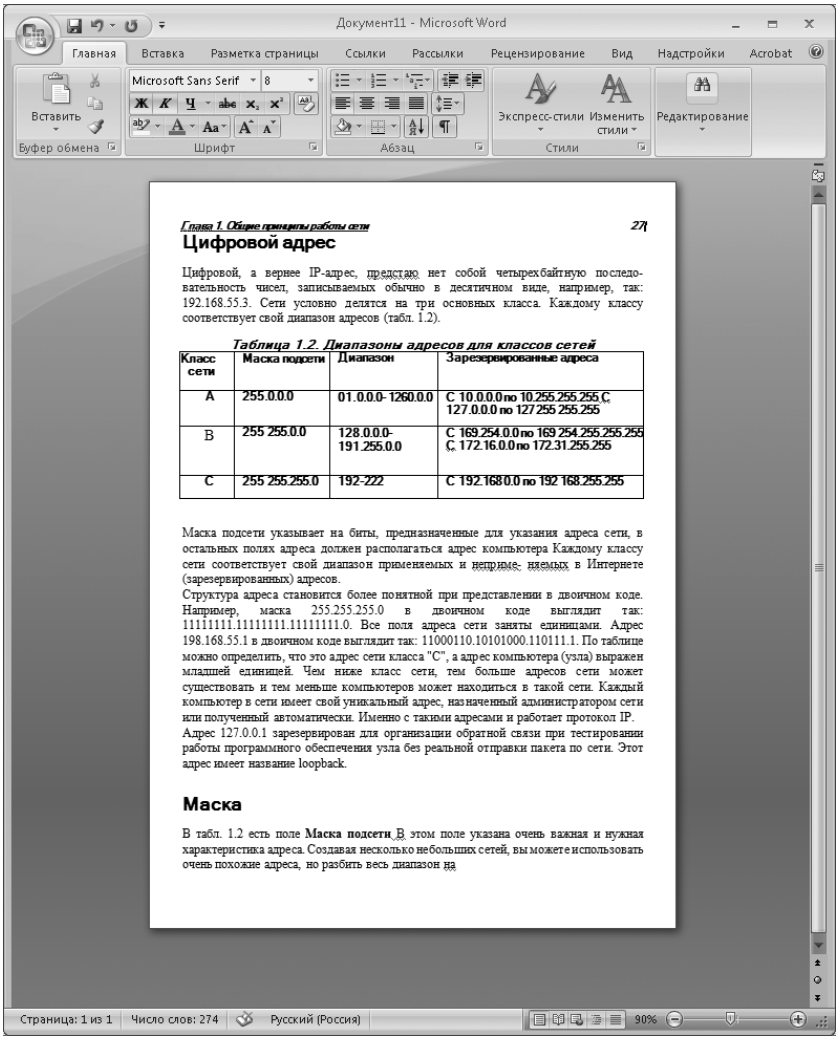
   Через некоторое время откроется окно приложения, в которое передан распознанный документ. Время распознавания зависит от производительности компьютера, а также объема текста, качества печати и сложности структуры документа.

По сценарию **Сканировать в Microsoft Word** в результате откроется окно программы Microsoft Word с распознанным документом (рис. 2.4). В результате выполнения сценария **Сканировать в Microsoft Excel** откроется окно программы Microsoft Excel с таблицей. Проверьте распознанный текст или таблицу на предмет ошибок, при необходимости исправьте их, а затем сохраните документ Word или таблицу Excel.

Если же был выбран сценарий **Сканировать в PDF**, откроется окно той программы, с которой на вашем компьютере ассоциирован этот формат файлов. Если для формата PDF программой по умолчанию у вас является Adobe Reader, то откроется это приложение (рис. 2.5).

Adobe Reader – действительно только «читатель»: эта программа не позволяет редактировать открытые в ней документы. Единственное, что вам остается сделать, – просмотреть получившийся документ PDF и сохранить его под нужным именем, «как есть».

1. В окне программы Adobe Reader выберите команду меню **Файл ➢ Сохранить копию...**(рис. 2.4).
2. В открывшемся окне сохранения документа выберите папку и укажите имя документа. В результате вы получите файл PDF в указанном месте.



Глава 1. Общие принципы работы сети 27

Цифровой адрес

Цифровой, а вернее IP-адрес, представляет собой четырехбайтную последовательность чисел, записываемых обычно в десятичном виде, например, так: 192.168.55.3. Сети условно делятся на три основных класса. Каждому классу соответствует свой диапазон адресов (табл. 1.2).

Таблица 1.2. Диапазоны адресов для классов сетей

Класс сети	Маска подсети	Диапазон	Зарезервированные адреса
A	255.0.0.0	01.0.0.0- 1260.0.0	С 10.0.0.0 по 10.255.255.255 С 127.0.0.0 по 127 255 255.255
B	255 255.0.0	128.0.0.0- 191.255.0.0	С 169.254.0.0 по 169 254.255.255.255 С 172.16.0.0 по 172.31.255.255
C	255 255.255.0	192-222	С 192.168 0.0 по 192 168.255.255

Маска подсети указывает на биты, предназначенные для указания адреса сети, в остальных полях адреса должен располагаться адрес компьютера Каждому классу сети соответствует свой диапазон применяемых и неприме- няемых в Интернете (зарезервированных) адресов.

Структура адреса становится более понятной при представлении в двоичном коде. Например, маска 255.255.255.0 в двоичном коде выглядит так: 11111111.11111111.11111111.0. Все поля адреса сети заняты единицами. Адрес 198.168.55.1 в двоичном коде выглядит так: 11000110.10101000.110111.1. По таблице можно определить, что это адрес сети класса "С", а адрес компьютера (узла) выражен младшей единицей. Чем ниже класс сети, тем больше адресов сети может существовать и тем меньше компьютеров может находиться в такой сети. Каждый компьютер в сети имеет свой уникальный адрес, назначенный администратором сети или полученный автоматически. Именно с такими адресами и работает протокол IP.

Адрес 127.0.0.1 зарезервирован для организации обратной связи при тестировании работы программного обеспечения узла без реальной отправки пакета по сети. Этот адрес имеет название loopback.

Маска

В табл. 1.2 есть поле **Маска подсети** В этом поле указана очень важная и нужная характеристика адреса. Создавая несколько небольших сетей, вы можете использовать очень похожие адреса, но разбить весь диапазон на

Рис. 2.4 ▾ Распознанный документ передан в Microsoft Word

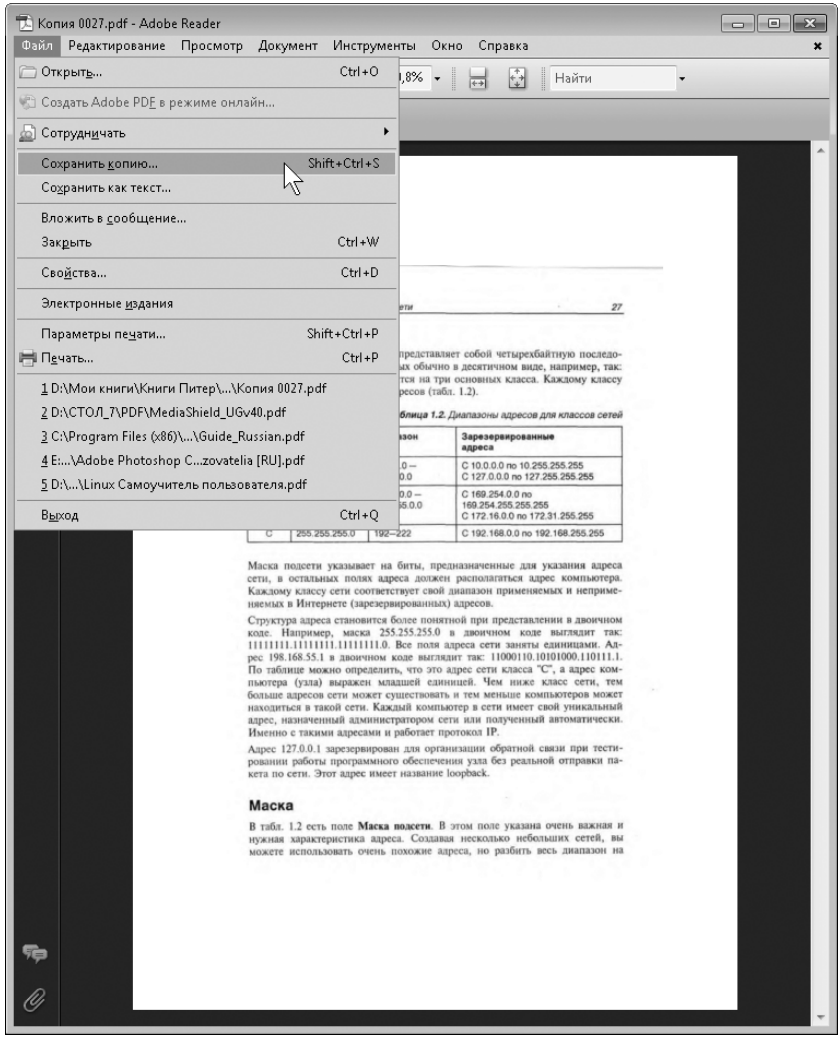


Рис. 2.5 ▾ Распознанный документ в окне Acrobat Reader

Когда на компьютере установлена программа Adobe Acrobat, то она, как правило, становится программой по умолчанию для открытия файлов формата PDF. Соответственно, результат распознавания будет передаваться в это приложение. Adobe Acrobat – полноценный редактор, позволяющий править текст практически так же, как Microsoft Word. Проверив и отредактировав в окне программы Adobe Acrobat распознанный документ, сохраните его в формате PDF.

Конвертирование изображений и PDF в документ Microsoft Word

Несколько сценариев позволяют распознать существующие изображения и PDF-документы, чтобы далее превратить их в редактируемые документы. Файлы с изображениями попадают в компьютер разными способами: сохраняются со сканеров, копируются из фотоаппаратов или скачиваются из Интернета и т. п.

Например, вы сфотографировали страницы книги или журнала. Подключите камеру к компьютеру кабелем USB. Карту памяти фотоаппарата компьютер определит как съемный диск. Скопируйте изображения с него в какую-либо папку. Иначе можно извлечь из камеры карту памяти и вставить ее в кард-ридер, подключенный к компьютеру. Карта памяти опять же определится как съемный носитель. Цифровые камеры охраняют изображения в различных графических форматах: как правило, в формате JPEG, либо в форматах TIFF или BMP. Оптимально, если на каждом снимке находится изображение одной страницы или книжного разворота.

В Интернете можно найти много отсканированных книг в формате DjVu. Часто изображения страниц сохраняют и в виде файлов PDF. Это многостраничные документы – в одном файле содержатся изображения всех страниц книги или журнала.

Чтобы распознать готовые изображения и передать результат в редактор Microsoft Word, выберите сценарий **PDF или изображения в Microsoft Word** или **Конвертировать фото в Microsoft Word**. Для **Конвертировать в PDF** откроется диалоговое окно выбора файлов (рис. 2.6).

В правой части окна находится область предварительного просмотра. Когда установлен флажок **Предварительный просмотр**, в ней выводится изображение открываемого документа.

Если вы открываете многостраничный документ (файл DjVu или PDF), в области просмотра появляется полоса прокрутки (рис. 2.7). Перемещая ползунок, вы можете просматривать разные страницы. В таком случае становится активен переключатель **Диапазон страниц**.

- Чтобы обработать и распознать все страницы, установите переключатель в положение **Все**.
- Чтобы распознать только некоторые страницы, установите переключатель в положение **Страницы:**. Введите в поле справа от переключателя

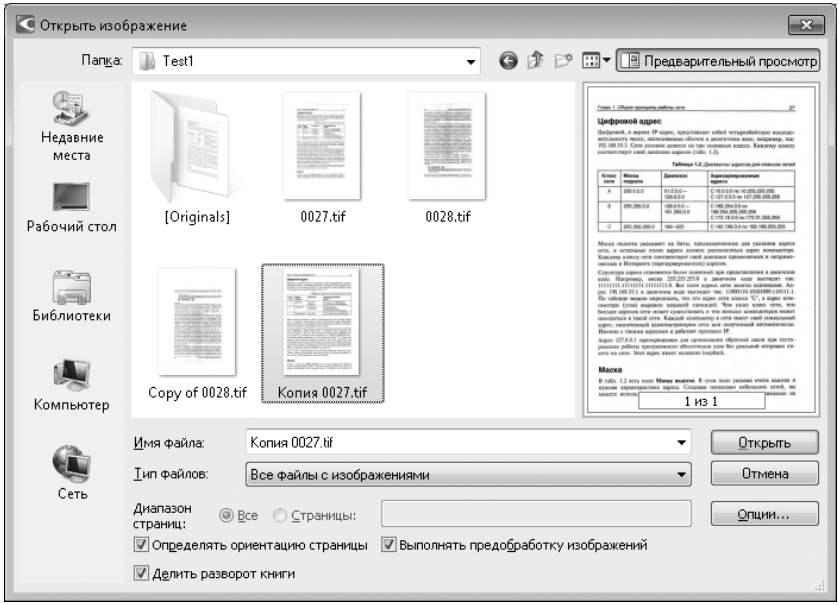


Рис. 2.6 ▾ Диалог **Открыть изображение**

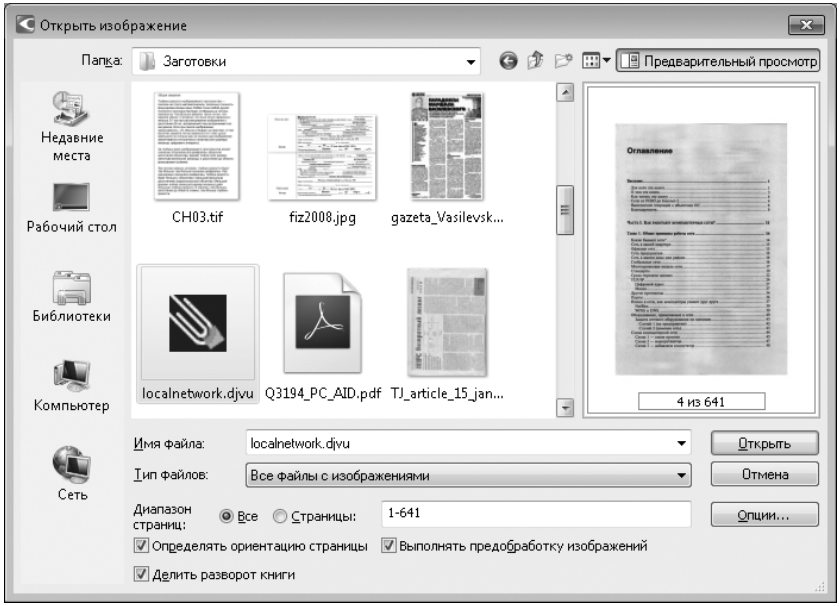


Рис. 2.7 ▾ Открытие многостраничного изображения

диапазон номеров страниц, например `2-10`, или номера отдельных страниц, например `2, 5-7`.

Когда документ PDF защищен автором от копирования, в области просмотра вы увидите предупреждение. Для открытия такого документа нужно ввести в появившемся диалоге пароль (рис. 2.8). Если этот пароль вам неизвестен, открыть и распознать документ не удастся. Естественно, документы защищают не для того, чтобы потом раздавать пароли всем желающим. Выход тем не менее есть! На помощь придет программа Screenshot Reader, которую мы рассмотрим в отдельной главе.

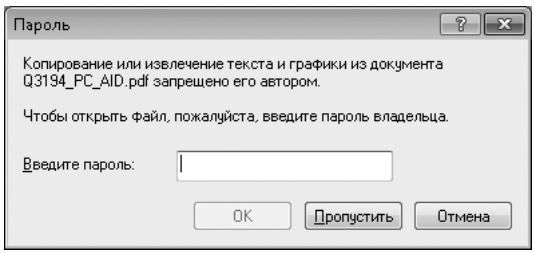


Рис. 2.8 ▾ Запрос пароля

Чтобы выбрать сразу несколько файлов, например фотографии нескольких страниц книги, щелкайте на значках файлов кнопкой мыши, одновременно удерживая клавишу **Ctrl** на клавиатуре. Выбрав файлы, нажмите кнопку **Открыть**.

Диалог **Открыть изображение** закроется. Изображения будут распознаны, а результат вы увидите в окне программы Microsoft Word. Проверьте, отредактируйте и сохраните документ средствами этого приложения.

Вызов сценариев из контекстного меню файла

Справедливо считают, что «продвинутый» пользователь работает правой кнопкой мыши гораздо чаще, чем новичок. Благодаря этому он совершает меньше лишних движений мыши! При щелчке правой кнопкой мыши на разных объектах открываются контекстные меню этих объектов, в которых, как правило, содержатся наиболее часто используемые команды.

Например, в окне **Проводника Windows** (иначе говоря – в окне открытой папки) щелкните правой кнопкой мыши на значке какого-нибудь файла. Откроется контекстное меню файла, в котором обязательно присутствуют команды **Открыть**, **Открыть с помощью...**, **Вырезать**, **Копировать**, **Удалить** и др.

Конкретный вид меню зависит от типа файла и того, какие приложения установлены в вашей системе. Многие программы добавляют в контекстные меню файлов собственные команды. После установки программы FineReader в контекстных меню файлов поддерживаемых графических форматов появляется группа из четырех команд (рис. 2.9).

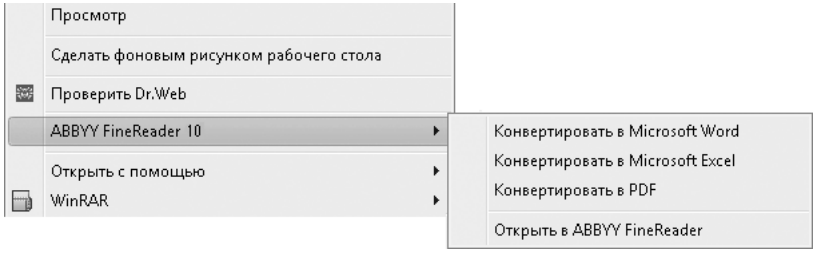


Рис. 2.9 ▾ Команды FineReader в контекстном меню файла

Первые три команды сразу запускают программу FineReader на выполнение названного сценария с выбранным файлом. Как работают эти сценарии, мы уже разобрались. Команда **Открыть в ABBYY FineReader** просто открывает файл в этой программе, чтобы затем вы могли выполнить обработку и распознавание в пошаговом режиме.

ПРИМЕЧАНИЕ

Команды ABBYY FineReader 10 добавляются в контекстное меню файлов только в 32-битных версиях операционной системы Windows. При использовании 64-битных версий Windows эти команды недоступны, и для запуска сценариев необходимо сначала запустить программу FineReader, а затем воспользоваться окном ***Новое задание****.*

Сканирование и сохранение изображений

Сценарий **Сканировать и сохранить изображение** предназначен для сохранения отсканированных изображений и выполняется без распознавания текста. Когда уже открыто окно программы FineReader, этот сценарий – самый простой и быстрый способ отсканировать какой-то документ и сохранить его изображение на диск компьютера. Это альтернатива «фирменным» программам, которыми обычно комплектуют сканеры, и стандартным средствам ОС Windows для получения изображений со сканера.

В операционной системе Windows для получения изображения со сканера предусмотрены встроенные средства. В Windows XP стандартная процедура сканирования запускается с помощью элемента **Пуск ➢ Панель управления ➢ Сканеры и камеры** и т. д., а в Windows 7 значок сканера находится в папке, открывающейся при выборе элемента меню **Пуск ➢ Устройства и принтеры**. По сравнению с запуском сценария программы FineReader, получение изображений встроенными средствами Windows потребует больше щелчков кнопками мыши, и этим способом пользуются довольно редко.

Фирменные утилиты сканеров, например Epson Presto Manager, позволяют детально настраивать параметры сканирования и сохранения изображений.

Благодаря этому для сканирования фотографий или других художественных оригиналов они зачастую оказываются удобнее.

Каким из трех перечисленных средств пользоваться – решайте сами, это дело привычек и личных предпочтений. Однако при наличии на компьютере программы FineReader очень удобно обратиться именно к ней.

1. Выберите сценарий **Сканировать и сохранить изображение**. Откроется диалоговое окно сканирования (рис. 2.2).
2. Отсканируйте документ или документы, как было показано выше. Закончив, нажмите кнопку **Закрыть**. Появится диалоговое окно сохранения изображений (рис. 2.10).

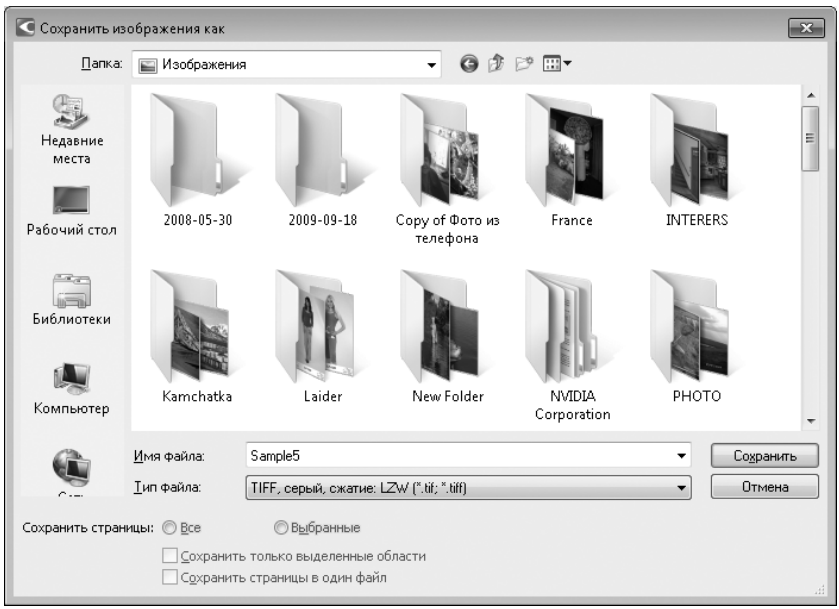


Рис. 2.10 ▾ Сохранение изображения в файл

3. Выберите папку, в которую надо сохранить изображения, и укажите имя файла.
   Чтобы создать новую папку непосредственно из этого диалогового окна, нажмите кнопку **Создание новой папки**. Заметьте, что в диалоговых окнах сохранения или открытия файлов можно удалять и переименовывать файлы и папки точно так же, как и в окне Проводника Windows.
4. В раскрывающемся списке **Тип файла** выберите желаемый формат. В отличие от многих других программ, в диалоге сохранения файлов программы FineReader 10 некоторые форматы приведены в списке несколько раз, с разными параметрами цветности и сжатия.

Если вы отсканировали несколько страниц, становятся активными элементы управления в нижней части окна сохранения файла. Переключатель **Сохранить страницы** при выполнении сценария использовать не удастся, да это и не нужно. Ведь если вы запустили сценарий и отсканировали несколько изображений, то, вероятно, хотите сохранить все полученные изображения.

Когда флажок **Сохранить страницы в один файл** снят или неактивен, каждое изображение сохраняется в отдельный файл. При этом к имени, которое вы ввели в поле **Имя файла**, автоматически добавятся четырехзначные номера страниц. Например, если вы отсканировали три страницы и указали имя файла **Sample 5**, то файлы получат имена **Sample 50001**, **Sample 50002** и **Sample 50003**.

Флажок **Сохранить страницы в один файл** становится доступен, если формат, выбранный в списке **Тип файла**, поддерживает многостраничность. Такими форматами являются TIFF и PDF. Установите флажок **Сохранить страницы в один файл**, и все изображения будут сохранены в один файл, как страницы.

5. Нажмите кнопку **Сохранить**.

   Изображения будут сохранены в файл (файлы) указанного формата. Впоследствии вы сможете просмотреть их, отправить по электронной почте, обработать в графическом редакторе или открыть в программе FineReader и распознать.

Резюме

Встроенные сценарии – быстрый способ начать работу с программой, не вникая в детали настроек. При распознавании с хороших и простых по структуре оригиналов этих сценариев бывает вполне достаточно.

Когда вам нужно отсканировать несколько страниц книги или журнала, чтобы потом использовать текст в реферате, в первую очередь попробуйте воспользоваться сценарием **Сканировать в Microsoft Word**. Для работы с оригиналом наподобие прайс-листа или накладной лучше подойдет сценарий **Сканировать в Microsoft Excel**. Если в получившемся документе практически нет ошибок, цель достигнута.

Если же вам перед вами оригинал нелучшего качества или у документа сложная структура (много иллюстраций, таблиц и т. д.), работать с ним лучше в пошаговом режиме. При этом появляется возможность настроить программу FineReader «под каждый конкретный случай» и тем самым улучшить качество распознавания. Эти шаги и все связанные с ними настройки рассмотрены в следующих главах нашей книги. Прежде всего выясним, что же такое «документ FineReader», и подробнее познакомимся с интерфейсом программы.

Глава 3 Работа в пошаговом режиме

С помощью встроенных сценариев хорошо работать с «удобными для распознавания» оригиналами. В предыдущей главе мы рассмотрели выполнение сценариев с такими настройками программы, которые приняты «по умолчанию» сразу после установки ABBYY FineReader 10. Во время работы сценария эти настройки изменить нельзя. Вы можете только выбрать некоторые параметры, когда при выполнении сценария на экране появляются диалоги сканирования, открытия или сохранения файла.

При работе со сложными для распознавания документами целесообразно держать процесс обработки под полным контролем. Тогда, в зависимости от характера оригинала и результатов, вы сможете настраивать программу в процессе работы с документом.

В этой главе мы познакомимся с устройством и назначением окон программы, некоторых элементов управления. В результате вы сможете настраивать внешний вид окон и панелей инструментов так, как вам удобно. Кроме того, выясним, что такое документ FineReader.

Окно программы и настройка рабочего пространства

Программа FineReader работает, как и любые системы OCR, в несколько этапов. При выполнении сценариев мы видим только первый и последний из них: все промежуточные операции происходят автоматически. Кратко рассмотрим, что это за этапы.

- Сначала в программу передается изображение: сканируется оригинал, или открывается уже существующий файл (файлы) с изображением. На

этом этапе настраивается сканер: можно изменить разрешение, цветность и яркость получаемого изображения.

- Изображение подготавливается к распознаванию: при необходимости оно поворачивается, выравнивается, устраняются геометрические искажения, изображение книжного разворота делится на две части, соответствующие страницам книги. В зависимости от настроек, заданных в диалоге **Опции** или диалоге открытия файла, предобработка изображения может выполняться автоматически. По умолчанию предлагается именно такой режим. В дополнение к автоматической предобработке, или вместо нее, пользователь может обработать изображение вручную с помощью встроенного редактора изображений.
- Затем на изображении определяются области, где расположены текст, таблицы и иллюстрации: происходит *анализ* страницы. По умолчанию анализ выполняется автоматически. Если в автоматическом режиме области были определены неправильно, то пользователь может самостоятельно переопределить эти области, изменить их границы.
- Далее происходит собственно распознавание текста. При этом изображения символов сравниваются с образцами (эталонами) шрифтов, которые могут встречаться в указанном языке или языках. Некоторые похожие по форме или нечетко изображенные символы могут распознаться неуверенно: в таком случае появляются несколько вариантов распознанного слова. Затем программа сравнивает распознанные слова со словарями и на этом основании предлагает наиболее вероятный вариант написания. При необходимости на этапе распознавания выбирают и настраивают эталоны и словари.
- Распознанный документ программа показывает пользователю. Видя результат распознавания, вы можете внести необходимые поправки в содержание и оформление. После корректировки вы можете сохранить результат в различных форматах, передать его в другое приложение для дальнейшего редактирования либо отправить его по электронной почте. Параметры сохранения в разные форматы также настраиваются.

Интерфейс программы FineReader построен так, чтобы вы могли одновременно видеть и входной документ (изображение), и результат распознавания (текст). В предыдущей главе мы уже бегло познакомились с окном **Новое задание**. Теперь рассмотрим интерфейс подробнее.

В верхней части главного окна программы расположена строка меню. С назначением пунктов меню ознакомимся на примерах в следующих главах, а пока остановимся на рабочих окнах, которые открываются в главном окне. Все окна, панели инструментов, их взаимное расположение и вид составляют *рабочее пространство* программы.

Рабочие окна

Чтобы увидеть все рабочие окна программы, необходимо сначала отсканировать хотя бы одну страницу или открыть хотя бы один файл с изображением. Поэтому в качестве примера отсканируем две страницы и оценим результат.

1. Запустите программу FineReader. Откроется главное окно программы (рис. 2.1).
2. Положите оригинал в сканер.
3. Нажмите на главной панели инструментов кнопку **Сканировать**. Откроется диалоговое окно сканирования (рис. 2.3).
4. Отсканируйте изображение, как было показано в предыдущей главе. Отсканируйте еще одну страницу. Закройте диалог сканирования.

По умолчанию программа сразу обрабатывает, анализирует и распознает полученные изображения документа. В результате в главном окне программы открываются три рабочих окна: **Страницы**, **Изображение** и **Текст** (рис. 3.1). Кроме того, в нижней части главного окна находится полоса, при нажатии на которую на экран выводится окно **Крупный план**.

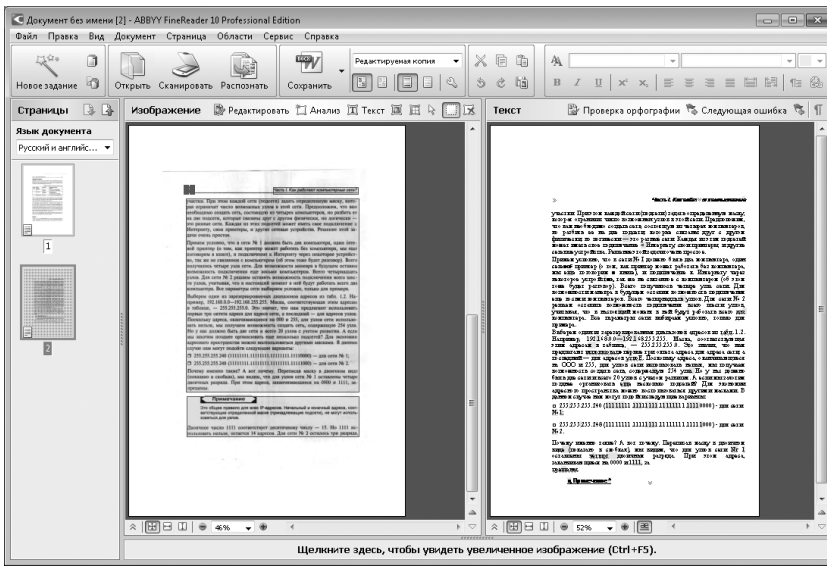


Рис. 3.1 ▾ Три рабочих окна в главном окне программы

В окнах **Изображение** и **Текст** отображается та страница оригинала, которая выбрана в окне **Страницы**. Чтобы работать с первой страницей, в окне **Страницы** щелкните кнопкой мыши на эскизе первой страницы, со второй – на эскизе второй страницы и т. д. Рассмотрим рабочие окна по порядку.

Окно Страницы

По умолчанию в окне **Страницы** страницы оригинала изображаются в виде эскизов, или пиктограмм. В левом нижнем углу эскиза каждой страницы выводится маленький значок, который показывает стадию обработки этой страницы.

- ❑ Если значок отсутствует, страница еще не обработана.

-  – страница проанализирована, на ней выделены области с текстом, таблицами и т. д.
-  – страница распознана.
- Если в процессе автоматического анализа или распознавания страницы возникли проблемы, в правом нижнем углу эскиза отображается желтый треугольник с восклицательным знаком. По умолчанию при выделении в окне **Страницы** страницы с таким значком всплывает панель предупреждений, которая содержит сообщения об ошибках, возникших в процессе обработки страниц документа.

В режиме Пиктограммы вы можете изменять порядок страниц простым перетаскиванием одной или нескольких выделенных страниц в нужное место в документе.

Другое представление страниц оригинала в окне **Страницы** – в виде таблицы. Чтобы изменить вид окна **Страницы**, щелкните кнопкой мыши на пункте **Вид** в строке меню. Раскроется меню (рис. 3.2), в котором выберите команду **Окно Страницы ➢ Таблица** (или воспользуйтесь контекстным меню окна **Страницы**).

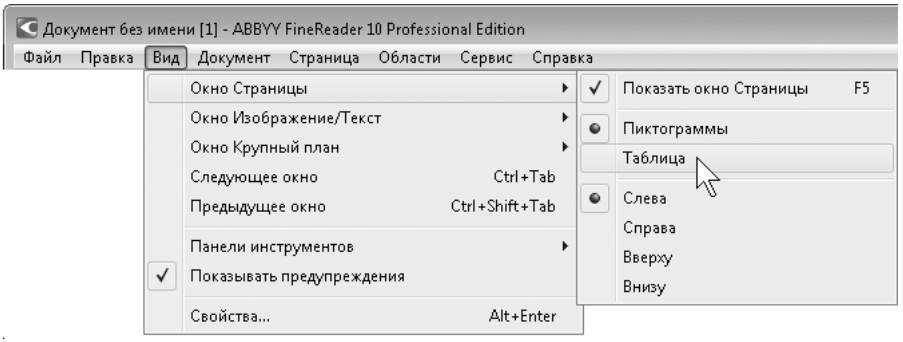


Рис. 3.2 ▾ Меню **Вид**

В результате в окне **Страницы** вы увидите таблицу, в которой напротив значка каждой страницы приводятся свойства этой страницы (рис. 3.3).

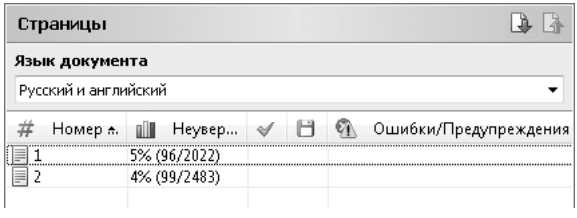


Рис. 3.3 ▾ Окно **Страница** – вид **Таблица**

Среди свойств страницы наиболее интересными являются сведения в колонке **Неуверенно распознанные символы**. В примере на рис. 3.3 видно, что на первой странице из 2022 символов неуверенно распозналось 96 (5% от общего числа знаков). На второй странице неуверенно распознано всего 4%, то есть на этой странице число возможных ошибок меньше.

В следующих колонках метки появляются после того, как страница была проверена, результаты распознавания сохранены в файл. В колонке **Ошибки/предупреждения** отображаются предупреждения о проблемах, возникших при распознавании страницы.

Еще правее расположены колонки **Комментарии** и **Источник**. В нашем примере они скрыты за правой границей окна **Страницы**. Чтобы увидеть эти колонки, следует воспользоваться горизонтальной полосой прокрутки в нижней части окна, либо расширить окно, перетаскивая мышью его правую границу. В колонку **Комментарии** вы можете вводить собственные замечания, а в колонке **Источник** приводится источник изображения: сканер либо имя открытого файла с изображением.

Если документ состоит из небольшого числа страниц, удобно видеть их в окне **Страницы** в виде пиктограмм: все наглядно и занимает минимум места на экране. Если же страниц в документе много, вид **Таблица** бывает полезнее. Вы сразу заметите страницы с большим количеством неуверенно распознанных символов. Такие страницы, скорее всего, придется отсканировать и распознать повторно. Кроме того, по таблице проще контролировать, какие страницы вы уже проверили, а какие еще нет, все ли распознанные страницы были сохранены. Чтобы удобнее было работать с таблицей, ее можно переместить в верхнюю часть рабочего пространства. Для этого в контекстном меню окна **Страницы** выберите команду **Окно Страницы ➢ Вверху**.

Переключать вид окна **Страницы** можно и другим способом: щелкните правой кнопкой мыши внутри этого окна. В контекстном меню выберите команду **Окно Страницы ➢ Пиктограммы** или **Окно Страницы ➢ Таблица**.

Окно Изображение

В окне **Изображение** (рис. 3.1) происходит работа с изображением оригинала. В нашем примере программа уже автоматически выделила на нем области, которые должны распознаваться как текст, таблицы, картинки или штрих-код.

В верхней части окна находится панель инструментов **Изображение**. С помощью ее кнопок вы можете вызывать редактор изображений, выделять области различных типов, запускать анализ страницы. Подробнее эти процедуры рассмотрены в главе «Обработка и анализ изображений».

Ниже находится рабочая область окна, в которой вы видите изображение оригинала. В процессе автоматического анализа документа или вручную на нем цветными рамками выделяются отдельные области (Текст, Таблица, Картинка, Штрих-код). Кроме того, пунктирная рамка со значком лупы обозначает участок изображения, который выводится в окне **Крупный план**, когда это окно включено.

В нижней части окна **Изображение** расположена небольшая панель для управления масштабом изображения. Кнопки на этой панели позволяют выбрать наиболее удобный размер изображения.

1. Нажмите кнопку **Целая страница**. Масштаб будет подобран так, чтобы в рабочей области окна была видна страница целиком.
2. Нажмите кнопку **По ширине**. Масштаб изменится, чтобы в рабочей области страница помещалась по ширине. В этом случае для просмотра невидимой части пользуйтесь полосой прокрутки или колесиком мыши. Вертикальная полоса прокрутки также позволяет переходить от страницы к странице. Кроме того, при нажатой клавише **Пробел** указатель мыши приобретает форму руки, и с помощью этого инструмента вы можете перемещать изображение внутри окна во всех направлениях.
3. Нажмите кнопку **По высоте**. Масштаб будет подобран так, чтобы в рабочей области окна страница уместилась по высоте.

Правее этих кнопок находится раскрывающийся список **Масштаб**. В нем вы можете выбрать одно из стандартных значений: от 25% до 200% – либо ввести нужное значение в процентах непосредственно в поле масштаба. Изменять масштаб можно и другими способами:

1. Щелкните правой кнопкой мыши внутри окна **Изображение**. В контекстном меню выберите команду **Масштаб**, а в дочернем меню выберите нужное значение.
2. Щелкните кнопкой мыши внутри окна **Изображение**. Нажмите клавишу **Ctrl**. Удерживая ее, вращайте колесико мыши. Изображение в окне будет плавно увеличиваться или уменьшаться.

По умолчанию масштаб изображения подбирается так, чтобы в рабочей области окна страница была видна целиком. В большинстве случаев такой масштаб является оптимальным. При необходимости рассмотреть мелкие детали или точно установить границы областей удобнее пользоваться окном **Крупный план**.

В нижнем левом углу окна **Изображение** находится кнопка-переключатель **Показать свойства области**. Она скрывает или разворачивает панель **Свойства**. Когда панель **Свойства** открыта, рисунок кнопки изменяется – стрелки на ней направлены вниз.

На двух вкладках панели **Свойства**, расположенной в нижней части окна **Изображение**, приводятся сведения об изображении в целом и отдельных его областях. Элементы управления, находящиеся на вкладках **Свойства области** и **Свойства изображения**, позволяют регулировать некоторые свойства областей и изображения в целом и тем самым влиять на качество распознавания. Скрыв панель **Свойства**, вы немного увеличиваете размер рабочей области окна, тем самым видимое изображение становится крупнее.

Окно Текст

В окне **Текст** выводится распознанный текст документа. Здесь производятся правка и форматирование текста, проверка правописания. В нашем примере распознавание уже выполнено в автоматическом режиме, документ отобража-

ется в таком виде, в каком он будет передаваться в другие приложения или сохраняться в файл.

Окно **Текст** организовано так же, как и окно **Изображение**. В верхней части – панель инструментов, кнопки на которой служат для проверки текста и отображения непечатаемых знаков, ниже – рабочая область, под рабочей областью – панель масштабирования. В самом низу окна находится панель свойств текста. Эта панель, подобно панели свойств окна **Изображение**, включается и выключается кнопкой **Показать свойства текста**.

Изменение масштаба производится кнопками панели масштабирования, командой **Масштаб** контекстного меню или вращением колесика мыши при нажатой клавише **Ctrl**. Для прокрутки текста, не умещающегося в окне, служат полосы прокрутки по краям окна или колесико мыши.

Возможны ситуации, когда распознанный текст из какой-то текстовой области не умещается на выделенном для него участке страницы. В таких случаях в окне **Текст** на границе текстового фрагмента появляются кнопки с красными стрелочками. Нажатие на эти кнопки позволяет просмотреть и отредактировать текст, выходящий за пределы участка страницы.

Окно Крупный план

Чтобы увидеть окно **Крупный план**, щелкните кнопкой мыши на полосе с надписью **Щелкните здесь, чтобы увидеть увеличенное изображение (Ctrl+F5)**. Как следует из подсказки, с той же целью вы можете нажать сочетание клавиш **Ctrl+F5**. Повторное нажатие сочетания клавиш скрывает это окно, и на экране вновь отображается полоса у нижней границы главного окна программы.

В окне **Крупный план** вы видите фрагмент изображения оригинала. В окне **Изображение** эта часть изображения выделяется пунктирной рамкой. Рабочая область окна **Крупный план** дублирует часть рабочей области окна **Изображение**. По сути, это окно играет роль «увеличительного стекла» для окна **Изображение**. Помимо рабочей области, в этом окне есть только панель управления масштабом изображения, а также возможность работы с областями при помощи контекстного меню.

Чтобы увидеть в окне **Крупный план** изображение в натуральную величину, нажмите кнопку **1:1 С точностью до пикселя**. Для уменьшения или увеличения масштаба выберите одно из значений в раскрывающемся списке **Масштаб**. Масштаб также изменяется через контекстное меню (оно точно такое же, как и в окне **Изображение**) или колесиком мыши при нажатой клавише **Ctrl**.

Изображение и тест в рабочей области окон **Изображение**, **Крупный план** и **Текст** прокручиваются синхронно. Благодаря этому вы всегда видите во всех трех окнах одну и ту же часть документа, что очень удобно при вычитке и проверке текста.

Щелкните кнопкой мыши в окне **Текст** на каком-нибудь слове в распознанном тексте. При этом в окнах **Изображение** и **Крупный план** будет показан участок изображения, где находится это слово. Точно так же щелкните кнопкой мыши на каком-нибудь участке оригинала в окне **Изображение**: распознан-

ный текст в окне **Текст** и изображение в окне **Крупный план** прокрутятся на соответствующее место.

Изменение расположения рабочих окон

Размер и расположение рабочих окон рассчитаны на монитор с минимальным разрешением 1024×768 точек. На таком экране хорошо видны и изображения, и текст, и все кнопки и панели инструментов помещаются в видимой области. Для удобства пользователя размер и взаимное расположение рабочих окон можно менять по ходу работы. На мониторе с меньшим экраном для комфортной работы, скорее всего, потребуется подстраивать расположение и вид рабочих окон.

Например, при разметке областей на изображении стоит расширить окно **Изображение**. При вычитке и корректировке распознанного текста, наоборот, удобнее расширить окно **Текст**, чтобы видеть в нем как можно большую часть текста. В любом случае, постарайтесь подстроить размеры и положение рабочих окон программы под свой монитор и под свое зрение. Также удобно временно скрывать не используемые в данный момент окна.

Для изменения размеров окон наведите указатель мыши на разделитель между рабочими окнами. Указатель превратится в двунаправленную стрелку. Нажмите кнопку мыши и перетащите разделитель. Отрегулируйте ширину и высоту окон так, как вам удобно.

Если вы перетащите вертикальный разделитель между окнами **Изображение** и **Текст** влево до предела, окно **Текст** максимально расширится, а окно **Изображение** скроется (рис. 3.4). Вместо него появится полоса с подсказкой: **Щелкните здесь, чтобы отредактировать области** (**F6**).

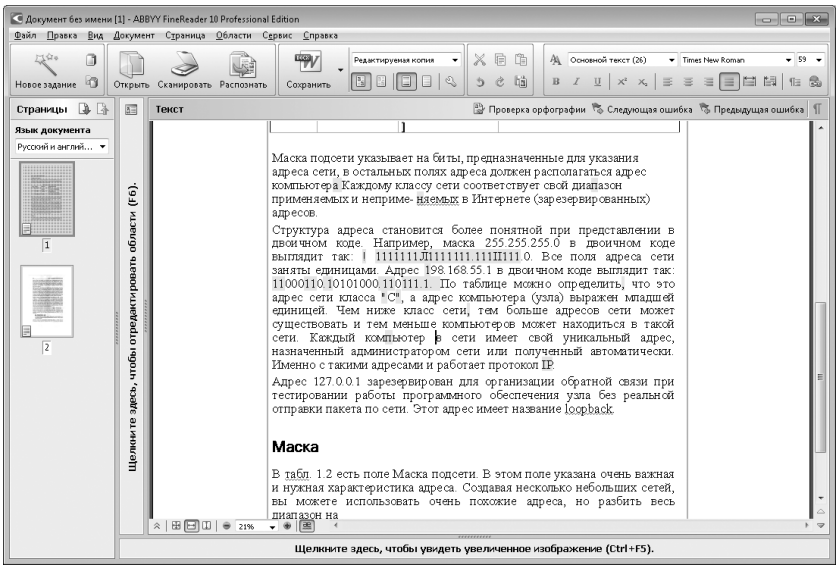


Рис. 3.4 ▾ Окно **Текст** развернуто на всю ширину

Последуйте подсказке и щелкните на этой полосе кнопкой мыши, либо нажмите клавишу **F6**. В результате окно **Изображение** развернется на полную ширину, а окно **Текст** свернется в полоску у правого края главного окна программы (рис. 3.5). На этой полосе написано: **Щелкните здесь, чтобы отредактировать результаты распознавания (F8)**.

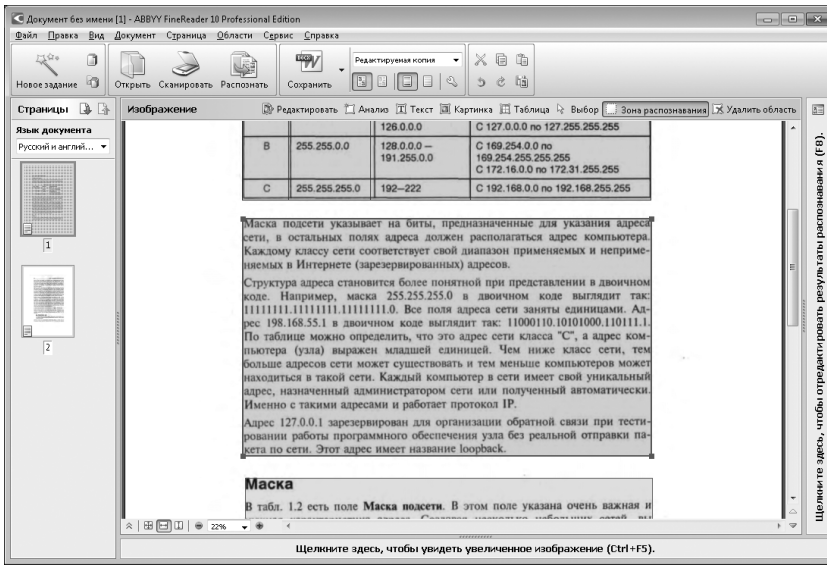


Рис. 3.5 ▾ Окно **Изображение** развернуто на всю ширину

Чтобы вновь увидеть рабочие окна **Изображение** и **Текст** рядом, как на рис. 3.1, перетащите мышью вертикальный разделитель к середине главного окна программы. Перетаскивая мышью горизонтальный разделитель, ограничивающий сверху окно **Крупный план**, отрегулируйте высоту этого окна.

Рабочее пространство удобно настраивать и через меню **Вид**. Щелкните кнопкой мыши в строке меню на пункте **Вид** и разверните подменю **Окно Изображение/Текст** (рис. 3.6).

Первые три команды меню приводят к тем же результатам, которые мы только что получили, перетаскивая мышью вертикальный разделитель между окнами. Рядом с командами приведены подсказки о том, какая клавиша служит для быстрого вызова этой команды.

- Чтобы развернуть окно **Изображение** в полную ширину, выберите команду **Показать окно Изображение**. Чтобы сделать это, не обращаясь к меню, нажмите клавишу **F6**.
- Чтобы расположить окна **Изображение** и **Текст** рядом, выберите команду **Показать окна Изображение и Текст**, или нажмите клавишу **F7**.
- Чтобы развернуть окно **Текст** в полную ширину, выберите команду **Показать окно Текст**, или нажмите клавишу **F8**.

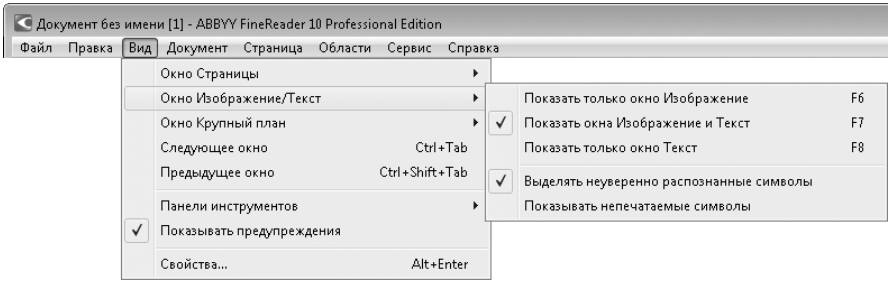


Рис. 3.6 ▾ Меню **Вид** ➢ Окно **Изображение/Текст**

Еще две команды настраивают отображение распознанного текста в окне **Текст**.

- Когда выбрана команда (установлен флажок) **Выделять неуверенно распознанные символы**, в окне **Текст** выделяются неуверенно распознанные символы и несловарные слова. По умолчанию они подсвечиваются голубым фоном, но цвет выделения можно настраивать в диалоге **Опции**.
- Команда **Показывать непечатные символы** дает указание программе показывать в распознанном тексте специальными значками такие символы, как пробелы, символы конца строки и абзаца. Отображение непечатаемых символов часто помогает в процессе окончательного редактирования распознанного текста.

Подменю **Окно Крупный план** содержит три команды и вложенное меню **Масштаб** (рис. 3.7). С их помощью настраивается вид рабочего окна **Крупный план**.

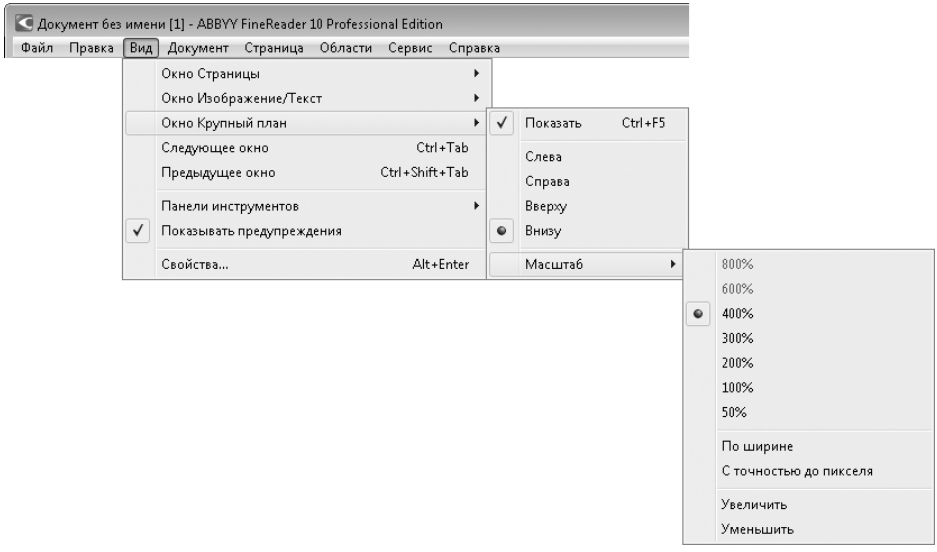


Рис. 3.7 ▾ Меню **Вид** ➢ Окно **Крупный план**

Команда **Показать** переключает видимость окна **Крупный план**. Чтобы скрыть окно, снимите флажок напротив этой команды. Чтобы вывести окно на экран, установите флажок. Кроме того, для скрытия и отображения этого окна вы можете одновременно нажимать на клавиатуре клавиши **Ctrl** и **F5**.

Чтобы расположить окно **Крупный план** в нижней части главного окна программы, под окнами **Изображение** и **Текст**, выберите команду **Внизу**. Это настройка, принятая по умолчанию. Чтобы окно **Крупный план** пристыковалось к левой, правой или верхней части главного окна программы, выберите команду **Слева**, **Справа** или **Вверху**.

Вложенное меню **Масштаб** изменяет масштаб изображения в окне **Крупный план**. Это четвертый способ регулирования масштаба изображения: еще три мы перечислили чуть раньше.

На обычном мониторе с соотношением сторон экрана 4:3 окно **Крупный план** помогает разглядеть мелкие детали изображения (рис. 3.8). Если же вы пользуетесь широкоэкранным монитором с диагональю 19" и более, окно **Крупный план** в большинстве случаев удобнее скрывать, как это и сделано по умолчанию. Высокое разрешение и большая ширина экрана позволяют детально рассмотреть все содержимое непосредственно в окнах **Изображение** и **Текст**.

Попробуйте по-разному настраивать рабочее пространство программы FineReader. Оптимальный для себя вариант вы найдете по мере дальнейшего знакомства с программой и практической работы.

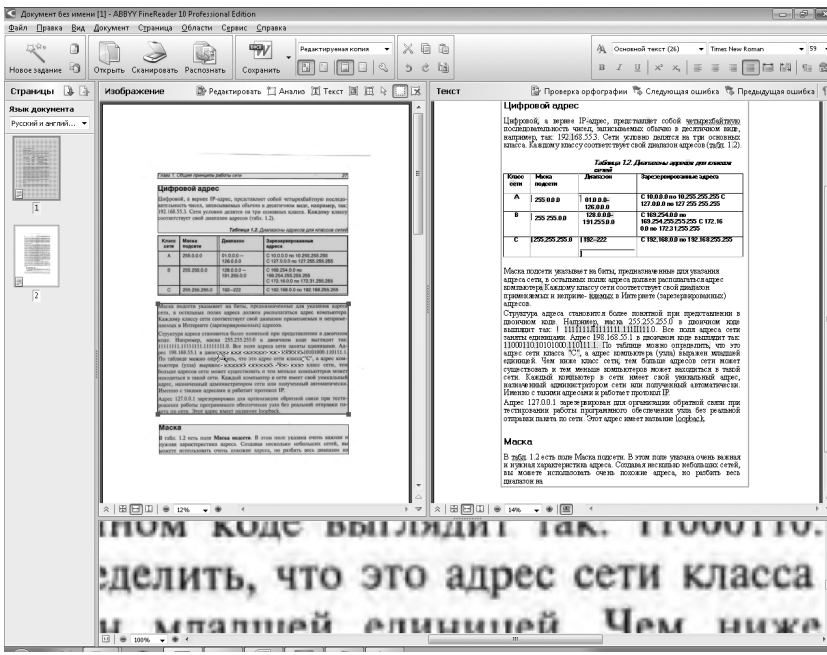


Рис. 3.8 ▾ Все рабочие окна открыты – настройка для монитора 4:3

Настройка панелей инструментов

По умолчанию в верхней части главного окна программы отображается главная панель инструментов. Расположенные на ней кнопки служат для обращения к наиболее часто используемым функциям программы. Все те же функции могут быть вызваны и из строки меню, но многим удобнее пользоваться именно кнопками на панелях инструментов.

При необходимости, например чтобы максимально использовать площадь экрана для размещения рабочих окон, главную панель инструментов можно скрыть. Для этого щелкните правой кнопкой мыши в любом месте панели инструментов и в контекстном меню снимите флажок у команды **Главная**. Главная панель инструментов исчезнет с экрана. Чтобы вновь показать главную панель инструментов, в меню **Вид** выберите команду **Панели инструментов ➢ Главная**.

В интерфейсе программы FineReader есть настраиваемая панель инструментов **Быстрый доступ**. По умолчанию она скрыта. На панель инструментов **Быстрый доступ** можно поместить кнопки в соответствии со своими предпочтениями.

Чтобы отобразить панель инструментов **Быстрый доступ**, выберите команду меню **Вид ➢ Панели инструментов ➢ Быстрый доступ**. Панель инструментов **Быстрый доступ** появится в главном окне программы между строкой меню и главной панелью инструментов (рис. 3.9).

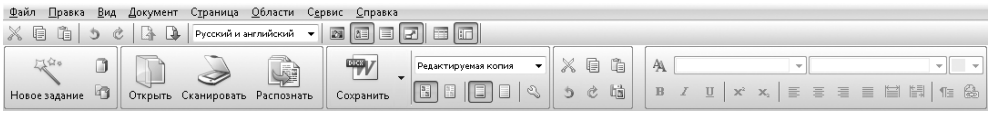


Рис. 3.9 ▾ Строка меню и панели инструментов

Рассмотрим кнопки, присутствующие на панели быстрого доступа по умолчанию.

- Первые три кнопки: **Вырезать** (**Ctrl+X**), **Копировать** (**Ctrl+C**) и **Вставить** (**Ctrl+V**) служат для вырезания, копирования и вставки фрагментов изображения или текста. Точно такие же кнопки есть и на главной панели инструментов.
- Следующие две кнопки: **Отменить** (**Ctrl+Z**) и **Восстановить** (**Ctrl+Enter**) – позволяют отменить последнее действие, выполненное в любом окне программы. или вернуть обратно результат отмененного действия.

  Назначение этих кнопок наверняка известно вам по приложениям Microsoft Office – во многих программах Windows операции копирования, вставки и отмены последнего действия являются почти стандартными. Точно так же фактическим стандартом стали сочетания клавиш, которые нужно нажать для выполнения этих действий. На практике при наборе или правке текста кнопками **Вырезать**, **Копировать** и **Вставить**

на панели инструментов пользуются редко – гораздо удобнее и быстрее нажать соответствующее сочетание клавиш.

Например, работая в окне **Текст**, вы хотите скопировать какое-то предложение (фрагмент текста) и вставить его в другое место этого же текста. Такая операция при редактировании текстов выполняется довольно часто.

1. Выделите фрагмент текста: наведите указатель мыши на его начало и щелкните кнопкой мыши. Удерживая кнопку мыши, переместите указатель на окончание выделяемого текста и отпустите кнопку. Слова выделены.
2. Нажмите кнопку **Копировать** на панели инструментов или сочетание клавиш **Ctrl** и **C** на клавиатуре. Выделенный фрагмент скопирован в буфер обмена.
3. Наведите указатель мыши на то место, куда вы хотите вставить текст, и щелкните кнопкой мыши. Нажмите кнопку **Вставить** на панели инструментов или сочетание клавиш **Ctrl** и **V** на клавиатуре. Фрагмент из буфера обмена вставлен в указанное место.

Если вы решите, что вставили текст не туда, операцию можно отменить. Для этого нажмите кнопку **Отменить** на панели инструментов или сочетание клавиш **Ctrl** и **Z** на клавиатуре. Вставленный текст исчезнет, и документ вернется в предыдущее состояние.

Точно так же текст, скопированный из окна программы FineReader, вставляется в другие приложения.

1. Скопируйте фрагмент текста, как показано выше.
2. Запустите другое приложение, например Блокнот или Microsoft Word. При запуске этих программ в их окнах по умолчанию открывается новый пустой документ.
3. Наведите указатель мыши на то место в документе, куда вы хотите вставить текст, и щелкните кнопкой мыши. Нажмите кнопку **Вставить** на панели инструментов или сочетание клавиш **Ctrl** и **V** на клавиатуре. Фрагмент из буфера обмена будет вставлен в указанное место.

- Кнопки **Предыдущая страница** и **Следующая страница** переключают страницы документа для отображения во всех рабочих окнах программы. Этими кнопками удобно пользоваться, когда окно **Страницы** скрыто и нельзя выбрать в нем страницу щелчком кнопки мыши. Кроме того, для навигации между страницами документа можно пользоваться колесиком мыши, полосой прокрутки в окне **Изображение**, клавишами **Page Up** и **Page Down**.
- Раскрывающийся список выбора языков дублирует такой же элемент управления в рабочем окне **Страницы**.
- Четыре кнопки управляют видимостью рабочих окон. Как мы уже убедились, скрывать и разворачивать окна можно разными способами – командами меню **Вид**, перетаскиванием разделительных полос, «быстрыми клавишами». Кнопки на панели **Быстрый доступ** – еще один способ настройки рабочего пространства.

- Две последние кнопки переключают режим отображения страниц в окне **Страницы**.

Панель быстрого доступа можно настраивать – добавлять на нее новые и удалять существующие кнопки. Возможно, вы захотите изменить набор кнопок сообразно своим целям и привычкам. Например, тем, кто привык пользоваться для копирования и вставки сочетаниями клавиш, соответствующие кнопки на панели инструментов попросту не нужны. Чем меньше кнопок на панели, тем проще в них ориентироваться. С другой стороны, если по характеру работы вы часто обращаетесь к каким-либо функциям, а кнопки для вызова этих функций на панели инструментов по умолчанию отсутствуют, такие кнопки целесообразно туда поместить. Пусть все команды присутствуют в меню, но нажать одну кнопку в непосредственной близости от рабочей области окна быстрее, чем каждый раз «тянуться мышью» к строке меню, а потом выбирать там нужный пункт.

Рассмотрим, как удалять и добавлять кнопки панели инструментов **Быстрый доступ**. Предположим, что вы решили убрать кнопки копирования и вставки и, наоборот, добавить две кнопки, весьма полезные при работе с низкокачественными оригиналами. Все это делается в диалоговом окне **Настройка**. Сначала уберем три кнопки с панели инструментов.

Вызовите диалоговое окно **Настройка**. Это можно сделать несколькими способами:

- выберите в меню **Вид** команду **Панели инструментов ➤ Настройка**;
- щелкните правой кнопкой мыши на любой панели инструментов и в контекстном меню выберите команду **Настройка**;
- в меню **Сервис** выберите команду **Настройка**.

Откроется диалоговое окно **Настройка панелей инструментов и горячих клавиш** (рис. 3.10) с двумя вкладками. На вкладке **Панели инструментов** на-

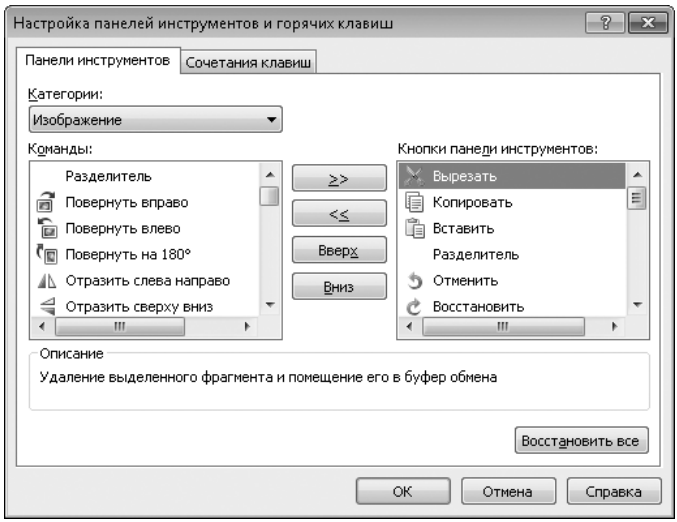
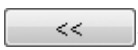


Рис. 3.10 ▾ Диалог **Настройка**, вкладка **Панели инструментов**

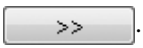
страиваются кнопки панели инструментов **Быстрый доступ**. На вкладке **Сочетания клавиш** задаются сочетания клавиш для выполнения любых действий, доступных в программе.

На вкладке **Панели инструментов** слева перечислены все существующие в программе FineReader команды (любую из них вы найдете в меню главного окна). Для удобства команды сгруппированы по категориям. Справа перечислены кнопки для вызова команд, присутствующие в данный момент на панели инструментов **Быстрый доступ**. В центре расположены кнопки, позволяющие добавлять кнопки (команды) на панель инструментов или убирать их оттуда. Кнопки **Вверх** и **Вниз** служат для сортировки и перемещения кнопок (команд) внутри панели инструментов. В списке кнопки перечислены в том же порядке, в каком они расположены на панели инструментов.

1. В перечне **Кнопки панели инструментов:** щелкните кнопкой мыши на записи **Вырезать**. Запись выделена.
2. Нажмите кнопку 

, удаляющую выбранную команду (кнопку) с панели инструментов. Запись **Вырезать** исчезнет из списка.
3. Точно так же выделите и удалите из списка записи **Копировать** и **Вставить**.
4. Нажмите кнопку **ОК**. Диалог **Настройка** закроется, а названные три кнопки пропадут с панели инструментов **Быстрый доступ**.

Теперь добавим две кнопки (команды).

1. Вызовите диалоговое окно **Настройка**.
2. В раскрывающемся списке **Категории** выберите категорию **Изображение**. В группе **Команды:** появится перечень команд, относящихся к этой категории. Если вы не помните, к какой категории относится нужная команда, выберите категорию **Все команды**. В этом случае в группе **Команды:** будут перечислены в алфавитном порядке все команды, существующие в программе.
3. В группе **Команды** щелкните кнопкой мыши на записи **Исправить искажение строк**.
4. Нажмите кнопку 

. Выбранная команда, точнее кнопка для ее вызова, будет перенесена на панель инструментов. В группе **Кнопки панели инструментов:** появится новая запись.
5. В группе **Кнопки панели инструментов:** щелкните кнопкой мыши на записи **Исправить искажение строк**. Она будет выделена. Нажимая кнопки **Вверх** и **Вниз**, переместите эту запись в нужное положение, например сделайте ее второй снизу.
6. Таким же образом поместите на панель инструментов **Изображение** команду (кнопку) **Исправить перекос изображения**.
7. Нажмите кнопку **ОК**. Диалог **Настройка** закроется, а на панели инструментов **Быстрый доступ** появятся две новые кнопки.

Чтобы быстро вернуть набор кнопок на панели инструментов к значениям по умолчанию, в диалоге **Настройка** предусмотрена кнопка **Восстановить все**. Нажмите ее, и на панели **Быстрый доступ** будет восстановлен набор кнопок,

принятый по умолчанию. Поэтому с настройкой панели **Быстрый доступ** можно смело экспериментировать – вы всегда сможете вернуть ее в первоначальный вид.

Как вы уже заметили, почти любое действие в программе FineReader, как и в других приложениях, можно выполнить несколькими разными способами. Какой способ предпочесть – зависит от ваших привычек и стиля работы.

- Одни пользователи предпочитают всегда вызывать любые команды из строки меню.
- Другие больше всего любят кнопки на панелях инструментов – это наглядно, хотя на кнопку еще нужно попасть указателем мыши.
- Третьи стараются при возможности вызывать команды из контекстных меню. Работает правая кнопка, мышь совершает минимум перемещений, однако в контекстных меню присутствуют не все команды. Если обе руки до этого лежали на клавиатуре, приходится брать мышь – это дополнительное движение.
- Четвертые для вызова команд активно пользуются сочетаниями «быстрых клавиш». Это особенно удобно, когда вы печатаете двумя руками: не нужно переносить правую руку на мышиный коврик. Проблема лишь в том, что сочетания клавиш нужно запомнить. Однако при регулярной работе затраченное на запоминание время окупается: меньше устают и глаза, и правая рука.

Пожалуй, самым профессиональным можно назвать стиль работы, когда человек равно владеет всеми четырьмя приемами. Такой пользователь в каждом случае выбирает наиболее экономный способ. Если правая рука в данный момент держит мышь – вызываем контекстное меню или нажимаем кнопки на панелях инструментов. Если обе руки лежат на клавиатуре – пользуемся быстрыми клавишами. Если нужна какая-то редкая команда, которой нет ни в контекстных меню, ни среди «быстрых клавиш», – обращаемся к строке меню.

Быстрые клавиши в программе FineReader задаются на вкладке **Сочетания клавиш** диалогового окна **Настройка панелей инструментов и горячих клавиш**. При этом для вызова любого действия может быть назначено одно или несколько сочетаний клавиш. Обычно сочетают нажатие клавиш **Alt**, **Shift** и одной из буквенных или цифровых клавиш. В качестве примера рассмотрим такую ситуацию.

В программе FineReader 10 по умолчанию для возврата отмененного действия служит любое из двух сочетаний: **Ctrl+Enter** либо **Ctrl+Y**. В программе Microsoft Word сочетание **Ctrl+Y** тоже означает возврат отмененного действия, но сочетание **Ctrl+Enter** служит для вставки разрыва страницы (в программе FineReader такое действие не предусмотрено). Если вы часто и поочередно работаете в этих приложениях, легко запутаться в двух комбинациях клавиш. Поэтому для возврата действия проще оставить лишь одно, «офисное» сочетание **Ctrl+Y**, а сочетание **Ctrl+Enter** пусть использует только программа Microsoft Word для вставки разрыва страниц.

1. Вызовите диалоговое окно **Настройка**. Перейдите в нем на вкладку **Сочетания клавиш** (рис. 3.11).

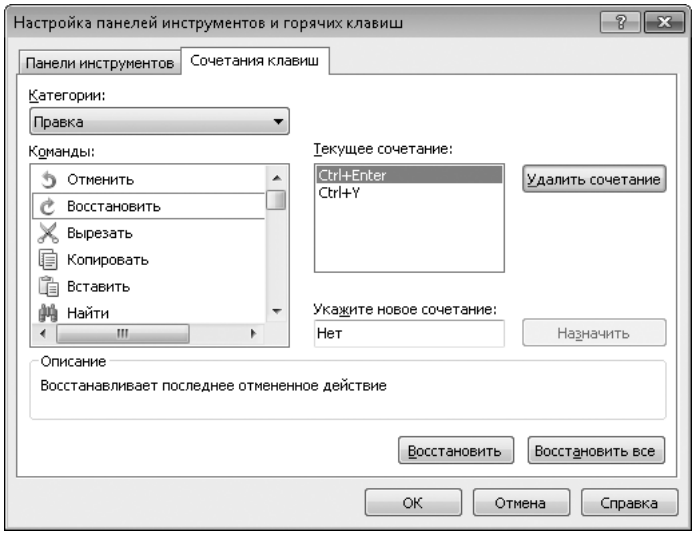


Рис. 3.11 ▾ Диалог **Настройка**, вкладка **Сочетания клавиш**

2. В раскрывающемся списке **Категории** выберите категорию **Правка** или **Все команды**. В группе **Команды:** появится перечень команд.
3. Выберите в группе **Команды** ту команду, для которой вы хотите изменить или назначить сочетание клавиш. В данном случае это команда **Восстановить**. В поле **Текущее сочетание** будут показаны присвоенные этой команде сочетания клавиш.
4. Выделите запись **Ctrl+Enter** и нажмите кнопку **Удалить сочетание**.
5. Нажмите кнопку **OK**. Диалоговое окно закроется, а настройки сохранятся.

Чтобы задать свое сочетание клавиш для вызова какой-либо команды, выберите эту команду в группе **Команды**. Затем щелкните кнопкой мыши в поле ввода **Укажите новое сочетание:** и наберите на клавиатуре желаемую комбинацию. Нажмите кнопку **Назначить**, и новая комбинация добавится в поле **Текущее сочетание**.

В программе FineReader 10 по умолчанию уже заданы «горячие клавиши» для всех часто вызываемых функций. Практика показывает, что менять или дополнять их обычно не требуется. Как правило, эти сочетания обоснованы с точки зрения эргономики: при правильной постановке рук для «слепой печати» использовать их удобно.

Чтобы вернуть сочетание клавиш для какой-либо команды к стандартному, выберите эту команду в списке и нажмите кнопку **Восстановить**. Чтобы вернуть к значениям по умолчанию сочетания клавиш для всех команд, нажмите кнопку **Восстановить все**.

Полный список стандартных сочетаний приведен в справочной системе программы. Возможно, его стоит крупно распечатать на листе бумаги и дер-

жать в пределах видимости. Периодически заглядывая в шпаргалку, вы легко запомните комбинации клавиш для тех команд, к которым вы обращаетесь чаще всего. Естественно, ощутимую пользу это принесет при регулярной и интенсивной работе с программой – если за месяц требуется распознавать два десятка страниц, то можно обойтись лишь кнопками панелей инструментов и контекстными меню.

Диалоговое окно Опции

Основные настройки программы FineReader сосредоточены в диалоговом окне **Опции** (рис. 3.12). При выполнении разнообразных операций может потребоваться изменение настроек программы, поэтому вызвать данный диалог можно различными способами:

- в строке меню выберите команду **Сервис ➢ Опции**;
- щелкните правой кнопкой мыши на любой панели инструментов и выберите в контекстном меню команду **Опции**;
- нажмите кнопку **Опции** на главной панели инструментов;
- нажмите сочетание клавиш **Ctrl+Shift+O**;
- в диалоге открытия изображений или сохранения документов нажмите кнопку **Опции**.

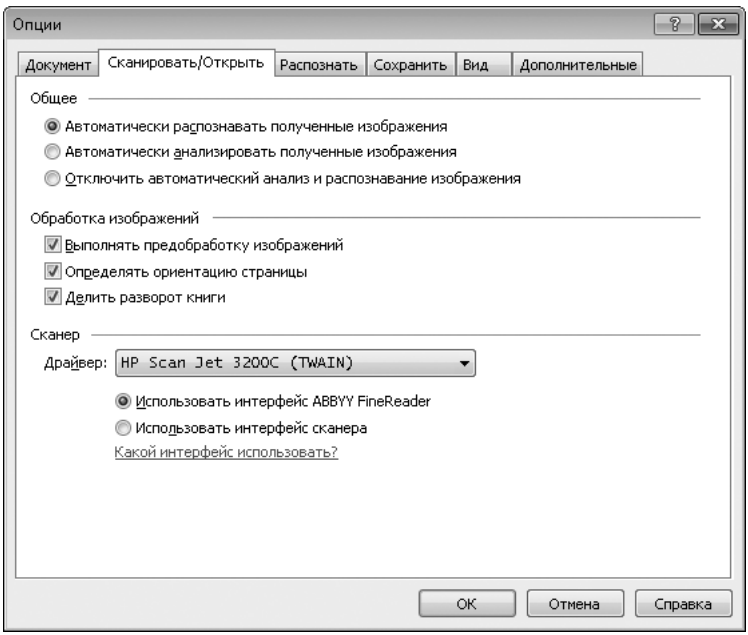


Рис. 3.12 ▾ Диалог **Опции**, вкладка **Сканировать/Открыть**

Диалоговое окно **Опции** включает в себя шесть вкладок. Основную часть настроек мы разберем, рассматривая те этапы обработки документа, на которые эти настройки непосредственно влияют. Пока же остановимся лишь на опциях, которые связаны с работой программы в целом.

На вкладке **Сканировать/Открыть** (рис. 3.12) расположены переключатель и группа флажков, определяющих режим автоматической обработки и распознавания полученных изображений.

- По умолчанию переключатель установлен в положение **Автоматически распознавать полученные изображения**. В таком случае, как только программа получила изображение со сканера или из графического файла, сразу запускается полный процесс анализа, а следом – и распознавания. На современном производительном компьютере это оптимальный вариант.
- Установите переключатель в положение **Автоматически анализировать полученные изображения**. В таком случае программа будет сразу анализировать изображения по мере их получения, но с дальнейшей обработкой повременит. Распознавание документа начнется только по вашей команде. Этот режим хорош при распознавании с обучением или настройкой эталонов и языков.
- Когда переключатель установлен в положение **Отключить автоматический анализ и распознавание изображения**, программа только получает изображения, а процессы анализа (разбивку на области) и распознавания вы должны будете запускать вручную. Как уже сказано, на менее мощных компьютерах с небольшим объемом памяти распознавание идет довольно долго, система при этом порой «задумывается». На таком компьютере при работе с многостраничными документами лучше поставить переключатель в положение **Отключить автоматический анализ и распознавание изображения**. Сначала вы спокойно отсканируете все страницы, а затем запустите обработку и распознавание.
- Назначение трех флажков в группе **Обработка изображений** мы рассмотрели при обсуждении диалогов сканирования и открытия файла. От того, установлены или сняты флажки в диалоге **Опции**, зависит, будут ли эти флажки по умолчанию установлены или сняты в диалогах сканирования и открытия файла.

На вкладке **Вид** (рис. 3.13) содержатся дополнительные настройки внешнего вида окон. Часть из них дублирует настройки, доступные в контекстных меню окон.

При работе с большинством оригиналов менять эти настройки нет никакой необходимости. Однако если на оригинале текст изображен на цветном фоне, а вы решили сканировать этот оригинал в цвете, в окнах некоторые рамки и метки могут быть плохо различимы на таком фоне или сливаться с ним.

В группе **Объекты** задаются цвета, которыми выделяются в рабочих окнах различные объекты. Например, по умолчанию неуверенно распознанные символы помечаются в окне **Текст** голубым цветом. Чтобы изменить цвет такого

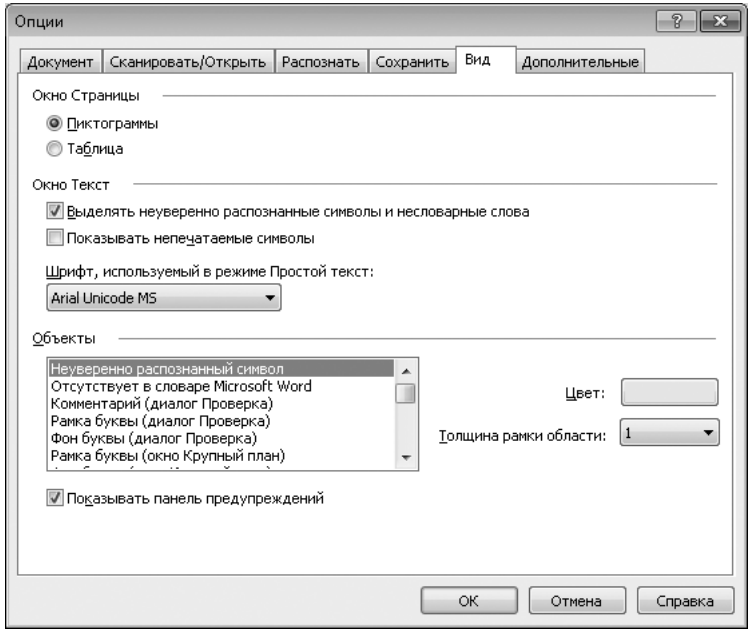


Рис. 3.13 ▾ Диалог **Опции** – вкладка **Вид**

выделения, в списке **Объекты** выберите объект **Неуверенно распознанный символ**, а затем щелкните кнопкой мыши на поле **Цвет**. Откроется палитра, в которой вы можете выбрать любой цвет. Теперь неуверенно распознанные символы будут выделяться этим цветом. На практике заданные по умолчанию цвета хорошо различимы, а главное, все привыкают к стандартным цветам меток. Поэтому менять эти цвета без особой необходимости не стоит.

На вкладке **Дополнительные** (рис. 3.14) пока отметим лишь три функции. Остальные мы рассмотрим, когда речь пойдет о распознавании и проверке документа.

ПРИМЕЧАНИЕ

Программа FineReader позволяет менять язык интерфейса прямо в ходе работы. Для этого на вкладке ***Дополнительно*** *существует кнопка* ***Язык интерфейса****, отображающая при нажатии список доступных языков.*

По умолчанию программа FineReader запускается с новым пустым документом, а на экран сразу выводится окно запуска встроенных сценариев. Чаще всего, при работе с небольшими документами, это оптимальный вариант.

Другое дело – создание электронных копий целых книг. Одно только сканирование нескольких сотен страниц займет не один час, а вместе с распознаванием, проверкой и тщательной вычиткой – еще больше. Разумеется, «в один присест» делать это тяжело. Скорее всего, такое занятие вы распределите на

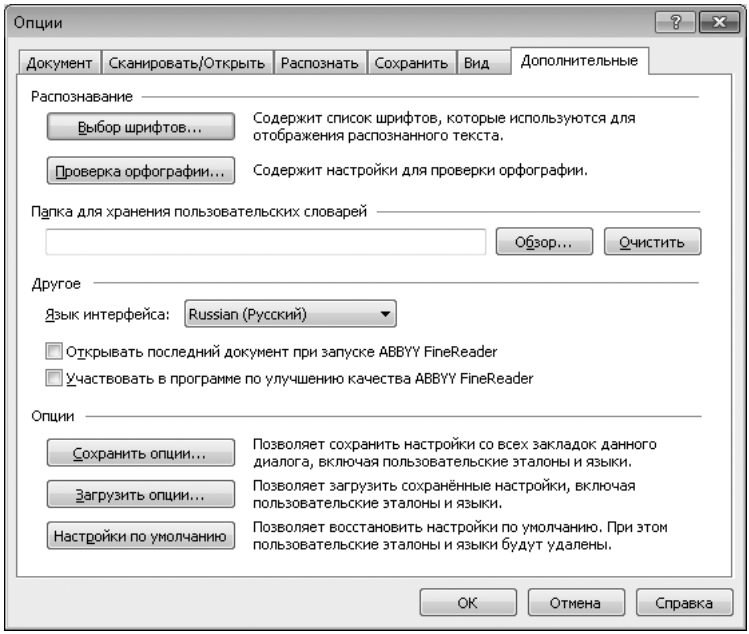


Рис. 3.14 ▾ Диалог **Опции** – вкладка **Дополнительные**

несколько дней. В таком случае установите флажок **Открывать последний документ при запуске ABBYY FineReader**. При следующем запуске программа сразу откроет последний сохраненный документ, и вы продолжите работу с того места, на котором остановились в прошлый раз.

Как мы вскоре увидим, для распознавания некоторых документов лучше подходят особые настройки. Если вам регулярно попадаются примерно однотипные оригиналы, например медицинская литература, радиотехнические справочники или старые иллюстрированные журналы, разумно однажды подобрать лучшие настройки для обработки оригиналов каждого рода и сохранить эти «заготовки» для повторного использования. В дальнейшем мы рассмотрим выбор оптимальных настроек для разных случаев.

Забежим вперед и предположим, что вы подобрали и уже опробовали на практике сочетание настроек, хорошо подходящее для распознавания справочников по электронике (в них много мелких таблиц, черно-белых схем, встречаются формулы, а ошибки в написании обозначений деталей крайне нежелательны). Эти настройки вы выполнили на разных вкладках диалога **Опции**.

1. Чтобы сохранить весь комплекс текущих настроек программы, включая пользовательские языки и эталоны, в диалоговом окне **Опции** на вкладке **Дополнительные** нажмите кнопку **Сохранить опции...**. Откроется стандартный диалог сохранения файла.
2. Выберите в нем папку, в которую надо сохранить файл настроек, например **Страницы**, и укажите имя файла, например **Для радиосправочни-**

ков. К имени файла будет автоматически добавлено расширение **.FBT** (тип файла – набор опций документа FineReader).

3. Нажмите кнопку **Сохранить**.

Таким образом, вы сохранили на диске файл с конкретными настройками программы FineReader. Настроив затем программу для другого случая распознавания, например для обработки иллюстрированных журналов (часть текста на цветном фоне, много иллюстраций, встречаются декоративные шрифты и т. д.), сохраните и этот набор настроек. Назовите его, допустим, **Для журналов**. У вас появилось уже два файла – два разных набора опций.

1. Когда вы соберетесь сканировать очередной похожий справочник, перед началом работы откройте диалоговое окно **Опции**, перейдите на вкладку **Дополнительные** и нажмите кнопку **Загрузить опции...**.
2. Появится предупреждение о том, что эта операция заменит текущий набор настроек программы. Нажмите кнопку **Да**.
3. Откроется стандартный диалог открытия файла. Выберите в нем тот файл, в который были сохранены настройки для радиосправочников. Нажмите кнопку **Открыть**. Сохраненный набор настроек будет загружен в программу.
4. Закройте диалоговое окно **Опции**, нажав в нем кнопку **ОК**. Окно закроется, а загруженные настройки будут применены.
5. Перед сканированием и обработкой оригинала типа глянцевого журнала таким же образом загрузите настройки из файла **Для журналов.fbt**.

Чтобы вернуть в любое время настройки программы FineReader к принятым по умолчанию, откройте диалоговое окно **Опции**, перейдите на вкладку **Дополнительные** и нажмите кнопку **Настройки по умолчанию...**. В результате все настройки сбрасываются к исходным значениям.

Документ FineReader

Документ ABBYY FineReader – это объект, который создается программой ABBYY FineReader для работы с одним входным документом с учетом его целостной структуры. В документе FineReader хранятся исходные изображения страниц, соответствующий им распознанный текст, настройки программы (опции сканирования, распознавания, сохранения, данные об оформлении текста (шрифтах, форматировании, расположении в тексте таблиц и рисунков), а также созданные в процессе работы пользовательские эталоны, языки и группы языков).

На диске такой документ может занимать значительное место. Например, документ, отсканированный с разрешением 300 DPI с нашей книги, будет занимать на диске около 3 Гб.

По умолчанию при запуске программы создается новый документ – **Документ без имени**. Пока вы его не сохранили, он находится в системной папке, предназначенной для хранения временных файлов.

Чаще всего работа с документом в программе FineReader заканчивается тем, что вы сохраняете результат распознавания в виде документа Word или

PDF, либо передаете результат в одно из приложений Microsoft Office или Adobe Reader (Adobe Acrobat) для дальнейшего редактирования. После этого вы закрываете программу FineReader, ведь конечный результат – документ Word, файл PDF или другого формата – уже получен.

При закрытии программа FineReader запрашивает, нужно ли сохранить изменения в *документе FineReader*. Если вы ответите утвердительно, откроется диалоговое окно **Сохранить документ FineReader** (рис. 3.15).

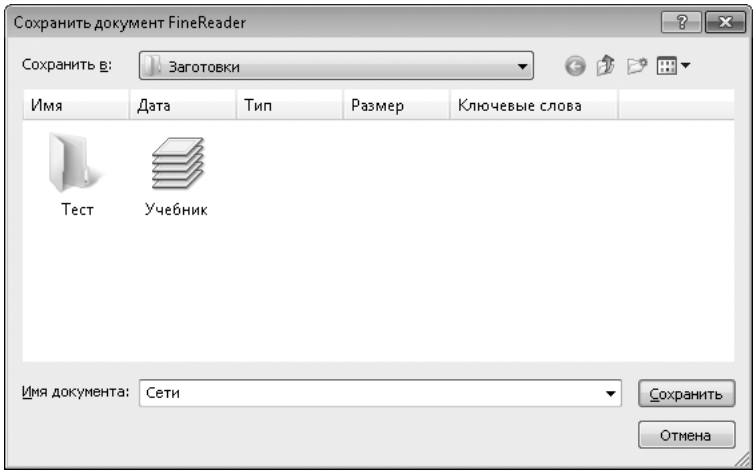


Рис. 3.15 ▾ Диалог сохранения документа FineReader

Выберите нужную папку в раскрывающемся списке **Сохранить в:** и задайте название сохраняемого документа в поле **Имя документа:**. По умолчанию предлагается сохранить документ FineReader в библиотеку Документы, но вы можете выбрать и любую другую папку. Нажмите кнопку **Сохранить**, и документ с заданным именем будет сохранен в указанном месте.

Если при выходе из программы вы не стали сохранять документ FineReader, все отсканированные изображения, а также эталоны и пользовательские словари, созданные при работе с этим документом, утрачиваются. На практике документы FineReader стоит сохранять в случаях, если вы намерены продолжить работу с ними в дальнейшем. Например, вы сканируете и распознаете книгу большого объема и хотите сделать это в несколько приемов.

1. Отсканируйте несколько страниц книги. Просмотрев результат распознавания в окне **Текст**, при необходимости измените настройки программы.
2. Продолжите сканирование, пока позволяет время.
3. Сохраните документ FineReader: меню **Файл ➢ Сохранить документ FineReader**.
4. Закройте программу: меню **Файл ➢ Выход**.

5. В следующий раз запустите программу и откройте ранее сохраненный документ: меню **Файл ➢ Открыть документ FineReader**.
   Программа «запоминает» несколько последних документов: их имена отображаются в меню **Файл**. Поэтому для открытия одного из последних сохраненных документов просто выберите его имя в меню **Файл**.
6. Продолжите работу – отсканируйте следующие страницы, откорректируйте часть распознанного текста и т. д. В однажды сохраненный документ все изменения и дополнения вносятся автоматически – этим работа с программой FineReader отличается от других приложений, где изменения в документе надо всякий раз сохранять вручную. Закройте программу.
7. В следующий раз запустите программу, откройте сохраненный документ – и так до тех пор, пока вы не распознаете и не проверите книгу или документ полностью.
8. Когда вся работа с распознанным текстом будет закончена, передайте результат распознавания в программу Microsoft Word либо сохраните его как документ Word или PDF. После этого закройте программу FineReader и удалите документ FineReader с диска.

Когда программа FineReader корректно завершает свою работу, она удаляет временные документы из временной папки. При следующем запуске приложения по умолчанию создается новый пустой документ.

Если же работа программы завершилась аварийно, например компьютер завис, или внезапно отключили электричество, документ во временной папке остается на диске. При следующем запуске программы появится запрос: **Вы хотите восстановить документ без имени?** Ответьте утвердительно, и документ будет восстановлен. Таким образом, вам удастся довести до конца случайно прерванную работу.

Резюме

Этапы обработки документа в программе FineReader отображаются в ее рабочих окнах. Расположение и вид окон настраиваются с помощью меню **Вид**, контекстных меню, вызываемых щелчком правой кнопки мыши внутри рабочих окон, а также кнопок на панели инструментов **Быстрый доступ**.

Документ FineReader – особая папка с файлами, в которых хранятся исходные изображения всех страниц документа, распознанный текст с учетом оформления, пользовательские настройки. Для сохранения на диске документа FineReader выберите в меню команду **Файл ➢ Сохранить документ FineReader**. Чтобы впоследствии открыть этот документ и продолжить работу с ним, выберите в меню команду **Файл ➢ Открыть документ FineReader**.

В отличие от документа FineReader, файл с сохраненными настройками программы не содержит ни изображений, ни распознанного текста – только сами настройки и пути к пользовательским словарям и эталонам. Чтобы сохранить настойки программы, откройте диалоговое окно **Опции** и на вкладке **Дополнительные** нажмите кнопку **Сохранить опции...**. Если вы захотите исполь-

зовать эти настройки при работе с очередным документом, вновь откройте диалоговое окно **Опции** и на вкладке **Дополнительные** нажмите кнопку **Загрузить опции...**.

Таким образом, мы рассмотрели общие принципы работы с программой. Теперь приступим к подробному описанию каждого этапа работы с документом, включая практические приемы и конкретные настройки программы.

Глава 4 Получение изображений

Первое, что надо сделать для создания редактируемой копии бумажного документа, – получить его изображение в электронном виде. Еще пять–семь лет назад основным и практически единственным устройством для получения изображений бумажных документов был сканер. Стремительный прогресс цифровых камер сделал фотографирование бумажных оригиналов реальной альтернативой их сканированию. Современный фотоаппарат начального или среднего уровня позволяет получать изображения с качеством, вполне пригодным для дальнейшего распознавания, а стоит он сегодня не дороже сканера среднего класса ($100–200).

Когда у вас уже есть сканер или цифровая камера – вопрос выбора отпадает сам собой. Что есть, тем и пользуйтесь! Видимо, при наличии камеры с матрицей более 2 Мпикс для работы с документами и программой FineReader специально покупать сканер не стоит. Если есть сканер, камеру стоит покупать не ради съемки книг, а только из других соображений – именно как фотоаппарат.

Если же пока нет ни того, ни другого, для личного пользования более оправданной тратой денег может оказаться покупка камеры средней ценовой категории. По крайней мере, камера заменит сканер практически во всех случаях, а вот у сканера область применения довольно узкая. Главное достоинство камеры – ее мобильность: компактный аппарат удобно носить с собой и использовать в любом месте.

Сканер как отдельное устройство, или объединенный с принтером, может оказаться удобнее там, где необходимо регулярно и быстро получать изображения большого количества документов. Камеру, после того как сделаны снимки, нужно подключать к компьютеру и переносить файлы с карты памяти на диск компьютера, затем удалять снимки и т. д. Эти процедуры занимают некоторое время, тогда как сканер постоянно подключен к компьютеру и всегда готов к работе.

Работа со сканером

В основе любого сканера лежат светочувствительный элемент (сенсор) и обслуживающая его оптико-механическая система. Самыми распространенными являются планшетные сканеры, в которых оригинал кладется на стекло (планшет), а под ним перемещается оптическая система. В зависимости от типа сенсора планшетные сканеры делятся на две большие группы.

CCD (ПЗС, прибор с зарядовой связью) – светочувствительная матрица, которая жестко закреплена в корпусе сканера. Вдоль оригинала перемещается каретка, которая несет лампу подсветки, а также систему линз и зеркал. Через эту оптическую систему изображение передается на сенсор CCD. Отличительная внешняя черта сканеров CCD – большая толщина корпуса, более 5 см. «Начинка» сканера довольно сложна, в ней присутствуют точная механика и столь же прецизионная оптика. Цена типичного сканера CCD составляет от $150 до $600.

CIS (Contact Image Sensor) – линейка, состоящая из светочувствительных ячеек. Такой сенсор вместе с линейкой светодиодов подсветки крепится на подвижную каретку и в процессе сканирования перемещается вдоль оригинала. По конструкции сканер с CIS гораздо проще сканера с матрицей CCD. Толщина корпуса может составлять всего 2–3 см. Стоят такие сканеры, как правило, не дороже $100. Однако по ряду параметров сканеры CIS заметно уступают устройствам на основе CCD.

Сканеры как самостоятельные устройства постепенно исчезают с рынка. Точно так же практически прекращен выпуск аналоговых копировальных аппаратов, в просторечии «ксероксов». На смену им приходят многофункциональные устройства (МФУ). МФУ – принтер, «на спину» которому установлен сканер. И принтер, и сканер МФУ определяются компьютером как отдельные устройства, хотя для подключения они используют общий кабель USB. Существенно, что копирование «с бумаги на бумагу» может происходить вообще без участия компьютера: на корпусе МФУ обязательно присутствует кнопка, запускающая процесс прямого копирования со сканера на принтер.

Как правило, в МФУ устанавливают самые простые сканеры на основе CIS. В результате появляются лазерные «комбайны» ценой до $200, а некоторые струйные агрегаты стоят дешевле $100. Поэтому правильнее сказать, что из продажи пропали не все, а лишь дешевые сканеры CIS – теперь все производители встраивают их в многофункциональные устройства.

Параметры сканеров

Самый известный параметр сканера – его **разрешающая способность**, или разрешение. Как уже сказано, это количество точек изображения, получаемого сканером с единицы размера оригинала (с 1 дюйма), и измеряется оно в точках на дюйм – DPI. Чем выше разрешение, тем более мелкие детали могут быть перенесены с оригинала на его изображение. Однако в нашем конкретном случае – при сканировании текста для распознавания – этот параметр не является реша-

ющим в выборе сканера! Даже самый простой и дешевый сканер дает разрешение никак не менее 600 DPI, в то время как для сканирования обычных печатных оригиналов достаточно 300 DPI. Стандартный книжный шрифт обычно сканируют с разрешением 150 DPI, и лишь для распознавания самый мелких шрифтов может понадобиться разрешение 600 точек на дюйм. Поэтому высокое разрешение сканера (а у некоторых моделей оно доходит до 6400 DPI) пригодится только для других задач, например сканирования фотографий или художественных изображений.

Глубина резкости – характеристика, хорошо знакомая любому фотографу. Смысл ее в том, что любой оптический прибор получает идеально резкое изображение только тех объектов, которые удалены от него на строго определенное расстояние. Другими словами, он фокусируется или наводится на определенную дальность. Все объекты, которые находятся ближе или дальше точки фокусировки, получаются на снимке более или менее размытыми. Однако в некотором диапазоне расстояний, зависящем от конструкции и настройки оптической системы, размытие изображения не превышает определенного предела и вполне удовлетворяет практическим требованиям. Этот диапазон и принято называть глубиной резкости (рис. 4.1).

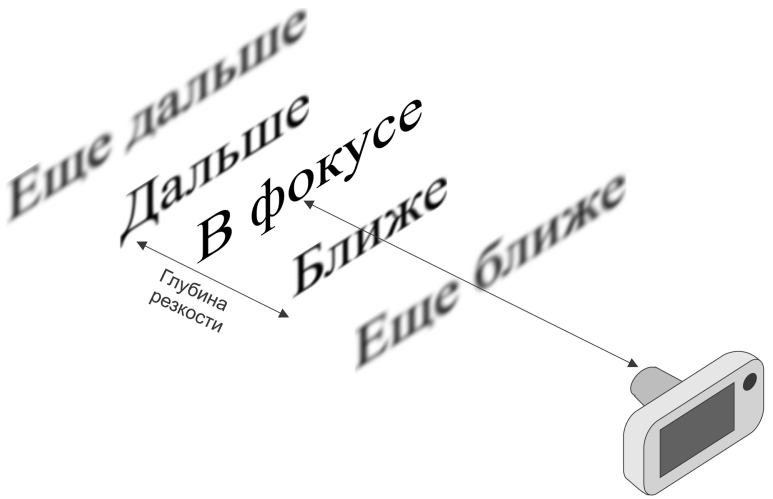


Рис. 4.1 ▾ Точка фокусировки и глубина резкости

Какое изображение можно считать «достаточно резким», а какое уже является размытым, описано математическими формулами – по ним рассчитывают глубину резкости для разных фотоаппаратов и другой подобной техники. Мы же ограничимся чисто практическими соображениями.

У фотокамер глубина резкости варьируется в пределах от десятков сантиметров до нескольких метров. Она зависит от конструкции камеры (главного фокусного расстояния объектива) и ее установок (фокусировки на определен-

ное расстояние, панорамирования, диафрагмы). Глубина резкости начинает сказываться лишь при фотографировании действительно объемных объектов. При съемке почти плоских оригиналов, какими являются отдельные листы бумаги или книги, достаточно сфокусировать камеру на центр страницы – все изображение получится четким. Для уменьшения краевых искажений книгу следует располагать на определенном расстоянии от объектива. По возможности разгладить книжный разворот, снимать горизонтально, а не под углом. При этом чем больше расстояние от камеры до снимаемого объекта, тем меньше искажения и потеря резкости на краях снимка.

Оптика любого сканера постоянно сфокусирована на поверхности стекла планшета. Для сканеров CCD глубина резкости колеблется в пределах от нескольких сантиметров до почти десяти сантиметров. Поэтому с помощью сканера CCD, в принципе, удается получить удовлетворительное изображение объемных предметов, находящихся в нескольких сантиметрах над стеклом. Шутки ради можно даже откинуть крышку сканера, уткнуться носом в стекло и получить некое подобие автопортрета!

Сканеры на основе CIS обладают гораздо меньшей глубиной резкости – всего несколько миллиметров. Когда оригинал, например отдельный лист бумаги, плотно прилегает к стеклу по всей площади, его изображение получится резким и на том, и на другом сканере. Глубина резкости в этом случае не играет заметной роли.

Другое дело, если оригинал местами отстоит от стекла на некоторое расстояние. При сканировании толстой книги в жестком переплете весь разворот практически невозможно полностью и равномерно прижать к стеклу. Участок вблизи корешка возвышается над планшетом как минимум на несколько миллиметров. На сканере с большой глубиной резкости (CCD) это почти не влияет на качество изображения, а вот многие сканеры CIS заметно «размазывают» картинку там, где бумага неплотно прилегала к стеклу (рис. 4.2).

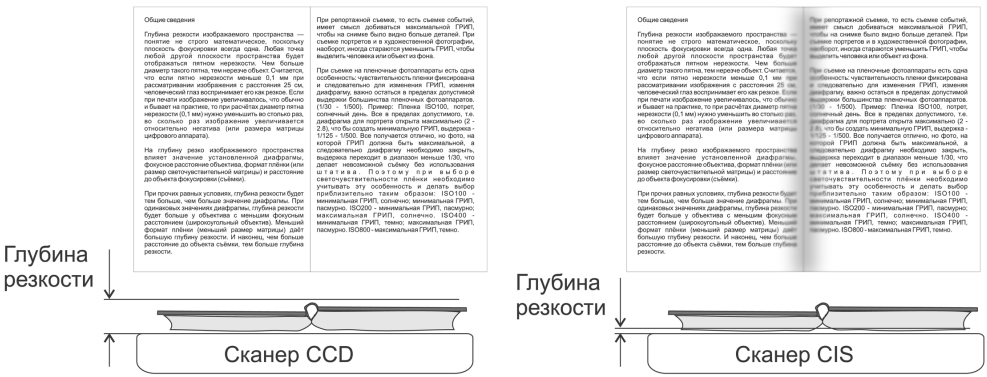


Рис. 4.2 ▾ Влияние глубины резкости на сканирование книжного разворота

Поэтому считается, что более дешевые сканеры с датчиком CIS – «офисный вариант», ведь в офисе чаще всего сканируют отдельные листы, самое большее – тонкие брошюры. Для сканирования книг лучше подходят «полупрофессиональные» и «профессиональные» сканеры с датчиками CCD. Они обеспечивают большую глубину резкости и хорошо справляются с оригиналами, которые нельзя плотно прижать к планшету.

Время сканирования зависит в первую очередь от модели сканера. На скорость получения изображения страницы одним и тем же сканером прямо влияют установленные разрешение и цветность. Чем выше разрешение, тем дольше будет сканироваться оригинал. Черно-белое изображение получается быстрее, чем изображение в градациях серого, а последнее – быстрее, чем цветное. Обычно в документации к сканеру приводится несколько значений – время сканирования страницы с разными разрешением и цветностью. Часто производители особо выделяют такой параметр, как время сканирования для предварительного просмотра: в этом случае используется черно-белый режим с наименьшим разрешением.

Применительно к работе с программой FineReader показательным является режим «Сканирование документа в оттенках серого, разрешение 300 DPI». Сканеры CCD начальной ценовой категории сканируют одну страницу почти минуту, а наиболее совершенные модели выполняют такую операцию в четыре-пять раз быстрее. Сканеры на основе CIS, например входящие в состав МФУ, обладают достаточным быстродействием: сканирование страницы формата A4 в оттенках серого с разрешением 300 DPI занимает примерно 10 с.

Быстродействие сканера особенно ощущается при работе с многостраничными документами. Чтобы открыть крышку, перевернуть лист, закрыть крышку и нажать кнопку, нужно 3–5 секунд. На «быстром» сканере при известной сноровке вы «прогоните» 100 страниц за полчаса, а на «медленном» эта же работа займет час, а то и больше.

Автоподатчик – приспособление, которым оборудуются некоторые большие офисные МФУ. Пачка оригиналов закладывается в лоток, как чистая бумага в принтер. Автоподатчик по очереди затягивает листы на стекло сканера, а после сканирования сбрасывает их во второй лоток. Реальную пользу автоподатчик приносит, когда нужно подряд сканировать много листов одинакового размера. Например, переводятся в электронный вид хорошо сохранившиеся архивные документы на стандартных машинописных листах, вынутых из скоросшивателя. При работе со сброшюрованными страницами, так же как и с ветхими или помятыми оригиналами, податчик не поможет. Подобные документы все равно придется укладывать на стекло вручную.

Ресурс сканера – число сканирований, которое устройство гарантированно может выполнить в течение своей «жизни». Механика постепенно изнашивается, и рано или поздно подшипники, зубчатые ремни и шестерни выходят из строя. По данным производителей, ресурс моделей младшей ценовой категории составляет порядка 10 000 копий, для более дорогих моделей он может доходить до 50 000 копий и более. При обычном «домашнем» использовании ска-

нер, как правило, успевает устареть морально еще задолго до физического износа. Однако при регулярном сканировании книг оказывается, что 10 000 страниц – не так уж и много! Это всего лишь около 50 книг средней толщины.

Таким образом, для обычной офисной работы хорошо подходит любой сканер, в том числе и встроенный в многофункциональное устройство. Тем же, кто собирается регулярно делать электронные копии книг, особенно если это толстые тома в жестком переплете, стоит обратить внимание на более дорогие модели с датчиком CCD.

Драйвер и настройки сканера

Любой сканер позволяет подстраивать яркость и контрастность получаемого изображения, выбирать разрешение и цветность. Такая регулировка производится через драйвер – специальную программу, посредством которой сканер взаимодействует с компьютером и другими программами. Драйвер разрабатывается производителем для каждой конкретной модели сканера, поставляется на входящем в комплект компакт-диске и устанавливается при подключении сканера к компьютеру. Как правило, вместе с драйвером устанавливается и фирменная утилита для настройки сканера и управления им. Фактически такая утилита является одним из компонентов драйвера.

Каждый производитель подходит к интерфейсу управления сканером по-своему. Как правило, в окне этой утилиты присутствуют область предварительного просмотра страницы, регуляторы яркости, контрастности, насыщенности, кнопки или переключатели для выбора разрешения и цветности, а также элементы управления дополнительными настройками. Иногда в окне, наоборот, выводятся лишь несколько кнопок для выбора предустановленных режимов с названиями наподобие «Фото», «Среднее качество» и «Черновое качество», а все детальные настройки спрятаны на дополнительных вкладках и диалогах. Подробное описание настроек содержится в документации к сканеру.

Программа FineReader умеет взаимодействовать со сканером двумя способами:

- через фирменный интерфейс драйвера конкретного сканера. В этом случае перед началом сканирования вызывается диалоговое окно программы настройки и управления сканером от производителя;
- через собственный интерфейс управления настройками сканера, который использует встроенную в операционную систему технологию TWAIN. С помощью этой технологии компьютер способен взаимодействовать с различными сканерами и другим устройствами получения изображений, соответствующими требованиям спецификации TWAIN. Функции выбора разрешения, цветности, регулирования контрастности и яркости являются в ней стандартными. Благодаря этому встроенный интерфейс программы FineReader работает с большинством моделей сканеров, как довольно старых, так и самых современных. Полный список поддерживаемых моделей приводится на сайте компании ABBYY (**http://www.abbyy.ru/support/TWAINscanners**).

По умолчанию программа FineReader использует свой встроенный интерфейс управления сканером. Если к компьютеру подключены несколько TWAIN-совместимых устройств, например сканер и веб-камера, либо при попытке обратиться к вашему сканеру из программы FineReader появляется сообщение об ошибке, следует уточнить настройки программы.

Сделать это нужно всего один раз – в дальнейшем программа каждый раз будет использовать указанное устройство и обращаться к нему через выбранный интерфейс. Для выбора интерфейса управления сканером откройте диалоговое окно **Опции** (меню **Сервис ➢ Опции**) и перейдите на вкладку **Сканировать/Открыть** (рис. 3.12).

1. В группе **Сканер** в раскрывающемся списке **Драйвер:** выберите драйвер сканера, с которым вы будете работать. В списке, в зависимости от конфигурации компьютера и модели сканера, могут присутствовать несколько записей. Если это так, возможно, нужную придется выбрать экспериментальным путем.
2. Убедитесь, что переключатель стоит в положении **Использовать интерфейс ABBYY FineReader**. Такова рекомендованная настройка, и если программа будет правильно работать с такими настройками, то переключатель следует оставить в этом положении.
3. Нажмите кнопку **OK**. Диалоговое окно **Опции** закроется, а программа будет использовать выбранные драйвер и интерфейс управления сканером.

Теперь можно проверить, что выбраны правильные настройки. Для этого отсканируйте какую-либо страницу.

На главной панели инструментов нажмите кнопку **Сканировать**. Другие способы – выберите ссылку **Сканировать** на закладке **Другие** окна **Новое задание**, или команду меню **Файл ➢ Сканировать страницы**, или нажмите сочетание клавиш **Ctrl+K** на клавиатуре. В результате должно открыться диалоговое окно сканирования документа (рис. 2.2).

Если вместо этого диалога появляется сообщение об ошибке, закройте сообщение, нажав в нем кнопку **OK**. Вновь откройте диалоговое окно **Опции** и на вкладке **Сканировать/Открыть** (рис. 3.12) выберите другой драйвер сканера.

Если и с другим драйвером программа FineReader не смогла обратиться к вашему сканеру, видимо, модель сканера и его драйвер не совместимы с программой. В таком случае откройте диалоговое окно **Опции** и на вкладке **Сканировать/Открыть** установите переключатель в положение **Использовать интерфейс сканера**. При такой настройке при нажатии кнопки **Сканировать** вместо диалогового окна сканирования программы FineReader будет открываться окно драйвера конкретной модели сканера. То, какие настройки и действия доступны в этом окне, уточните в документации к сканеру.

Процедура сканирования уже описана в главе «Быстрый старт». Если оригинал хорошего качества, лучше оставить все настройки в диалоге сканирования по умолчанию. Чтобы в любое время выставить такие настройки, нажмите кнопку **По умолчанию**. При выборе собственных настроек сканирования ориентируйтесь на качество оригинала и на то, что вы хотите получить в конечном

счете. Возможно, первый результат вас не устроит – тогда и нужно прибегать к изменению настроек сканирования.

1. Отсканируйте страницу. В окне **Страницы** выберите **Вид со свойствами** (рис. 3.3). Посмотрите на процент неуверенно распознанных символов и сообщения об ошибках. Если доля неуверенно распознанных символов не превышает 3–5%, все нормально. Если качество распознавания хуже – попробуйте выяснить причину и изменить настройки.
2. Просмотрите разные участки изображения в рабочем окне **Крупный план**. Обратите внимание на несколько признаков, по которым можно судить о хорошем качестве изображения текста. Разумеется, изображение не может быть намного лучше, чем оригинал. Если сам оригинал низкого качества, на нем присутствуют пятна, размазанный или слабо пропечатанный шрифт, исправить такие дефекты с помощью настройки сканера достаточно сложно, но иногда возможно. Например, если изображение слишком темное, его можно осветлить.
   - Все буквы отделены друг от друга, не «слипаются» между собой. Чрезмерная «жирность» бывает следствием недостаточной яркости изображения при сканировании. Кроме того, при недостаточной яркости на изображении часто появляется «мусор» в виде мелких крапинок и штрихов.
   - В контурах букв отсутствуют разрывы. Такие разрывы могут появиться, если при сканировании была выставлена избыточная яркость изображения.
   - В буквах и цифрах четко различимы засечки и «хвостики».
   - Хорошо различимы точки и запятые.

Когда вы видите, что изображение не соответствует этим требованиям, попробуйте изменить настройки сканирования и вновь отсканировать страницу. Если при распознавании нового изображения того же оригинала число ошибок уменьшилось – вы на правильном пути. В противном случае испробуйте другие сочетания разрешения, режима сканирования и яркости, пока не получите лучший результат.

Разрешение

Для выбора разрешения выберите нужное значение в раскрывающемся списке **Разрешение**. Программа FineReader предлагает три стандартных значения.

Рекомендованное разрешение при сканировании обычного книжного текста – 300 DPI. Его можно считать стандартной и универсальной настройкой. Для работы с документом, отпечатанным достаточно крупным шрифтом на мелованной бумаге, приемлемым может оказаться и меньшее разрешение – 250 DPI. Если с уменьшением разрешения процент ошибок распознавания остается прежним, целесообразно остановиться именно на меньшем разрешении – сканироваться документ будет быстрее, а на диске понадобится меньше места для временных файлов.

Оригиналы с мелким шрифтом попробуйте сканировать при разрешении 600 DPI. Такое разрешение может помочь и при распознавании плохо пропечатанных документов, например после ксерокопирования.

Режим сканирования

Цветной режим уместен лишь тогда, когда вам важно получить цветные иллюстрации в выходном документе. Если же таких иллюстраций в оригинале нет, а вы лишь хотите сохранить какое-то цветовое оформление текста, цветной режим сканирования – не лучшее решение. В цвете сканирование происходит дольше, документ FineReader занимает на диске больше места. В то же время «раскрасить» распознанный документ при последующей обработке в окне **Текст** или программе Microsoft Word – буквально секундное дело.

Сканирование в оттенках серого считается предпочтительным режимом при работе в программе FineReader 10. С изображениями в оттенках серого работает технология адаптивного распознавания. При этом яркость и контрастность автоматически подстраиваются в процессе обработки изображения, и программе удается распознавать даже не совсем четкие знаки. Кроме того, режим «оттенки серого» позволяет сохранять в выходном документе полутоновые иллюстрации, а во многих случаях это бывает кстати.

Черно-белый режим сканирования выгоден в первую очередь своей скоростью. В черно-белом режиме любая точка изображения, которая светлее заданного порога, расценивается как «белая», а которая темнее – как «черная». Порог же задается регулятором **Яркость**.

В черно-белом режиме целесообразно сканировать оригиналы хорошего качества, не содержащие иллюстраций. Пример такого оригинала – обычный офисный документ, напечатанный на лазерном принтере. Черно-белый режим можно выбрать и тогда, когда из оригинала вам требуется взять только текст, а иллюстрации не нужны.

Однако здесь подстерегает маленькая «ловушка». Если отсканировать в черно-белом режиме документ, содержащий иллюстрации, то от иллюстраций останутся только самые темные фрагменты, в том числе мелкие. В ходе распознавания изображения программа FineReader может принять отдельные части иллюстраций за какие-то символы, и в распознанном тексте появятся совершенно бессмысленные «вставки». Поэтому документы, содержащие иллюстрации, лучше все-таки сканировать в оттенках серого: удалить из распознанного текста целые рисунки проще, чем выискивать в нем «обрывки» таких рисунков, ошибочно распознанные как текст.

Яркость

Ручная регулировка яркости изображения действует при любом режиме сканирования. По умолчанию принята 50% настройка – ползунок стоит посередине полосы регулировки. Как правило, при таком значении яркости изображение типографского шрифта на обычной белой бумаге получается оптимальным. Подстройка чаще всего требуется при сканировании менее контрастных оригиналов. Примеры таких оригиналов легко найти среди изданий, вышедших в «перестроечные годы», – бледная типографская краска на низкокачественной сероватой или зеленоватой бумаге.

При подборе яркости сканирования следует стремиться к результату, показанному на рис. 4.3 на среднем образце. В верхней строке – изображение, полученное с избыточной яркостью: буквы истончены, и их контуры местами разорваны. Нижний образец отсканирован с недостаточной яркостью: некоторые буквы сливаются между собой.

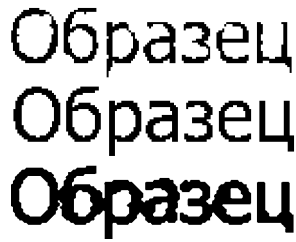


Рис. 4.3 ▾ Изображение, отсканированное с разной яркостью

Оценивать качество изображения удобнее всего в рабочем окне **Крупный план**, установив в нем масштаб 300% или 600%. Когда установленная вручную яркость сканирования значительно отличается от оптимальной, при распознавании появляется сообщение о необходимости изменить яркость сканирования.

Параметры страницы

В группе **Параметры страницы** настраивается размер и положение области, изображение которой будет получать сканер. В раскрывающемся списке **Размер бумаги:** доступны три варианта.

- **Размер области сканирования**. По умолчанию предлагается этот вариант. Сканер будет получать изображение со всей площади планшета. У большинства распространенных моделей размер планшета чуть больше стандартного листа бумаги формата A4 (297×210 мм). Эта установка хорошо подходит для сканирования и офисных документов, и книжных разворотов.
- **Размер текущего выделения**. Поверх изображения в области предварительного просмотра вы увидите рамку с маркерами по углам. Чтобы задать область сканирования, перетаскивайте мышью края или углы рамки – сканироваться будет только часть оригинала внутри рамки. Ограничивать область сканирования есть смысл, когда по размеру оригинал значительно меньше планшета: зачем сканировать пустые поля вокруг?
- **Настроить**. Выберите этот вариант, и откроется диалог **Пользовательский размер бумаги** (рис. 4.4).

Введите в поля **Высота** и **Ширина** точный размер области сканирования в миллиметрах или дюймах. Нажмите кнопку **OK**. Диалог закроется, а сканер будет получать изображение только из области указанного размера. Учтите, что один угол этого воображаемого прямоугольника всегда находится в том месте планшета, где пересекаются нулевые метки на линейках или на стекле нанесен габаритный «уголок». Обычно это левый угол планшета, обращенный к петлям крышки.

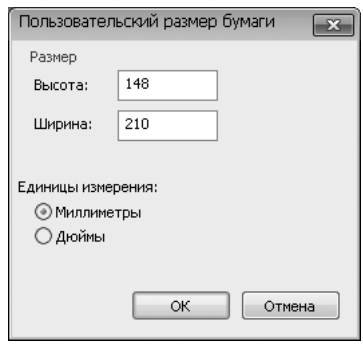


Рис. 4.4 ▾ Задание размера области сканирования

Сканирование многостраничных документов

Два переключателя в группе **Многостраничное сканирование** активны только в том случае, если сканер оборудован автоподатчиком. Чтобы задействовать автоматическую подачу оригиналов, установите флажок **Загружать страницы из автоподатчика бумаги**. Положите пачку оригиналов в лоток автоподатчика и нажмите кнопку **Сканировать** в диалоговом окне программы FineReader. Сканер автоматически обработает всю пачку документов без дополнительных запросов. Когда будет отсканирован последний лист, нажмите в диалоге кнопку **Закрыть** и продолжите работу с программой.

В наиболее совершенных сканерах предусмотрена функция автоматического сканирования двухсторонних оригиналов. При установленном флажке **Двустороннее сканирование** автоподатчик будет переворачивать каждый лист и сканировать его с обеих сторон.

При сканировании большого количества документов подряд движения рук в определенной мере доходят до автоматизма. Чтобы положить в сканер следующую страницу, выровнять ее на стекле и закрыть крышку, требуется всего несколько секунд. Необходимость каждый раз поворачиваться от сканера к мыши и монитору отнимает лишнее время.

Чтобы не нажимать кнопку **Сканировать** после укладки на планшет очередной страницы, установите флажок **Пауза между страницами:** и укажите подходящую задержку между сканированием страниц. По умолчанию предлагается задержка в 10 секунд. В этом случае, отсканировав страницу, сканер автоматически приступит к сканированию следующего оригинала по истечении указанной паузы. Чтобы завершить сканирование последней страницы оригинала, нажмите в диалоговом окне кнопку **Закрыть**.

Работа с цифровой камерой

У цифровой камеры есть одно неоспоримое преимущество перед любым сканером – портативность. Фотоаппарат легко взять с собой и переснять любую книгу или другой документ практически в любом месте. Цифровой аппарат весьма полезен и в стационарных условиях. Во-первых, есть оригиналы, не очень удобные для сканирования из-за габаритов, – возьмем хотя бы энциклопедии или альбомы, подшивки газет. Во-вторых, сфотографировать много страниц подряд иногда быстрее, чем сканировать их. Не случайно профессиональные системы на основе цифровых камер для библиотек и книгохранилищ успешно строили еще тогда, когда любительские цифровые фотоаппараты считались крайне дорогой экзотикой.

Параметры цифровых камер

Рассмотрим некоторые параметры современных камер, относящиеся к получению изображений для последующего распознавания. Те качества, которые обсуждают чаще всего, в основном важны в других ситуациях: съемке пейзажей,

портретов, репортажей. Съемку печатных материалов с целью дальнейшего распознавания текста до недавнего времени почти никто не рассматривал в числе типичных применений «цифровика».

Камеры разделяют на аппараты начального уровня, любительские камеры среднего ценового диапазона и дорогие полупрофессиональные и профессиональные модели. Технических характеристик недорогих камер, выпускаемых с 2004–2005 года, уже достаточно, чтобы снимать бумажные оригиналы для распознавания их программой FineReader. Современные аппараты хорошо подходят для этой цели все без исключения. Чем более «профессиональным» считается аппарат, тем больше в нем заложено тонких ручных настроек. Фотографировать печатные оригиналы, которые лежат на месте и никуда не убегут, удобно с ручными настройками камеры. Однако автоматика тоже позволяет получить прекрасные результаты, и использование для съемки печатных оригиналов более дорогих и совершенных моделей едва ли принесет какие-то дополнительные преимущества по сравнению с камерами начального и среднего уровня.

Требования к снимку, предназначенному для распознавания, несколько отличаются от того, чего обычно ожидают от художественной фотографии. При художественной съемке важны широкий динамический диапазон, точная цветопередача, а на геометрические искажения обращают меньше внимания. При фотографировании печатных оригиналов приоритеты иные:

- кадрирование: края оригинала в идеале должны совпадать с границами кадра;
- минимум геометрических искажений. Строки должны быть ровными и параллельными, а прямые углы – прямыми. Необходимое условие – держать аппарат нужно над центром оригинала параллельно плоскости документа;
- максимальная четкость и высокая контрастность;
- равномерная яркость. На снимке не должно быть теней и бликов;
- баланс белого: белая бумага на снимке должна быть действительно белой. В то же время правильная цветопередача никакой роли не играет.

Фокусное расстояние и трансфокатор. Большинство цифровых аппаратов снабжено объективами с переменным фокусным расстоянием. Иначе эта функция называется Zoom («зум»), или «приближение». Многие начинающие фотографы уверены, что трансфокатор служит исключительно для приближения удаленных объектов, чтобы, не сходя с места, можно было снимать такие объекты «крупным планом». На самом деле, регулируя степень «приближения», фотограф изменяет очень важные характеристики объектива, такие как угол охвата и главное фокусное расстояние. В зависимости от положения трансфокатора один и тот же объектив будет работать как широкоугольный, «нормальный» или телескопический (рис. 4.5).

Во всех камерах сразу после включения объектив устанавливается на минимальное фокусное расстояние, то есть работает как широкоугольный или панорамный. Такая установка при обычной бытовой съемке является самой удоб-

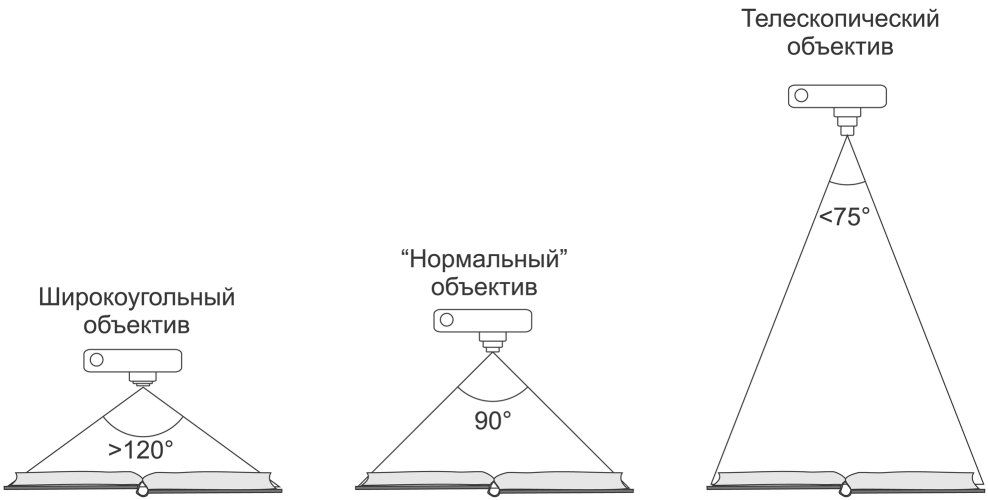
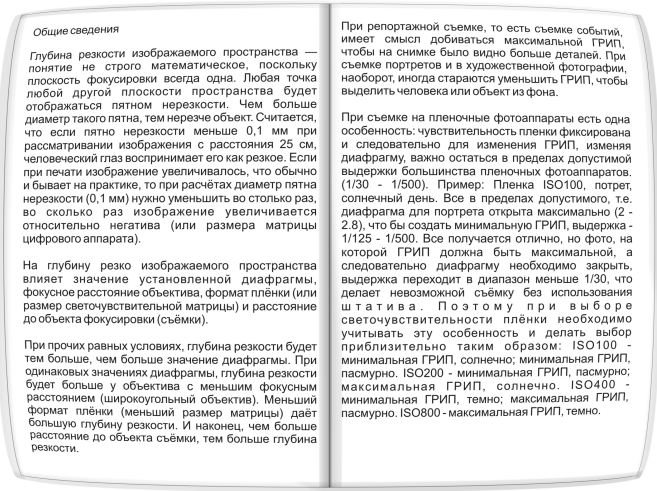


Рис. 4.5 ▾ Применение трансфокатора

ной и универсальной. Если при этом положении трансфокатора поймать во весь кадр стандартный лист бумаги или развернутую книгу, окажется, что фотоаппарат нужно держать примерно в 20–40 см над центром оригинала. Главная проблема короткофокусных широкоугольных объективов – геометрические искажения по краям кадра. На снимке, сделанном широкоугольным объективом с близкого расстояния, кажется, что бумага натянута на поверхность шара (рис. 4.6). Это закономерное оптическое явление – широкоугольный объектив совершенно не предназначен для съемки репродукций.



Общие сведения

Глубина резкости изображаемого пространства — понятие не строго математическое, поскольку плоскость фокусировки всегда одна. Любая точка любой другой плоскости пространства будет отображаться пятном нерезкости. Чем больше диаметр такого пятна, тем нерезче объект. Считается, что если пятно нерезкости меньше 0,1 мм при рассматривании изображения с расстояния 25 см, человеческий глаз воспринимает его как резкое. Если при печати изображение увеличивалось, что обычно и бывает на практике, то при расчётах диаметр пятна нерезкости (0,1 мм) нужно уменьшить во столько раз, во сколько раз изображение увеличивается относительно негатива (или размера матрицы цифрового аппарата).

На глубину резко изображаемого пространства влияет значение установленной диафрагмы, фокусное расстояние объектива, формат плёнки (или размер светочувствительной матрицы) и расстояние до объекта фокусировки (съёмки).

При прочих равных условиях, глубина резкости будет тем больше, чем больше значение диафрагмы. При одинаковых значениях диафрагмы, глубина резкости будет больше у объектива с меньшим фокусным расстоянием (широкоугольный объектив). Меньший формат плёнки (меньший размер матрицы) даёт большую глубину резкости. И наконец, чем больше расстояние до объекта съёмки, тем больше глубина резкости.

При репортажной съемке, то есть съемке событий, имеет смысл добиваться максимальной ГРИП, чтобы на снимке было видно больше деталей. При съемке портретов и в художественной фотографии, наоборот, иногда стараются уменьшить ГРИП, чтобы выделить человека или объект из фона.

При съемке на пленочные фотоаппараты есть одна особенность: чувствительность пленки фиксирована и следовательно для изменения ГРИП, изменяя диафрагму, важно остаться в пределах допустимой выдержки большинства пленочных фотоаппаратов. (1/30 - 1/500). Пример: Пленка ISO100, потрет, солнечный день. Все в пределах допустимого, т.е. диафрагма для портрета открыта максимально (2 - 2.8), что бы создать минимальную ГРИП, выдержка - 1/125 - 1/500. Все получается отлично, но фото, на которой ГРИП должна быть максимальной, а следовательно диафрагму необходимо закрыть, выдержка переходит в диапазон меньше 1/30, что делает невозможной съёмку без использования штатива. Поэтому при выборе светочувствительности плёнки необходимо учитывать эту особенность и делать выбор приблизительно таким образом: ISO100 - минимальная ГРИП, солнечно; минимальная ГРИП, пасмурно. ISO200 - минимальная ГРИП, пасмурно; максимальная ГРИП, солнечно. ISO400 - минимальная ГРИП, темно; максимальная ГРИП, пасмурно. ISO800 - максимальная ГРИП, темно.

Рис. 4.6 ▾ Сферическое искажение

Основной способ избежать таких искажений, которые ухудшают качество распознавания, – снимать книги и другие оригиналы при «среднем» или «дальнем» положении трансфокатора. При этом для большинства камер расстояние от объектива до оригинала составит от 50–60 см до метра.

Получить изображения страниц удовлетворительного качества с помощью мобильного телефона практически нельзя. Некоторые телефоны оборудованы камерой 2 Мпикс и даже более – формально такого разрешения должно хватить для фотографирования, например, страницы книжного формата. Однако из-за крошечного фокусного расстояния объектива на краях снимка возникают значительные геометрические искажения. Попробовать можно, но, скорее всего, распознавание текста с такого снимка даст слишком большое количество ошибок.

Выдержка (экспозиция), диафрагма и чувствительность. Эти три параметра взаимосвязаны, и в камерах начального уровня, как правило, их раздельная регулировка не предусмотрена. Выбирается один из нескольких предустановленных режимов: например, «Пейзаж», «Портрет», «Спорт», «Ночь» и др. Автоматика камеры согласованно отрабатывает значения всех трех настроек, но по разным критериям. Для «спортивного» режима, например, определяющим является подбор самой короткой выдержки, в «портретном» прежде всего полностью открывается диафрагма, а к ней подстраиваются выдержка и, иногда, чувствительность матрицы.

Хорошо, если в камере есть предустановленный режим с названием «Текст» (Text). Такую функцию, например, стала закладывать в некоторые модели своих аппаратов компания Samsung. Если в вашей камере нет специального режима съемки текста, для этой цели лучше других подойдет режим «Портрет».

Если камера позволяет задавать режим съемки вручную («полуавтомат»), выберите режим с приоритетом диафрагмы. Задайте меньшее значение диафрагмы (2,3–4,5). Выдержку камера установит автоматически, в зависимости от освещения.

Формат и качество выходного изображения. Камера сохраняет изображения в файлы во встроенной флэш-памяти или на сменных картах. В большинстве камер, за исключением самых простых, одной из настроек является выбор формата, в котором будут сохраняться полученные снимки.

В профессиональных моделях в меню настройки обычно прямо называется формат файлов, например RAW, BMP, TIFF, JPEG. Дополнительно для каждого из форматов можно выбрать глубину цвета, алгоритм и степень сжатия. Оптимальным для распознавания считается формат TIFF или BMP с глубиной цвета 8 bit (grayscale) и сжатием LZW. Если возможность сохранять файлы с глубиной цвета 8 бит (оттенки серого) в настройках камеры отсутствует, выберите глубину цвета 24 бит. Файлы при этом получатся большего размера.

В более простых моделях вместо этого часто задается один из трех или четырех вариантов качества изображения: от «самого высокого» (Very High) до «низкого» (Low). Как правило, «самому высокому» качеству соответствует формат TIFF, а остальным – JPEG с разной степенью сжатия. Для наших целей лучше подойдет «самое высокое» или «высокое» качество.

Техника съемки

Прежде всего оборудуйте рабочее место. Для съемки нескольких страниц в «полевых условиях» достаточно найти ровную поверхность с хорошим освещением. Как вариант – фотографировать печатные материалы в светлое время суток можно на подоконнике, лишь бы на него не падали прямые солнечные лучи. Снимая без использования штатива, делайте по два-три снимка каждой страницы: если какие-то фотографии получатся нерезкими, у вас будет, из чего выбрать.

Расстояние

При съемке дома или на работе желательно закрепить фотоаппарат на штативе и наладить подсветку оригинала. Использование штатива помогает избежать случайных перемещений камеры во время съемки. Для съемки печатных оригиналов довольно удобен штатив со струбциной. Его можно закрепить на краю стола (оригинал в таком случае кладется на пол), можно соорудить конструкцию из стульев или табуретов и закрепить штатив на ней.

Если штатива нет, как минимум постарайтесь принять при съемке наиболее устойчивое положение. Например, положите на стол две стопки книг подходящей высоты и опирайтесь на них локтями.

Если камера оборудована сильным «зумом», например 8x или 12x, лучше ограничиться средним положением трансфокатора. Конечно, при съемке книги с расстояния 2–3 метра сферические искажения исчезнут практически полностью. Однако здесь вступает в действие другой неприятный фактор. Самое легкое дрожание камеры во время съемки приведет к заметному смазыванию изображения. Поэтому во всех случаях целесообразно выбрать «золотую середину», когда изображение оригинала во весь кадр получается с расстояния *около метра*.

Расположите камеру на расстоянии около 1 метра от оригинала. Объектив должен находиться точно над центром листа или разворота. Работая диском или кнопками трансфокатора, приближайте и удаляйте изображение так, чтобы оно приблизительно вписалось в кадр. Затем, поднимая или опуская фотоаппарат на штативе, добейтесь точного совпадения краев страницы или разворота на изображении с границами кадра. «Прицеливаться» желательно по изображению на дисплее фотоаппарата, так как оптический видоискатель дает большую погрешность.

Освещение

Хорошие условия съемки текста – вблизи окна днем в ясную погоду. Когда естественного освещения недостаточно, оригинал обязательно надо подсветить. Подсветка должна быть равномерной и достаточно рассеянной. Нужно, чтобы свет падал на оригинал почти перпендикулярно. Боковое или скользящее освещение выявит все неровности и шероховатости бумаги: на снимке появятся дополнительный «мусор» и тени. Удачное решение – две люминесцентные настольные лампы, расположенные по бокам от оригинала.

При съемке документов с расстояния меньше метра использование встроенной вспышки ставится под большое сомнение. Многое зависит от модели фотоаппарата и конструкции вспышки: в одних случаях оригинал освещается вполне равномерно, в других на нем появляются резкие светлые пятна. Глянцевая бумага при освещении вспышкой часто дает резкие блики. В общем случае, при фотографировании документов вспышку рекомендуют отключать. Окончательное решение лучше принимать на собственном опыте: попробуйте снять одну страницу со вспышкой и без нее, с разного расстояния при разных установках трансфокатора. Сравнив снимки, выберите среди них лучший по качеству.

Баланс белого

Определившись с расстоянием и освещением, настройте баланс белого (WB, White Balance). Упрощенный вариант – выбор одного из готовых профилей, например «Ясно», «Пасмурно», «Лампы накаливания», «Лампы дневного света». Лучшие результаты дает полностью ручная настройка (MWB, Manual White Balance).

Для настройки положите в качестве оригинала чистый лист такой же бумаги, на которой отпечатан документ. В начале или конце книги почти всегда можно найти чистую страницу или даже разворот. Выполните настройку в соответствии с инструкцией к камере. Важно, чтобы при установке баланса белого использовалось такое же освещение, что и при съемке.

Повышение четкости изображения

Четкость контуров букв на снимке может снижаться по двум основным причинам: это либо неточная фокусировка, либо шевеление аппарата в момент съемки. На таких контрастных объектах, как страницы с текстом, автоматическая фокусировка почти всегда срабатывает правильно. Когда съемка ведется с рук, перед нажатием на спуск нужно выждать секунду или чуть более – даже самой быстродействующей автоматике на отработку фокусировки и экспозиции требуется определенное время. О том, что режим съемки установился, свидетельствует индикатор на дисплее камеры. При установке камеры на штативе каждый раз, когда вы перекладываете оригинал, в поле зрения объектива попадают ваши руки. При этом автоматика будет менять фокусировку, наводясь то на бумагу, то на руки! Следовательно, при установке камеры на штативе и автоматической фокусировке все равно надо давать некоторое время на отработку автофокусировки и следить за индикатором. Чтобы исключить сбой фокусировки во время замены оригиналов, при возможности лучше отключить автофокус и навести камеру на резкость вручную.

При съемке со штатива рекомендуется задействовать таймер спуска (автоспуск). Выставьте таймер на минимальный интервал, например на 1 секунду. Нажмите кнопку затвора, отпустите, и камера сделает снимок через указанное время. Если спускать затвор непосредственно кнопкой без задержки, аппарат вместе со штативом может дернуться в момент снимка, и изображение будет смазано.

Некоторые камеры оборудованы оптическим стабилизатором изображения. Если в вашей камере такая функция предусмотрена, ее желательно включить. Тем не менее для изображений со множеством мелких деталей стабилизация вовсе не заменяет установку аппарата на штативе. Это, скорее, дополнительная и вынужденная мера для случаев, когда штатива нет.

Работа над ошибками

Каждое следующее изображение, как полученное со сканера, так и открытое из графического файла, по умолчанию добавляется в конец текущего документа FineReader. То, как добавляются страницы, видно в рабочем окне **Страницы.**

Если в процессе работы вы обнаружили, что какая-то страница распознана со слишком большим количеством ошибок, возможно, качество изображения не очень хорошее. Просмотрите изображение в окне **Крупный план**. Самые распространенные дефекты изображений – расплывчатые первые или последние символы в каждой строке (со стороны переплета книги), тени и блики, смазанность всего изображения (неточная фокусировка или шевеление камеры при съемке), низкая контрастность слишком светлого или темного изображения (неправильная экспозиция). В таком случае повторно отсканируйте или сфотографируйте эту страницу. Как правило, когда оригинал есть «под рукой», это получается быстрее, чем исправлять большое количество ошибок в распознанном тексте.

Новое изображение будет добавлено в конец открытого документа FineReader. Например, в документе 10 страниц, и вас не удовлетворяет качество страницы № 5. Повторно отсканированная, эта страница пока станет в документе одиннадцатой. Если новое изображение распозналось лучше, замените исходную страницу:

1. В окне **Страницы** щелкните правой кнопкой мыши на странице № 5 и в контекстном меню выберите команду **Удалить страницу из документа**. Выбранная страница будет удалена.
2. Перетащите мышью последнюю (добавленную) страницу туда, где она должна находиться, – на место удаленной.

Во многих случаях программа FineReader сама выявит дефекты при сканировании (съемке) и обратит ваше внимание на такие страницы. На проблемы, возникшие при анализе и распознавании страниц, указывают значки в виде восклицательного знака, появляющиеся на эскизах страниц в окне **Страницы**. Когда программа сочла, что какое-то изображение было отсканировано со слишком низким разрешением, или решила, что для такого мелкого текста разрешение при сканировании нужно увеличить, об этом будет выведено сообщение (рис. 4.7). Подобные сообщения появляются при разных дефектах изображения – недостаточной контрастности из-за неправильной экспозиции, смазанности, сбитой резкости и т. д. Визуально оцените оригинал и, если дефекты изображения действительно возникли в процессе сканирования или фотографирования, получите его изображение заново.

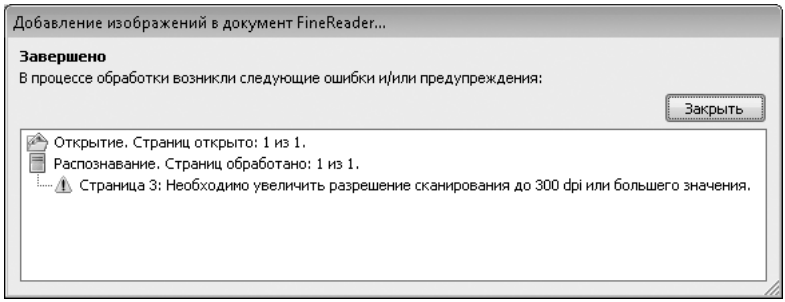


Рис. 4.7 ▾ Сообщение о недостаточном разрешении

В окне **Страницы** щелкните кнопкой мыши на значке страницы, при распознавании которой возникли ошибки. В нижнем левом углу главного окна программы появится полупрозрачное всплывающее сообщение (рис. 4.8).

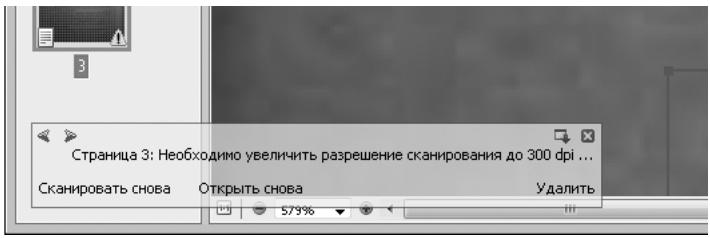


Рис. 4.8 ▾ Всплывающее сообщение об ошибке

Во всплывающем сообщении выберите ссылку **Сканировать снова.** Откроется диалоговое окно сканирования (рис. 2.2). Отсканируйте страницу еще раз, обращая внимание на укладку оригинала в сканер, выбранный режим, разрешение и яркость. Вновь полученное изображение заменит прежнее – порядок страниц в документе не изменится.

Если вы пользуетесь камерой, перефотографируйте страницу, опять же уделив внимание ровной укладке оригинала, освещению и настройкам камеры. Скопируйте файл из камеры на жесткий диск компьютера.

Щелкните кнопкой мыши в окне **Страницы** на значке неудачно распознанной страницы. Во всплывающем сообщении выберите ссылку **Открыть снова** – появится диалоговое окно открытия файла с изображением (см. рис. 2.3). Выберите новое изображение, нажмите кнопку **Открыть**. В документе FineReader изображение страницы будет заменено.

Кроме того, заменить любое изображение всегда можно из диалогового окна свойств страницы. Когда страниц много, это удобнее, чем добавлять по-

вторно отсканированную страницу в конец документа и перетаскивать ее на место по порядку.

В рабочем окне **Страницы** щелкните правой кнопкой мыши на той странице, изображение которой вы хотите заменить. В контекстном меню выберите команду **Свойства**. Откроется диалоговое окно свойств страницы (рис. 4.9).

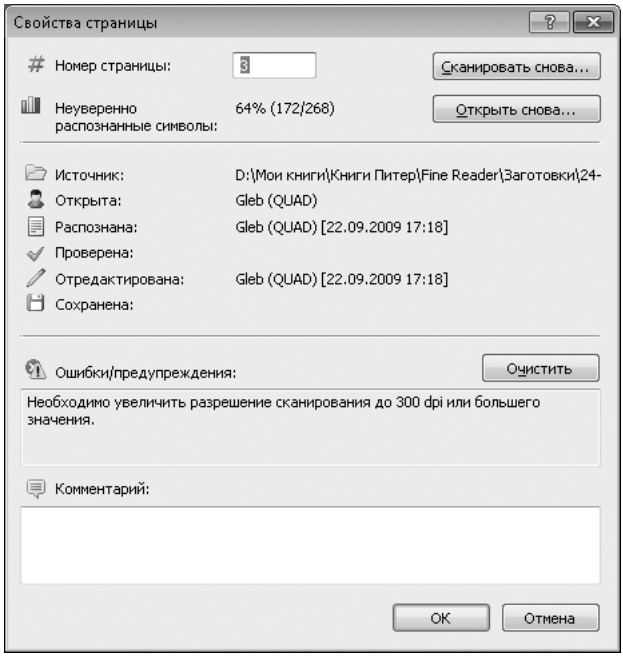


Рис. 4.9 ▾ Диалоговое окно **Свойства страницы**

В верхней части диалогового окна есть кнопки **Сканировать снова** и **Открыть снова**. Чтобы заменить изображение, нажмите одну из них.

Частные случаи

Проще всего сканировать и фотографировать документы, отпечатанные на отдельных листах стандартных форматов. Распространенные недорогие сканеры как раз и рассчитаны на работу с бумагой формата A4 (297×210 мм). Некоторые трудности возникают, когда надо получить изображения крупноформатных оригиналов, например газет или плакатов. Еще один случай «неудобного» оригинала – толстая книга в переплете, который мешает плотно прижать страницы к планшету. Однако практика подсказала несколько простых решений и для таких ситуаций.

Крупноформатные оригиналы

Газетный лист целиком бесполезно фотографировать даже самой современной и дорогой камерой – шрифт на снимке все равно получится слишком мелким для распознавания. То, что нельзя отсканировать или сфотографировать целиком, очевидно, следует снимать по частям.

У большинства сканеров крышка съемная. Для работы с большими оригиналами просто снимите крышку вместе с шарниром. Возможно, придется отыскать с нижней стороны корпуса какие-то фиксаторы или защелки. Уточните в документации к сканеру, как правильно снять крышку. Чтобы прижимать оригинал к стеклу, возьмите подходящий по размеру кусок картона, на него положите книгу или другой не слишком тяжелый груз.

Положите газету на планшет сканера и прижмите ее к стеклу. В диалоговом окне сканирования нажмите кнопку **Просмотр** и получите предварительное изображение первого фрагмента. В группе **Параметры страницы** укажите вариант **Размер текущего выделения**. Перетаскивая мышью границы кадрирующей рамки, задайте область сканирования так, чтобы в нее попала вся колонка по ширине. По высоте же текст, естественно, будет выходить за пределы сканируемой области. Нажмите кнопку **Сканировать** и получите изображение первого фрагмента.

Переложите газету на сканере. Вновь выполните предварительный просмотр. Выделите второй фрагмент. При этом захватите рамкой несколько строк текста, отсканированного в прошлый раз. Отсканируйте очередной фрагмент и т. д.

Таким образом, в несколько приемов вы получите изображение всего газетного листа (рис. 4.10). В документе FineReader каждый фрагмент окажется отдельной страницей. Важно, чтобы фрагменты немного, хотя бы на несколько строк, перекрывали друг друга. Повторяющиеся строки будут распознаны дважды. При окончательном редактировании в окне **Текст** эти повторы легко найти и удалить.

При фотографировании крупноформатного оригинала поступают почти так же. Каждый раз постарайтесь располагать камеру точно над центром снимаемого фрагмента и направлять линию съемки перпендикулярно плоскости листа. Тщательно кадрировать фрагменты не надо – лишнее «обрежется» потом, в процессе распознавания. Для удобства последующего выделения областей постарайтесь только удерживать границы кадра строго параллельно и перпендикулярно строкам и колонкам.

Перед тем как открыть полученные изображения фрагментов в программе FineReader, в диалоговом окне **Опции** (рис. 3.12) на вкладке **Сканировать/Открыть** установите переключатель в положение **Отключить автоматический анализ и распознавание изображения.** Как будет показано в следующей главе, разбейте изображение на области вручную. При выделении областей на каждом снимке вы включите в область распознавания только колонки, которые полностью поместились в кадр по ширине.

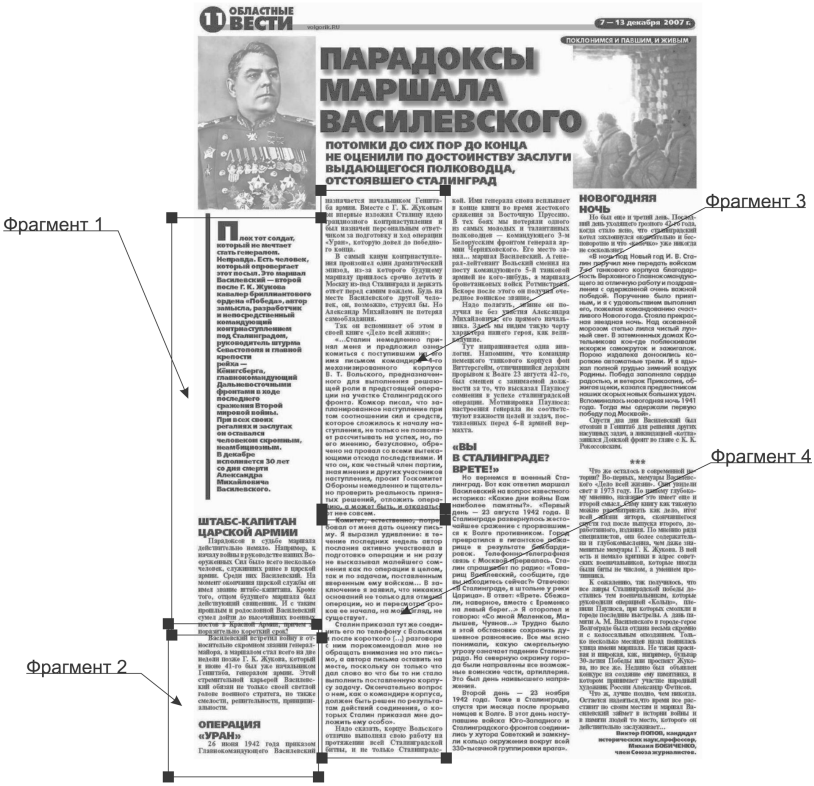


Рис. 4.10 ▾ Сканирование большого оригинала (газеты)

Книги

И при сканировании, и при фотографировании книг главная проблема – как выровнять и зафиксировать разворот страниц. Видимо, жесткий корешок и узкие внутренние поля можно считать своеобразной «защитой от несанкционированного копирования». На сканах и фотографиях правый край текста на четной странице и левый – на нечетной часто выходят нерезкими и затененными. В этом отношении со старыми, слегка «зачитанными» книгами работать проще. Новую книгу перед сканированием можно даже перелистать с начала и до конца, осторожно разглаживая каждый разворот ладонью – так вы сделаете корешок мягче. Правда, из современных изданий с клееным переплетом после такой процедуры порой начинают выпадать отдельные страницы.

Крышка сканера хорошо и равномерно прижимает к стеклу тонкие оригиналы. Хорошо прижать разворот книги, особенно в ее начале и ближе к концу, штатной крышкой сканера, скорее всего, не удастся. Придерживать книгу рукой тоже не стоит – за десяток секунд, пока ползет каретка, вы наверняка ее

хоть чуточку, но сдвинете. В таком случае лучше откинуть крышку и прижимать сканируемую книгу импровизированными средствами – например, другими книгами (рис. 4.11).

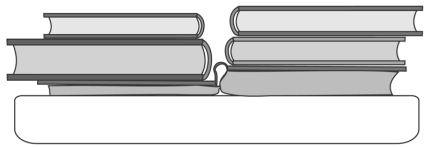


Рис. 4.11 ▾ Стопка книг вместо крышки сканера

По опыту можно сказать: дополнительные 5 секунд, потраченные на тщательную укладку книги на сканер или под фотокамеру, сберегают 5 минут при правке распознанного текста. Такие дефекты, как перекос и трапециевидные искажения, удается практически полностью исправить с помощью автоматической предобработки и встроенного редактора изображений, но искажение строк хуже поддается коррекции. Поэтому постарайтесь избежать появления подобных искажений еще на этапе сканирования или фотографирования оригинала.

Укладка некоторых книг для фотографирования тоже требует определенных ухищрений. Если раскрытую книгу не прижать чем-то, она, естественно, постарается захлопнуться в самый неподходящий момент. В «полевых» условиях придерживать страницы можно и пальцем. При этом позаботьтесь о хорошем освещении: камеру придется держать одной рукой, и короткая выдержка становится чрезвычайно важной.

В стационарных условиях для фиксации фотографируемой книги применяют разные приспособления. Это либо тяжелое стекло – получается «сканер вверх ногами», либо тяжелые линейки.

Прижимать разворот книги стеклом – самое очевидное решение. Однако оно сопряжено с несколькими проблемами.

- Снимать стекло, переворачивать страницу и класть его обратно довольно неудобно. Тем более кусочек обычного оконного стекла размером с книгу вряд ли хорошо расправит разворот, а толстое стекло большого формата еще надо где-то найти.
- Царапины, пыль и следы от пальцев на стекле ухудшают качество снимка.
- Когда оригинал фотографируют через стекло, на снимке могут появиться блики – от ламп подсветки, окон, верхнего освещения. Вспышка даст резкий отблеск почти обязательно!

Из школьного курса физики известно, что «угол падения равен углу отражения». Поэтому существует зона, в которой однозначно не нужно располагать лампы для подсветки оригинала. На рис. 4.12 это пространство заштриховано. Если установить лампу в заштрихованной области, то, отразившись от стекла, лучи от нее могут дать блик на снимке.

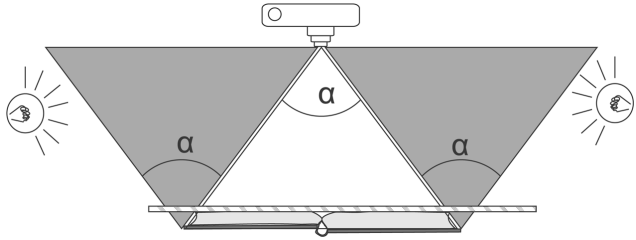


Рис. 4.12 ▾ Расположение ламп для подсветки оригинала без бликов

Схема наводит еще на один практический вывод. При съемке широкоугольным объективом «запретная зона» широка. Лампы придется расставлять так, что свет от них будет падать на оригинал почти сбоку. Под боковым освещением выявляется рельеф бумаги, что в нашем случае совершенно нежелательно. Если же трансфокатором увеличить главное фокусное расстояние объектива, фотоаппарат удастся отдалить от оригинала, а «запретные зоны» сузятся. Лампы можно будет переместить ближе к центру. В результате и свет будет падать на книгу более отвесно, и бликов на стекле удастся избежать.

Можно прижимать страницы двумя линейками, укладывая их вдоль верхнего и нижнего обрезов книги. Весить такие линейки должны не менее килограмма – иначе толка от них нет. В качестве прижимов подойдут два стальных уголка длиной около полуметра, или другие похожие предметы. При фотографировании начальных и последних страниц подложите под обложку с «тонкой» стороны поролон или свернутую ткань, чтобы обе страницы разворота оказались на одной высоте.

Если вы собираетесь переснять цифровой камерой целую книгу, стоит заранее позаботиться о собственном удобстве, так как работа займет не один час. Прежде всего нужен штатив. Если штатива нет, в крайнем случае устройте удобную и устойчивую опору для локтей из двух стопок книг, коробок или других подходящих предметов. Такая подставка позволит при каждом снимке держать камеру почти на одной и той же высоте, да и руки устанут меньше.

Резюме

Качественное изображение – залог высокого качества распознавания. Основные требования, которые программа FineReader предъявляет к изображениям, – достаточное разрешение, четкость, контрастность и отсутствие бликов или теней.

Точность распознавания текста с отсканированных оригиналов или с их фотографий практически одинакова. Выбирая способ получения изображения, ориентируйтесь в первую очередь на то, какая техника есть в наличии, и на то, где вы будете этой техникой пользоваться.

Для сканирования в программу FineReader оптимальным является разрешение 250 или 300 DPI и режим «оттенки серого». При сканировании книжных

страниц позаботьтесь о том, чтобы оригинал плотно прилегал к стеклу сканера всей поверхностью. В противном случае на изображении возможны дефекты.

При фотографировании установите трансфокатор («зум») так, чтобы изображение оригинала во весь кадр получалось с расстояния около метра. Съемка с близкого расстояния дает сильные геометрические искажения по краям кадра. Желательно закрепить камеру на штативе и пользоваться автоспуском. При съемке «с рук» особенно важно хорошее освещение, чтобы выдержка была как можно короче. Чтобы выяснить истинные возможности своего фотоаппарата, обязательно прочитайте инструкцию к нему – при обычной бытовой съемке о наличии многих настроек и функций мы зачастую даже не подозреваем.

Однако изображение идеального качества удается получить не всегда. Качество изображений ограничивают и характер оригиналов, и технические характеристики оборудования, и конкретные условия работы. В программе FineReader 10 заложен целый ряд средств для исправления дефектов полученных изображений и повышения качества распознавания. В следующей главе мы обсудим возможности этих инструментов и то, как их использовать наилучшим образом.

Глава 5 Обработка и анализ изображений

После сканирования оригинала или открытия графического файла программа FineReader обрабатывает полученное изображение. Несколько операций предобработки выполняются автоматически, если были установлены соответствующие флажки в диалоге сканирования или открытия файла. Кроме того, встроенный редактор изображений позволяет выполнить еще несколько операций в полуавтоматическом режиме: вы указываете, какие исправления следует внести в изображение, а программа FineReader 10 выполняет указанные действия.

Строго говоря, в обработке не нуждаются лишь изображения, отсканированные или снятые с идеальным качеством. Во всех остальных случаях обработка изображений в большей или меньшей степени помогает улучшить качество распознавания. Однако надо понимать, что возможности улучшения качества изображений средствами программы FineReader 10 не безграничны. Поэтому при низком качестве изображения какой-то страницы, если в вашем распоряжении есть оригинал, эту страницу правильнее отсканировать или сфотографировать заново.

Анализ изображения заключается в том, что программа FineReader разбивает изображение на отдельные области. При этом учитываются характерные признаки фрагментов текста и тех участков, которые текстом явно быть не могут. Программа решает, что нужно сделать с каждой областью: распознать ее содержимое как слитный текст, оформить распознанный текст в виде таблицы или штрих-кода, либо передать содержимое области в выходной документ без распознавания, как картинку. При автоматическом анализе документов со сложной структурой возможны ошибки, поэтому в окне **Изображение** вы можете обозначать области самостоятельно и указывать программе, что каждая из этих областей содержит: текст, таблицу и т. д.

«Мнение» программы о том, каким должно быть «хорошее» изображение и какова сложность документа, может ощутимо отличаться от вашего собственного. Поэтому оценивать необходимость той или иной обработки изображения и коррекции разбивки на области стоит не столько по своему впечатлению от картинки в рабочих окнах **Изображение** и **Крупный план**, сколько по конечному результату – проценту неуверенно распознанных символов. На точность распознавания ощутимо влияют правильный выбор языков, использование подходящих эталонов и словарей. Речь об этом пойдет в следующей главе. Пока же просто условимся, что в диалоговом окне **Опции** на вкладке **Сканировать/Открыть** переключатель установлен в положение **Автоматически распознавать полученные изображения**, а язык в раскрывающемся списке рабочего окна **Страницы** выбран правильно.

Кроме того, в диалоговом окне **Опции** на вкладке **Документ** переключатель **Тип печати документа** должен быть установлен в положение, соответствующее шрифту распознаваемого документа:

- **Авто** – программа сама старается определить тип печати. Это настройка по умолчанию, лучше всего подходящая для документов, отпечатанных типографским способом, на струйном или лазерном принтере;
- **Пишущая машинка** – распознавание текста, напечатанного на пишущей машинке. Программа в этом случае использует особый эталон, ведь все пишущие машинки печатали одним из двух стандартных шрифтов («канцелярским» или «портативным»);
- **Факс** – настройка для распознавания текстов, переданных по факсу. В этом случае программа также будет использовать особый эталон.

Тогда после открытия или сканирования изображения вы увидите в рабочем окне **Текст** результат распознавания и сможете оценить число ошибок. Подчеркнем, что эти настройки нужно проверить до того, как вы начнете сканировать оригиналы или открывать фотографии.

Обработка изображения

Автоматическая обработка включает в себя поворот и устранение перекоса изображения, деление книжного разворота на две страницы, определение ориентации страницы, а также при необходимости корректировку разрешения. Какие из этих действий будут выполняться, зависит от того, какие флажки установлены на вкладке **Сканировать/Открыть** в группе **Обработка изображений**. Такие же флажки будут установлены или сняты в диалогах сканирования или открытия файлов. Перед сканированием или открытием файла вы сможете изменить настройку непосредственно в диалоге сканирования или открытия файла.

Кроме того, в любое время любое загруженное в программу FineReader изображение можно дополнительно обработать вручную. Встроенный редактор изображений позволяет принудительно повернуть изображение, исправить перекос и искажение строк, обрезать края, стереть часть изображения, уменьшить шум и устранить размытие.

Для большинства оригиналов наилучшие результаты распознавания достигаются, когда в процессе обработки выполняется только автоматический поворот изображения и устраняется перекос – это настройки программы FineReader по умолчанию. С настройкой остальных функций обработки есть смысл экспериментировать, если оригинала в вашем распоряжении нет, а изображения страниц были сделаны где-то раньше, или вы скачали их из Интернета.

Настройка автоматической обработки

Настройки, заданные на вкладке **Сканировать/Открыть** диалогового окна **Опции**, влияют на обработку изображений, получаемых после того, как были выбраны эти настройки. На изображения, полученные раньше, изменение текущих установок не действует. Поэтому включать или отключать отдельные функции обработки целесообразно перед тем, как вы отсканируете или откроете первую страницу документа.

Вызовите диалоговое окно **Опции** и перейдите на вкладку **Сканировать/Открыть** (см. рис. 3.12). В группе **Обработка изображений** установите или снимите нужные флажки. Рассмотрим, для чего нужна каждая из функций обработки и в каких случаях стоит ее использовать.

- **Выполнять предобработку изображений**. По умолчанию флажок установлен – функция включена. Программа автоматически старается повернуть изображение так, чтобы строки стали строго горизонтальными, боковые границы текстовых фрагментов – вертикальными, а также исправить разрешение, если это необходимо. Такая коррекция изображения полезна практически всегда, поскольку даже при самой тщательной укладке оригинала в сканер или кадрировании снимка небольшой перекос все равно возможен.
- **Определять ориентацию страницы**. По умолчанию флажок установлен – функция включена. Программа при необходимости поворачивает изображение на 90° или 180°, чтобы «верх» оказался вверху. Стандартный лист на фотографиях, а книжный разворот при сканировании по определению получается «лежащим на боку», а при сканировании очень часто оригинал кладут на планшет «вверх ногами». Поэтому функцию автоматической ориентации страницы рекомендуется включать всегда.
- **Делить разворот книги**. Когда флажок установлен, текст с каждой страницы разворота будет размещен на отдельной странице выходного документа. Если флажок снят, текст со страниц разворота будет размещен в выходном документе в две колонки на одной странице.

Принятая по умолчанию настройка (установлены все три флажка) является оптимальной для работы с подавляющим большинством оригиналов. Если вы работаете с простыми по структуре оригиналами высокого качества, например сканируете обычный текст с листов, распечатанных на лазерном принтере, или страниц предварительно расшитой книги, флажки можно и снять. За счет исключения отдельных операций работа с документом несколько ускорится.

Как правило, предобработка изображений занимает в несколько раз меньше времени, чем следующие за ней анализ и распознавание.

Обработка в Редакторе изображений

Рассмотренная ранее обработка задается в диалоге **Опции** перед началом работы с документом, и ей автоматически подвергаются все сканируемые или открываемые изображения. Для дополнительной обработки изображения после того, как оно получено со сканера или открыто из файла, откройте **Редактор изображений** (меню **Страница ➢ Редактировать изображение страницы...**, кнопка **Редактировать** на панели инструментов рабочего окна **Изображение** или сочетание клавиш **Ctrl+Shift+C**).

В левой части окна редактора (рис. 5.1) показывается текущее изображение. Под ним находятся кнопки со стрелками для пролистывания всех изображений, содержащихся в документе FineReader. Правую часть окна редактора занимают кнопки групп инструментов. При нажатии на любую кнопку под ней раскрывается небольшая панель, на которой размещены кнопки инструментов указанной группы.

Такая организация панели очень удобна – она компактна, но позволяет вызвать любой из многих инструментов. Группы расположены в том порядке, в котором логично использовать функции ручной обработки.

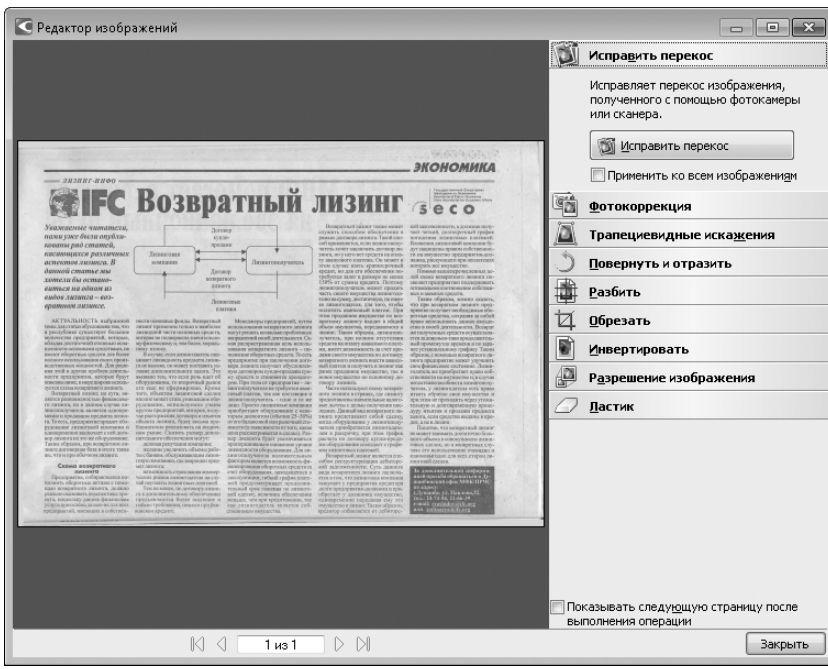


Рис. 5.1 ▾ Редактор изображений, группа **Исправить перекос**

- Группа **Исправить перекос**.
  - **Исправить перекос изображения**. При нажатии на эту кнопку программа старается исправить перекос изображения в целом, ориентируясь на его границы. Фотографии бывают перекошены, если при съемке объектив смещен относительно центра оригинала. На отсканированных изображениях такой дефект появляется, если страницы были неровно уложены на планшет сканера.

    Кнопку **Исправить перекос изображения**, как и кнопки в группе **Фотокоррекция**, рекомендуется нажимать однократно. Повторные попытки коррекции, скорее всего, не устранят дефекта.

    В нижней части этой и некоторых других групп находится флажок **Применить ко всем изображениям**. По умолчанию он снят: все действия выполняются только с текущей страницей документа. Прежде чем установить этот флажок, подумайте: нужно ли выполнять такую же обработку всех страниц, или проблема присутствует только на текущей странице?
- Группа **Фотокоррекция**.
  - **Исправить искажение строк**. При нажатии на эту кнопку программа старается выровнять искривленные строки. Такие дефекты чаще всего появляются при фотографировании или сканировании книг, когда концы строк изгибаются вблизи переплета.
  - **Устранить размытие**. При нажатии на эту кнопку программа старается сделать границы элементов изображения более четкими и контрастными. Фотографии получаются размытыми из-за неточной фокусировки или дрожания камеры в момент съемки.
  - **Уменьшить шум**. При нажатии на эту кнопку программа сглаживает мелкие неоднородности изображения. Шум, или «мусор», может появиться на фотографии из-за грубой структуры бумаги, особенно при косом или боковом освещении оригинала, а также при неудачно выбранной экспозиции в условиях недостаточного освещения.
- Группа **Трапециевидные искажения**.

  Перспективные искажения обычно возникают из-за того, что камера при съемке была смещена относительно центра оригинала, а сама съемка велась под углом. Когда выбрана эта группа, в рабочей области окна над изображением появляется рамка с маркерами по углам.

  Перетаскивая рамку мышью, вы можете перемещать ее над изображением страницы без изменения формы и размеров. При наведении на угловые маркеры указатель мыши принимает вид двунаправленной стрелки. Перетаскивая по очереди каждый из углов рамки, постарайтесь совместить их с углами изображенной на снимке страницы (рис. 5.2).

  Таким образом, рамка приобретет форму трапеции или даже неправильного четырехугольника. Нажмите кнопку **Применить**. Часть изображения, оставшаяся за границами рамки, будет обрезана, а часть, ограниченная рамкой, будет вписана в прямоугольник. В результате перспективные

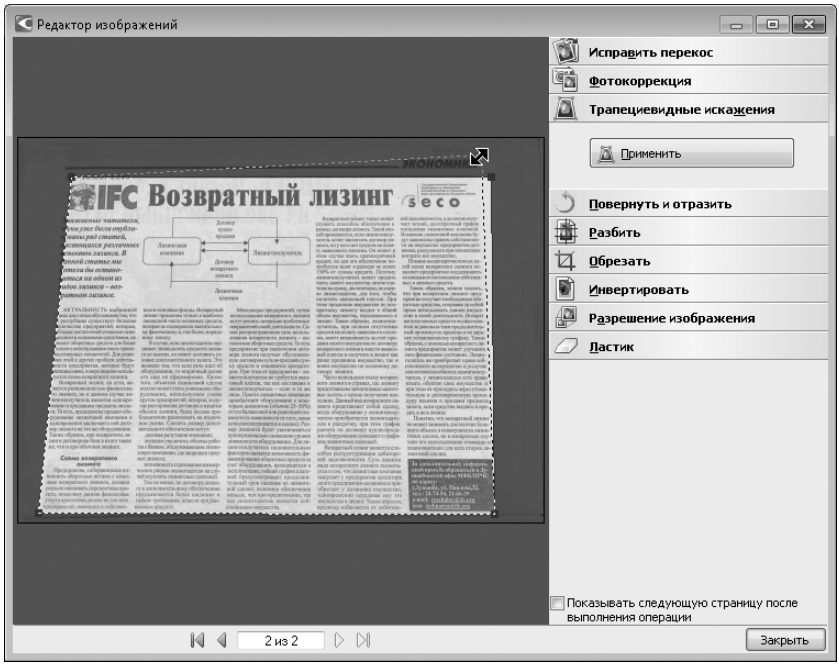


Рис. 5.2 ▾ Коррекция трапециевидных искажений

изображения компенсируются, а углы изображенной на фотографии страницы станут близки к прямым.

- Группа **Повернуть и отразить**.
  Три кнопки поворачивают изображение на 90° влево или вправо, или на 180°. Еще две кнопки, **Отразить сверху вниз** и **Отразить слева направо**, выполняют зеркальное отражение картинки относительно горизонтальной или вертикальной осей. Вращать изображение нужно, если программа не повернула его автоматически, например был снят флажок **Определять ориентацию страницы** на вкладке **Сканировать/Открыть** диалогового окна **Опции**. Зеркальное отражение требуется редко, но на всякий случай в редакторе предусмотрена и такая возможность.
- Группа **Разбить**.
  Инструменты этой группы позволяют вручную разделить изображение на несколько частей, чтобы программа далее работала с каждой частью как с отдельным изображением страницы. Типичная задача – деление книжного разворота. По идее, когда на вкладке **Сканировать/Открыть** диалогового окна **Опции** установлен флажок **Делить разворот книги**, программа должна разбивать разворот автоматически. Однако на практике это происходит не всегда: на изображении разворота должна быть четко заметна линия, на которую программа могла бы ориентироваться. Кнопка **Разбить изображение** запускает процесс разрезания текущего

изображения по предварительно установленным разделительным линиям. Перед тем как нажать эту кнопку, необходимо нанести на изображение один или несколько разделителей:

– чтобы разбить разворот вручную, нажмите кнопку **Добавить вертикальный разделитель**. Наведите указатель мыши на изображение в окне редактора. Указатель примет вид карандаша, проводящего вертикальную черту (рис. 5.2);
– установите черту на то место, где должны разделяться смежные страницы, и щелкните кнопкой мыши. Разделительная линия закрепится на изображении (рис. 5.3).

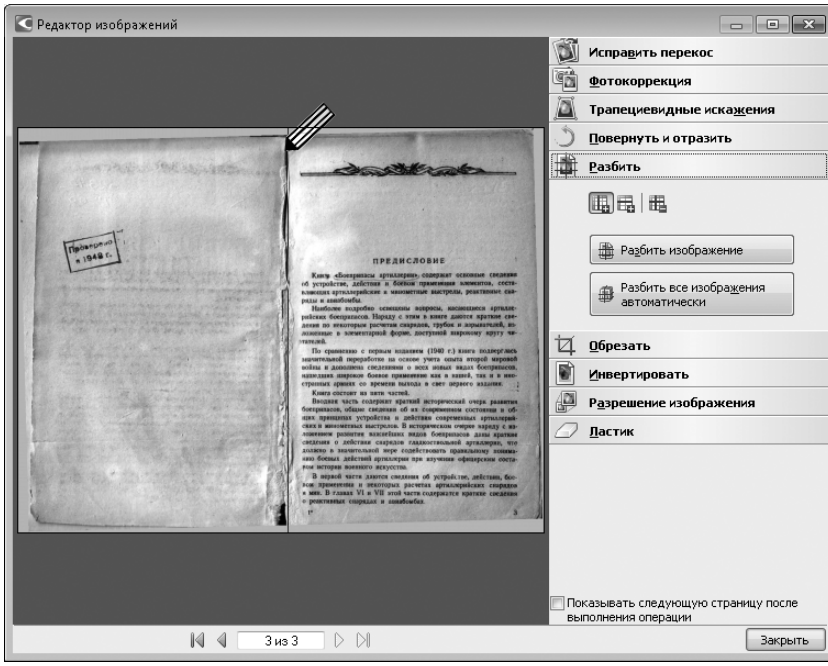


Рис. 5.3 ▾ Деление изображения страницы

При необходимости откорректируйте положение разделителя, перетаскивая его мышью влево или вправо. Когда указатель мыши находится не над разделителем, он вновь приобретает форму карандаша – так вы можете установить еще один разделитель.

Иногда изображение нужно разбить не только по горизонтали, но и по вертикали. Например, на снимке изображены сразу четыре страницы малого формата. Чтобы установить горизонтальные разделители, нажмите кнопку **Добавить горизонтальный разделитель**. Од-

новременно на изображении можно нарисовать сколько угодно разделителей, и вертикальных, и горизонтальных. Если вы случайно нарисовали лишнюю разделительную линию, нажмите кнопку **Удалить все разделители**, а затем установите разделители снова. Вплоть до следующего шага положение разделителей только размечается: само изображение пока не изменяется;

– окончательно уточнив положение разделителей, нажмите кнопку **Разбить изображение**. Появится предупреждение, что это необратимая операция. Нажмите кнопку **OK** для подтверждения действия. Изображение будет разрезано по разделительным линиям, и каждая часть станет отдельной страницей документа FineReader.
  - Кнопка **Разбить все изображения автоматически** запускает такое же деление всех изображений, которое программа выполняет при установленном флажке **Делить разворот книги**. Эта операция не всегда проходит успешно: если программа не находит на изображении явных ориентиров для деления, она выводит сообщение о невозможности разбить страницу. В таком случае следует установить разделители вручную и нажать кнопку **Разбить изображение**. Выполните эту процедуру для каждого изображения, которое нужно разделить на части.

❑ Группа **Обрезать**.

Как и при задании области сканирования, инструменты этой группы позволяют оставить лишь часть изображения. Все, что находится за пределами выделенного участка, удаляется. Обрезка полезна, если в кадр попал какой-то неоднородный фон, который может быть распознан как рисунок. Особенно это актуально, когда результат потом передается в файл PDF – проще сразу обрезать исходное изображение, чем потом править выходной документ. Разумеется, с тем же успехом изображение можно кадрировать еще в ходе сканирования, но фотографии чаще приходится обрезать именно таким способом.

– В раскрывающемся списке **Формат** выберите один из стандартных форматов бумаги. На изображении появится прямоугольная кадрирующая рамка с маркерами по углам (рис. 5.4).
– При необходимости уточните размеры и положение рамки, перетаскивая мышью ее границы и углы.
– Нажмите кнопку **Обрезать края изображения**. Часть изображения, находящаяся за пределами кадрирующей рамки, будет удалена.

Если при этом был установлен флажок **Применить ко всем изображениям**, то точно по тем же границам будут обрезаны все изображения документа. При обработке фотографий такую возможность надо использовать очень осторожно. Даже если камера была установлена на штативе, от кадра к кадру положение «полезного» изображения может немного меняться, и есть риск на отдельных страницах случайно срезать часть

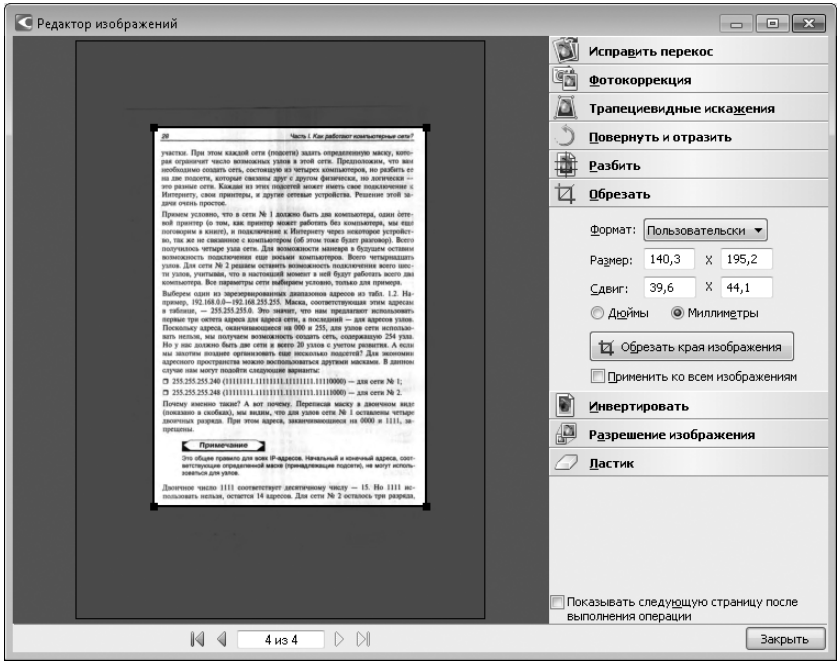


Рис. 5.4 ▾ Обрезка изображения

текста. Поэтому надежнее обрезать каждую страницу по очереди и зрительно контролировать, что попало в рамку, а что – нет.

- Группа **Инвертировать**.
  Кнопка **Инвертировать цвета** превращает изображения в негатив. Ту же функцию несет флажок **Инвертировать цвета изображения** на вкладке **Сканировать/Открыть** диалогового окна **Опции**. Подобное преобразование бывает нужно, когда светлый шрифт напечатан на темном фоне.
- Группа **Разрешение изображения** содержит переключатель и кнопку **Применить**. Выберите переключателем одно из трех стандартных разрешений, либо установите переключатель в положение **Другое** и введите в текстовое поле значение разрешения. Нажмите кнопку **Применить**. Для данного изображения будет принято указанное разрешение.
  Как уже было сказано, изображение само по себе никаким разрешением не обладает – данные о разрешении содержатся в служебной информации файла. Программы, руководствуясь этими данными, определяют, каким должен быть размер изображения при выводе его на экран или при печати. От изменения сведений о разрешении число точек-пикселов в изображении не меняется. Однако программе FineReader сведения о разрешении нужны, чтобы правильно выбрать размер страницы и шрифта в выходном документе.

Например, если вы отсканируете оригинал на листе формата A4 с разрешением 300 DPI и не будете его менять в редакторе, программа выведет документ на такую же по размеру страницу. При просмотре в окне **Текст** в масштабе 1:1 буквы на экране получатся примерно такой же высоты, как и в оригинале. Если же вы при обработке укажете разрешение равным 150 DPI, программа посчитает, что отсканированный оригинал был напечатан на листе формата A2, и выведет документ на страницу такого формата. Соответственно, в выходном документе все шрифты будут в два раза крупнее, чем на оригинале.

- **Ластик**.

  Когда выбран этот инструмент, указатель мыши над изображением превращается в квадратик, изображающий стирательную резинку. Выделяя прямоугольные области, вы можете стирать отдельные участки изображения до фона страницы.

  Операция требует внимания и отнимает не меньше времени, чем правка распознанного текста. Очевидно, ластиком есть смысл пользоваться тогда, когда вы хотите после распознавания передать документ в формат PDF и получить визуально похожую копию оригинала. Например, так иногда копируют разные бланки, убирая с них ненужные надписи и подписи. В остальных случаях гораздо проще отредактировать итоговый документ в окне **Текст** или программе Microsoft Word.

Анализ изображений

Когда в диалоговом окне **Опции** на вкладке **Сканировать/Открыть** переключатель установлен в положение **Автоматически распознавать полученные изображения** или **Автоматически анализировать полученные изображения**, программа автоматически разбивает каждое получаемое изображение на области. Когда переключатель установлен в положение **Отключить автоматический анализ и распознавание изображения**, программа только открывает изображение в окне **Изображение** и ждет команды от пользователя.

Области изображения

При настройке, принятой по умолчанию (**Автоматически распознавать полученные изображения**), вскоре после сканирования страницы в рабочем окне **Изображение** вы видите, как программа автоматически разметила области: на изображении появляются цветные рамки (рис. 5.5), а в окне **Текст** – результат распознавания. Если автоматический анализ был отключен, нажмите кнопку **Анализ страницы** на панели инструментов окна **Изображение** либо щелкните правой кнопкой мыши на изображении и в контекстном меню выберите команду **Анализ страницы.** Программа поделит изображение на области и при нажатии сочетания клавиш **Ctrl+E**. Целесообразно сразу же оценить, как программа FineReader разбила изображение на области.

Рис. 5.5 ▾ Деление изображения на области

Существуют четыре типа областей:

- Текст. Содержимое распознается как слитный текст, состоящий из абзацев. Область обводится зеленой рамкой;
- Таблица. Содержимое распознается как таблица, состоящая из ячеек. Ячейки образуют строки и столбцы. Область помечается синей рамкой;
- Картинка. Содержимое области не распознается и передается в выходной документ «как есть», то есть как иллюстрация. Область помечается красной рамкой;
- Штрих-код. Штрих-код по сути является особым «алфавитом», который предназначен для считывания торговым оборудованием. Комбинации полосок разной толщины обозначают разные цифры, а под ними значение штрих-кода обычно дублируется арабскими цифрами. Программа FineReader способна распознавать штрих-коды. Область помечается ярко-зеленой рамкой.

Особый тип области – *Зона распознавания*. Она помечается серой рамкой. Выделяя область такого типа, вы обозначаете, какую часть изображения нужно проанализировать и распознать. Программа автоматически решит, на какие области (Текст, Таблица, Картинка, Штрих-код) нужно разбить выделенную зону распознавания.

Каждой области присваивается номер. Области автоматически нумеруются слева направо и сверху вниз. В выходном документе распознанные фрагменты будут скомпонованы именно в таком порядке. Номер выделенной в настоящий момент области отображается в поле со счетчиком **№** на вкладке **Свойства обла-**

сти. С помощью этого поля со счетчиком вы можете изменить номер выделенной области.

В простейшем случае (страница текста без иллюстраций, заголовков, колонтитулов и т. п.) все изображение программа считает одной областью типа текст. В документе со сложной структурой каждая колонка текста рассматривается как отдельная область, другие области содержат рисунки или таблицы. Если вы видите, что при анализе страницы программа допустила ошибки, необходимо вручную изменить тип и границы некоторых областей. Возможно, какие-то области проще удалить, а потом обозначить их заново.

Исправление разбивки на области

В качестве примера возьмем документ со сложной структурой – статью в газете (рис. 5.5). Логотип в верхнем левом углу программа совершенно справедливо восприняла как картинку. В отношении заголовка и основного текста статьи все тоже в порядке – они размечены как области с текстом. Однако анализ некоторых участков изображения можно оспорить.

Схему программа посчитала несколькими текстовыми фрагментами, а рамки и стрелки этого рисунка игнорировала (рис. 5.6). Программа, в принципе, совершенно права: она нашла символы и расценила их как текст. Содержимое областей с 8 по 11 будет распознано как текст, и в выходном документе вместо схемы помещены несколько не связанных между собой абзацев (рис. 5.7).

Рис. 5.6 ▾ Схема разбита на области неудачно

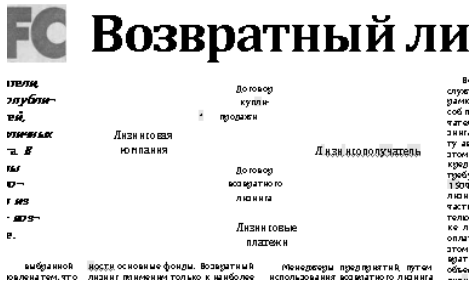


Рис. 5.7 ▾ Результат распознавания неудачно определенных областей

Чтобы исправить ситуацию, необходимо переопределить области вручную. Сделать это можно разными способами. Чтобы продемонстрировать доступные действия, рассмотрим два варианта – результат получится один и тот же.

1. Удалите области 8–11:
   1) щелкните кнопкой мыши на области 8. Область будет выбрана – рамка вокруг нее станет ярче, а на ее углах появятся «квадратики»;
   2) щелкните на выбранной области правой кнопкой мыши и в контекстном меню выберите команду **Удалить область**. Или нажмите на клавиатуре клавишу **Delete**. Область удалена;
   3) таким же образом удалите области 9, 10 и 11.
2. Измените тип области 7 с *Текст* на *Картинка*:
   1) щелкните кнопкой мыши на области 7. Область будет выбрана;
   2) щелкните на выбранной области правой кнопкой мыши и в контекстном меню выберите команду **Изменить тип области ➢ Картинка.** Тип области изменится, и рамка вокруг нее станет красной.
3. Измените границы области 7 так, чтобы в нее попала вся схема. Перетаскивайте мышью края рамки или маркеры на ее углах.

Можно поступить чуть иначе. Сначала удалите области с 7 по 11, а потом обозначьте на изображении новую область типа *Картинка*.

1. Нажмите на панели инструментов кнопку **Картинка**. Указатель мыши превратится в стрелку со значком рисунка. Это режим «рисования» областей.
2. Щелкните кнопкой мыши на изображении и, удерживая кнопку, перемещайте мышь. Дойдя до противоположного угла обозначаемой области, отпустите кнопку мыши. Границы новой области обозначены.
3. При необходимости откорректируйте границы области, перетаскивая их мышью.
4. Чтобы вернуться из режима рисования областей к обычному режиму выбора объектов, нажмите на панели инструментов кнопку **Выбор**. Указатель мыши вновь приобретет вид простой стрелки.

ПРИМЕЧАНИЕ

*Для рисования области типа Текст нажмите на панели инструментов кнопку **Текст**, а для рисования области типа Таблица нажмите на панели инструментов кнопку **Таблица**. Перемещать границы любых областей и вызывать контекстные меню можно в любом режиме. Вся разница в том, что в режиме рисования областей щелчок кнопкой мыши с перетаскиванием служит для рисования области* указанного *типа.*

В результате в окне **Изображение** на месте пяти областей типа Текст появится одна область типа Картинка. При распознавании страницы содержимое этой области будет перенесено в выходной документ без распознавания как иллюстрация (рис. 5.8).

Рис. 5.8 ▾ Часть изображения перенесена в документ как иллюстрация

Как правило, проблемы с автоматическим разбиением на области возникают при работе с документами, в которых содержатся логотипы, разного рода схемы, а также формулы. Логика работы программы FineReader такова, что все, похожее на текст, она старается распознать именно как текст. Однако некоторые элементы целесообразно переносить в итоговый документ без распознавания. В частности, не стоит распознавать «многоэтажные» формулы с дробями, корнями, дифференциалами и т. п. В выходном документе вы получите вместо них смесь символов.

Основное решение в таких ситуациях – корректировка областей вручную. В зависимости от количества и расположения на странице элементов, которые распознавать не стоит, можно рекомендовать две тактики.

Когда формул и схем на странице много и они разбросаны по всему тексту (например, в учебнике физики или радиотехническом справочнике), предоставьте программе FineReader сначала выполнить автоматический анализ. В подобных случаях она часто обозначает области сложной формы – с уступами, выступами и врезками (рис. 5.9).

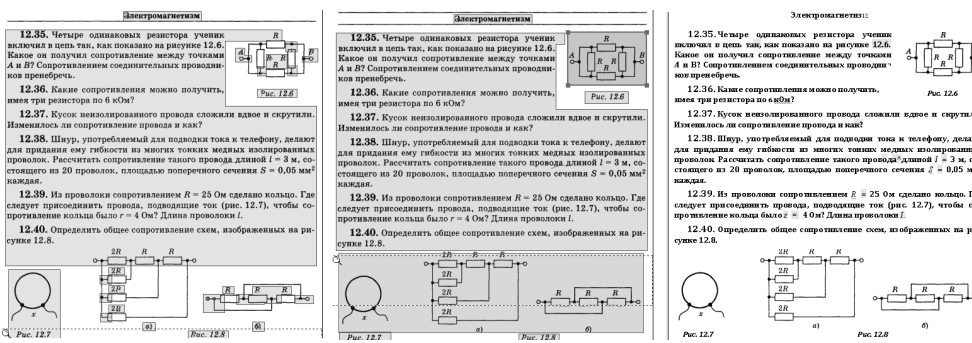


Рис. 5.9 ▾ Автоматическая разбивка, ее корректировка и результат распознавания

Затем откорректируйте области вручную. Уточните границы областей типа Текст, удалите лишние, а на объекты, которые нужно оставить «как есть», нанесите области типа Картинка.

Если вставок на странице одна-две и все они примыкают к краям страницы, целесообразно нарисовать области вручную. Заранее отключите в диалоговом окне **Опции** автоматический анализ и распознавание. Иначе для удаления всех областей щелкните на изображении правой кнопкой мыши и в контекстном меню выберите команду **Удалить все области и текст**. Обозначьте области на изображении. Проще и быстрее обозначать области прямоугольной формы, так чтобы весь текст попал в несколько прямоугольных областей.

ПРИМЕЧАНИЕ

Нанося области типа Текст, не бойтесь их взаимного перекрытия по высоте! Если области немного заходят друг на друга, программа совершенно правильно воспринимает такую ситуацию. В распознанном документе текст будет аккуратно состыкован. Важно захватить границами области весь текст по ширине – если какой-то символ окажется за ее пределами, распознан он не будет.

Чтобы вручную создать область сложной формы, сначала нарисуйте прямоугольную область, захватывающую по высоте весь текст, а по ширине – абзацы наименьшей ширины. Затем добавьте к этой области прямоугольные фрагменты, чтобы захватить текст полностью.

1. Переключитесь в режим рисования области типа Текст одним из двух способов:
   - в строке меню выберите команду **Области ➢ Выделить область ➢ Выделить область Текст**;
   - нажмите на панели инструментов **Изображение** кнопку **Текст**.

   Указатель мыши приобретет вид стрелки со значком выделения текстовой области.
2. Выделите на изображении прямоугольную область (рис. 5.9 слева).
3. Переключитесь в режим добавления нового участка к существующей области одним из трех способов:
   - щелкните на области, которую вы хотите изменить. Над правым верхним углом области появятся две полупрозрачные пиктограммы: **Добавить часть к области** и **Удалить часть области**. Щелкните кнопкой мыши на первой из них;
   - в строке меню выберите команду **Области ➢ Выделить область ➢ Добавить часть к области**;
   - нажмите и удерживайте на клавиатуре клавишу **Shift**.

   Указатель мыши приобретет вид стрелки со значком «плюс».
4. Добавьте к области новые части, обводя их указателем мыши (рис. 5.10 справа).

Чтобы удалить какую-либо часть из существующей области, в строке меню выберите команду **Области ➢ Выделить область ➢ Удалить часть области** или нажмите и удерживайте на клавиатуре клавишу **Alt**. Указатель мыши приоб-

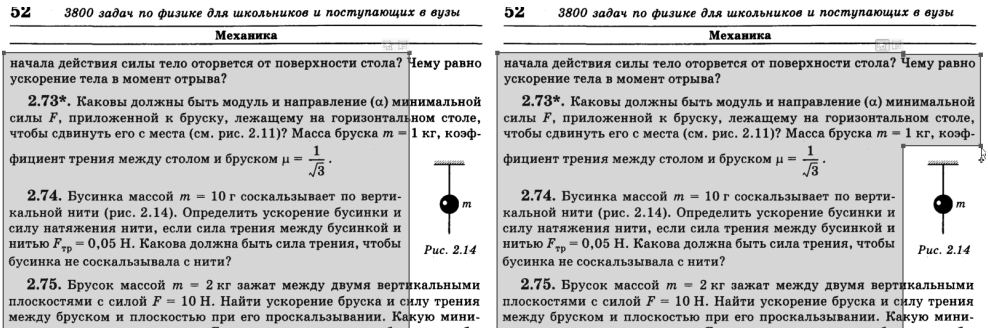


Рис. 5.10 ▾ Рисование области сложной формы

ретет вид стрелки со значком «минус». Обведите указателем мыши части области, и эта часть будет исключена (рис. 5.11). В данном примере из области исключена формула: ее невозможно распознать как текст, и лучше заключить эту формулу в область типа Картинка.

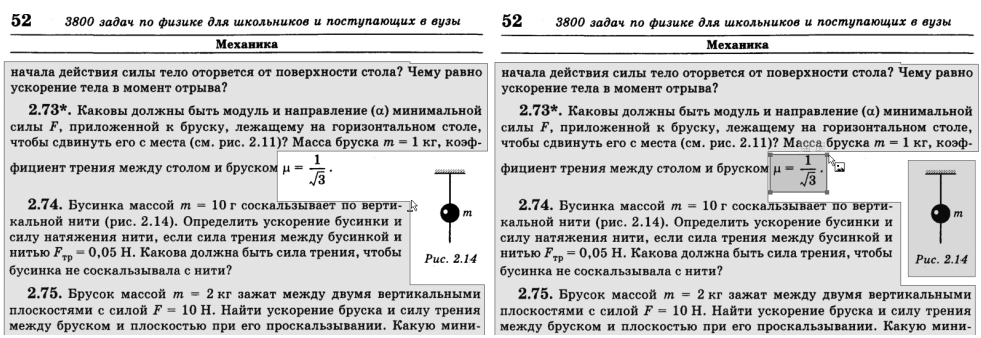


Рис. 5.11 ▾ Завершение разметки страницы на области

В программе FineReader 10 предусмотрен еще один тип области – *Зона распознавания*. По сути, это предварительная разметка: вы помечаете какую-то часть изображения, а программа затем анализирует и распознает все, что вошло в эту область, но игнорирует остальную часть изображения.

Использовать такой прием удобно, когда автоматический анализ в диалоге **Опции** отключен, а из каждого изображения надо распознать лишь определенную часть. Например, зачастую нет нужды распознавать на каждой странице колонтитулы, которые расположены в верхней и нижней частях листа.

В окне **Изображение** нажмите кнопку **Зона распознавания**. Выделите мышью часть изображения, которую следует проанализировать и распознать. Вокруг этой части появится серая рамка.

Щелкните правой кнопкой мыши на выделенной области распознавания и в контекстном меню выберите команду **Анализ области**. Программа проана-

лизирует выделенную часть изображения и поделит ее на области по собственному усмотрению.

Свойства области

В нижней части окна **Изображение** находится панель свойств. Она состоит из двух вкладок: **Свойства области** и **Свойства изображения**. На вкладке **Свойства области** можно изменить любые свойства области, выбранной в настоящий момент.

Чтобы изменить тип выделенной области, нажмите соответствующую кнопку в группе **Тип области**. Набор остальных элементов управления на этой вкладке меняется в зависимости от типа области. На рис. 5.12 показаны элементы управления, доступные для области типа Текст.

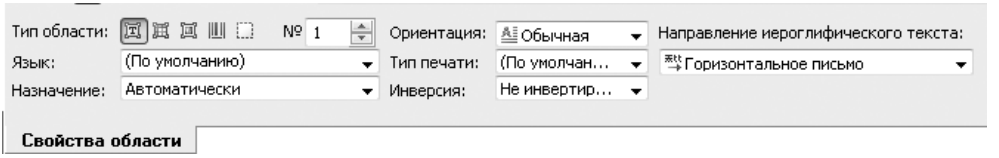


Рис. 5.12 ▾ Панель свойств области

- Чтобы изменить язык распознавания для конкретной области, на изображении щелкните кнопкой мыши на этой области и выберите на панели свойств нужный язык в раскрывающемся списке **Язык:**. Значение **По умолчанию** соответствует тому языку (или тем языкам), который был выбран для всего документа в раскрывающемся списке окна **Страницы**. Если в русском тексте присутствует абзац на другом языке (пример или цитата), выделите этот абзац как отдельную область и выберите для нее этот язык. Так вы уменьшите возможность ошибок при распознавании по сравнению со случаем, когда для документа в целом выбраны два языка.
- В раскрывающемся списке **Назначение:** задаются особенности передачи распознанного содержимого области в выходной документ. Забегая вперед, заметим, что пункты этого раскрывающегося списка связаны со *стилями*: в зависимости от того, какой пункт выбран, в выходном документе к этому фрагменту будет применен соответствующий стиль оформления.
- В раскрывающемся списке **Ориентация** выбирается направление текста. В некоторых документах присутствуют фрагменты, набранные под углом 90° по отношению к основному тексту, например подписи к схемам. Для области, которая содержит такой фрагмент, выберите значение **Вертикальная (сверху вниз)** или **Вертикальная (снизу вверх)**.
- **Тип печати** – параметр, уже знакомый нам по вкладке **Документ** диалога **Опции**. В данном случае вы можете задать для определенных областей тип печати, отличающийся от такового для документа в целом. Характер-

ный пример – типографский бланк, в котором отдельные поля заполнены на пишущей машинке.

- В раскрывающемся списке **Инверсия** выбирается соотношение цвета текста и фона. Обычное значение – **Авто** или **Не инвертирован**. Для областей, в которых текст напечатан светлым шрифтом на темном фоне, выберите значение **Инвертирован**. Например, такие области есть в документах на рис. 5.1, 5.18.
- Выбор одного из вариантов из раскрывающегося списка **Направление иероглифического текста** влияет на распознавание текста, набранного иероглифами на китайском, корейском или японском языке.

Программа FineReader почти всегда сама правильно определяет и ориентацию, и инверсию для тех областей, где они отличаются от обычных. В отличие от этого, язык и *тип печати*, отличающиеся от принятых по умолчанию для всего документа, обычно приходится задавать вручную.

Десятками лет в делопроизводстве заполняли на пишущей машинке бланки, отпечатанные типографским способом. Так велись личные дела, учетные карточки, составлялись протоколы и другие документы, а в правоохранительных органах и армии бланки иногда заполняют подобным образом и в наши дни. При «оцифровке» старых архивных материалов часто возникает необходимость распознавать оригиналы, в которых чередуются типографский и машинописный шрифт (рис. 5.13).

ПРОТОКОЛ
допроса потерпевшего

г.Саратов «10» августа 1975 г.
(место составления)

Допрос начат в 15 ч 20 мин
Допрос окончен в 16 ч 10 мин

следователь Кировского РОВД г. Саратова
(должность следователя (дознавателя),
ст. лейтенант милиции Степанов А.И.
классный чин или звание, фамилия, инициалы)
в помещении хирургического отделения 3 городской клинической больницы
(каком именно)
в соответствии со ст. 189 и 190 (191) УПК РФ допросил по уголовному делу № 125081
в качестве потерпевшего:

Рис. 5.13 ▾ Бланк, заполненный на пишущей машинке

Для наилучшего распознавания подобного документа вручную выделите в нем области типа Текст. Постарайтесь, чтобы в каждую область попадал только типографский или только машинописный текст.

По очереди выберите все области с машинописными фрагментами и на панели свойств укажите для этих областей тип печати **Пишущая машинка** (рис. 5.14). Для областей, которые содержат типографский текст, оставьте тип печати **По умолчанию** или **Авто**.

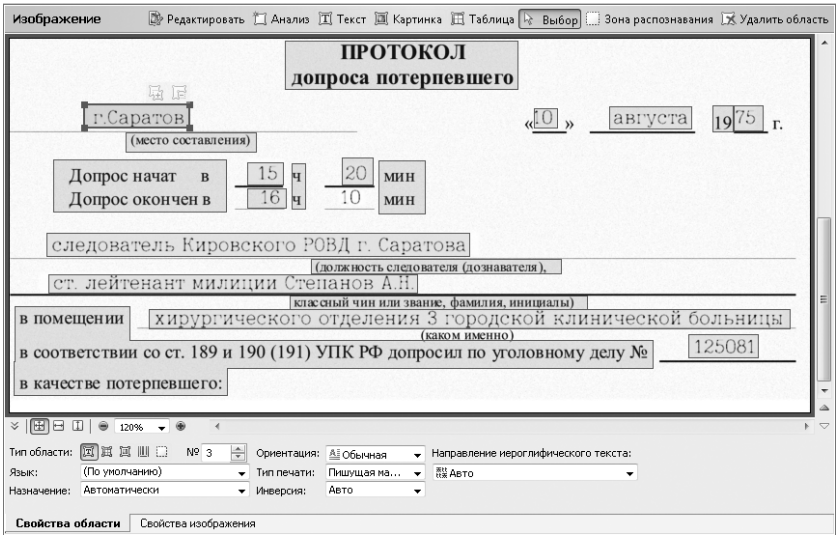


Рис. 5.14 ▾ Разметка областей на заполненном бланке

Подобная разметка на области вручную и уточнение свойств каждой области позволяет ощутимо поднять качество распознавания, хотя занятие это довольно кропотливое. Возможно, ради распознавания единственного документа заниматься этим и не стоит – быстрее будет распознать документ «на полном автомате», а потом выправить ошибки в окне **Текст**.

Однако при сканировании целой пачки однотипных бланков тщательная разбивка на области первого документа поможет автоматизировать и улучшить процесс обработки следующих документов. Для этого служат шаблоны областей.

Использование шаблонов областей

Предположим, вам нужно регулярно сканировать и распознавать какие-то однотипные по содержанию и оформлению документы. Такая задача часто встает при переводе в электронный вид архивных материалов. Другой типичный случай – использование программы FineReader для ввода в компьютер данных со стандартных печатных форм, например квитанций, платежных поручений и т. п. При автоматизации ввода серии однотипных оригиналов иногда достаточно распознавать из целого документа лишь определенную часть – например, нет необходимости каждый раз распознавать «шапку» или стандартные примечания. Во всех этих случаях использование шаблонов областей ускоряет и упрощает работу.

1. Отключите автоматический анализ и распознавание. В диалоговом окне **Опции** на вкладке **Сканировать/Открыть** установите переключатель в положение **Отключить автоматический анализ и распознавание изображения**.

2. Отсканируйте оригинал, который будет взят за образец при создании шаблона областей.
3. В окне **Изображение** разметьте области. Внимательно уточните размер и положение, свойства каждой области.
4. В меню **Области** выберите команду **Сохранить шаблон областей**. Откроется диалог сохранения файла. Выберите в нем папку и введите имя сохраняемого шаблона. Нажмите кнопку **Сохранить**.

Шаблон областей сохраняется в файл с расширением **BLK**. Этот файл содержит сведения о положении каждой области относительно левого верхнего угла изображения, а также о свойствах каждой области. Шаблон применяется к одной странице.

Используйте сохраненный шаблон областей при распознавании последующих документов. Это позволит избежать проблем распознавания, вызванных ошибками при автоматической разметке областей.

1. Отсканируйте оригиналы.
2. В окне **Страницы** выделите страницы, к которым надо применить шаблон. Для этого нажмите клавишу **Ctrl** и, удерживая ее, щелкайте кнопкой мыши на эскизах страниц.
3. В меню **Области** выберите команду **Загрузить шаблон областей**. В диалоге **Открыть шаблон областей** выберите файл шаблона (рис. 5.15).
4. Чтобы применить шаблон только к выбранным страницам, установите переключатель **Применить к:** в положение **Выделенным страницам**. Чтобы использовать шаблон для всех страниц документа, установите пе-

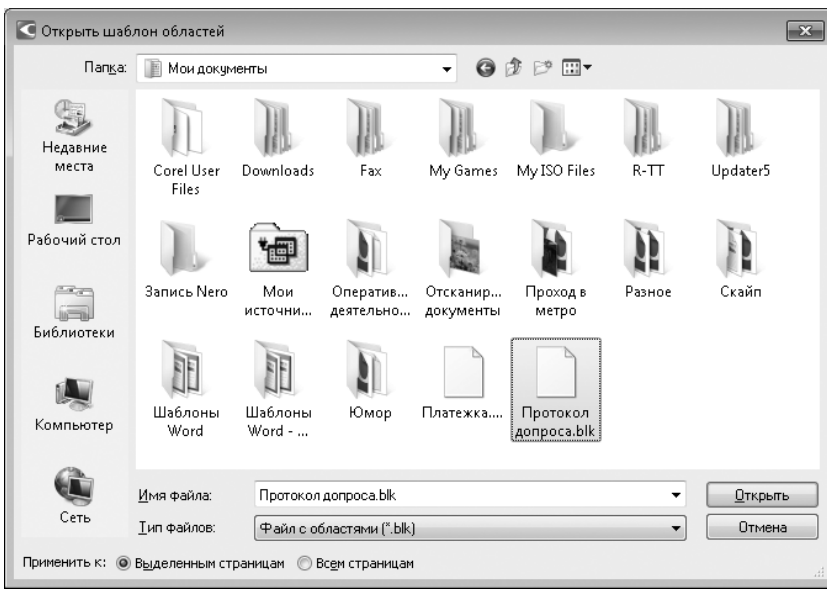


Рис. 5.15 ▾ Диалог открытия шаблона

реключатель в положение **Всем страницам**. Нажмите кнопку **Открыть**. Изображение будет разбито на области в соответствии с сохраненным шаблоном.

Области следует размечать так, чтобы они покрывали всю ширину поля на бланке: в одном документе в это поле может быть вписано два слова, а в другом – целых десять. Высоту областей также стоит делать максимально возможной – при заполнении бланков отдельные строки могли слегка «гулять» по высоте.

При сканировании каждый документ необходимо очень точно и совершенно одинаково укладывать на стекло сканера. В прицеливании помогут линейки планшета. Только при этом условии текст на очередном изображении каждый раз будет попадать в соответствующие области шаблона.

ПРИМЕЧАНИЕ

При работе с фотографиями документов шаблоны областей использовать, скорее всего, не удастся. Проблема в том, что несколько снимков практически невозможно кадрировать одинаково, с точностью до миллиметра.

Иногда из документов достаточно распознавать только отдельные фрагменты. Например, нужно каждый день заносить в базу данных сведения из нескольких десятков платежных поручений. Эту процедуру можно частично автоматизировать: сканировать и распознавать «платежки», а потом копировать данные из распознанного документа и вставлять их в поля базы данных.

Очевидно, при этом из всего платежного поручения надо извлекать только номер, дату, реквизиты плательщика и сумму. Поэтому при создании шаблона областей достаточно обозначить лишь те участки платежного поручения, где находятся эти сведения (рис. 5.16).

Более того, если какая-то из областей может содержать только дату и ничего, кроме даты, выберите для нее особый язык. То, как создать пользователь-


СБЕРБАНК РОССИИ
Основан в 1841 году
Форма № ПД-4сб (налог)
УФК по г. Москве (НУК ИУ МГТУ им. Н.Э.Баумана, л/с 06073441620)
КПП 770102011
7701002520
(наименование получателя платежа)
45286555000
ИНН налогового органа*
и его сокращенное наименование
(Код ОКАТО)
40503810600001009079
(номер счета получателя платежа)
в Отд. №1 Московского ГТУ
(наименование банка)
БИК: 044583001
Кор./сч.: Банка России г. Москва 705
П.1, р.2 от 01.04.2005
Оргвзнос за участие в межд. симпозиуме "Интеллектуальные системы-2008"
07302010100100000130
(код бюджетной классификации КБК)
Плательщик (Ф.И.О.) Иванов Иван Иванович
Адрес плательщика: Москва, Ленинский пр-т, д. 123, кв. 456
ИНН плательщика:
№ л/с плательщика
Сумма: 767 руб. 00 коп. В том числе НДС 18%
Плательщик (подпись): Иванов
Дата: 10 марта 200 8 г.
* или иной государственный орган исполнительной власти

Рис. 5.16 ▾ Пример разметки областей на платежном поручении

ский язык «Date» для дат вида «Число.Месяц.Год», рассмотрено в следующей главе. Точно так же для областей, в которых должны находиться исключительно числовые данные, выберите язык «Цифры». Так вы еще уменьшите вероятность ошибок при распознавании.

В полной мере возможности автоматизации с использованием шаблонов областей и пользовательских языков раскрываются при работе со сценариями. Пока мы рассмотрели первый компонент – шаблоны областей. О пользовательских языках сказано в следующей главе. В предпоследней главе книги мы соберем все компоненты воедино и покажем, как превратить программу FineReader в автомат для практически безошибочного распознавания документов определенного рода.

Анализ таблиц

Программа FineReader способна автоматически распознавать на изображении данные, оформленные в виде таблицы. Лучше всего программа воспринимает таблицы с явно прорисованной сеткой и темным текстом на светлом фоне во всех ячейках – обычное оформление таблиц в книгах или официальных документах. При автоматическом анализе изображения программа находит такую таблицу, обозначает ее как область типа Таблица и правильно разбивает на ячейки (рис. 5.17). Это самый простой случай, автоматический анализ и распознавание подобных таблиц почти всегда проходят успешно.

Таблицы, в которых есть объединенные ячейки, или в части ячеек содержится светлый текст на темном фоне, – более сложный случай. При анализе таких таблиц программа может совершить ошибки. Например, на рис. 5.18 мы

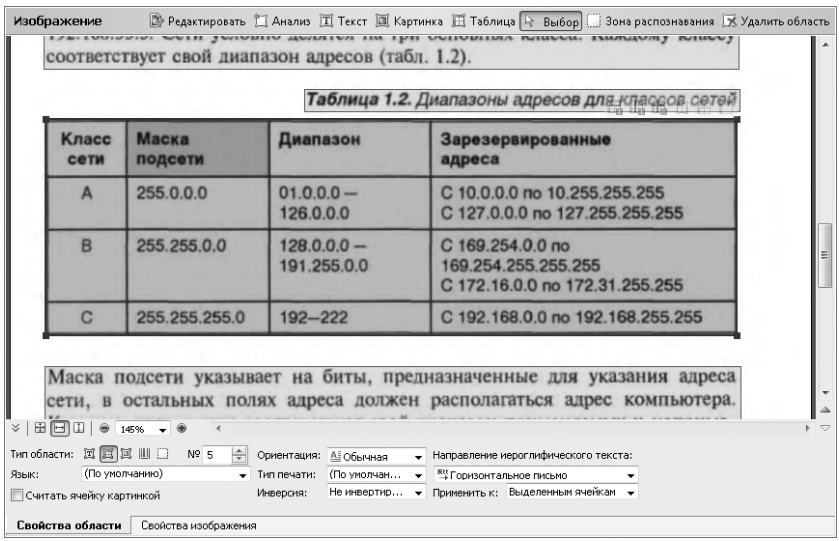


Рис. 5.17 ▾ Автоматический анализ простой таблицы прошел успешно

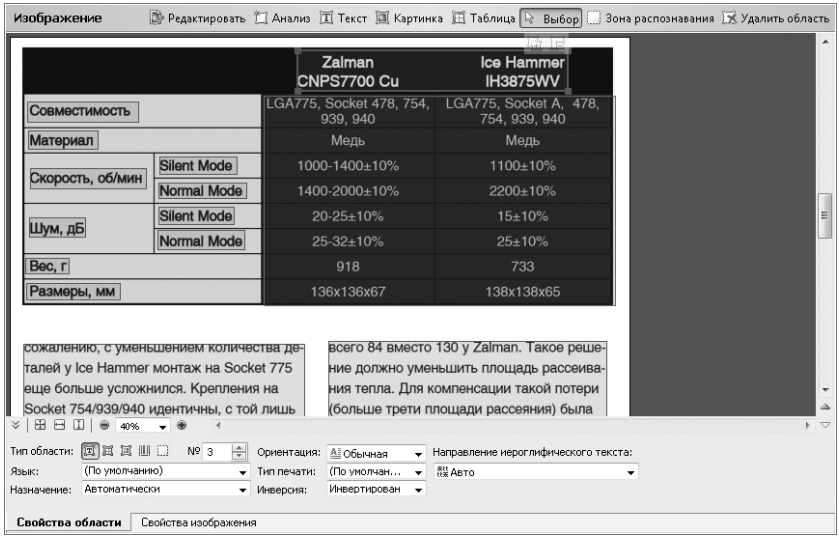


Рис. 5.18 ▾ Анализ сложной таблицы выполнен с ошибками

видим, что при автоматическом анализе программа посчитала часть ячеек текстовыми фрагментами, а часть расценила как область типа Картинка.

В такой ситуации разбивку таблицы для последующего распознавания нужно выполнить вручную. При этом стоит дать программе «вторую попытку» – выделите всю таблицу, а затем выполните автоматический анализ структуры этой области.

1. Предварительно удалите все области, которые программа обозначила на месте таблицы.
2. Нажмите на панели инструментов кнопку **Таблица** и мышью выделите всю область, в которой находится таблица. На изображении появится область типа Таблица, обозначенная синей рамкой.
3. Выполните автоматический анализ структуры таблицы. Щелкните на области типа Таблица правой кнопкой мыши и в контекстном меню выберите команду **Анализ структуры таблицы** (рис. 5.19). Область будет разбита тонкими синими линиями на отдельные ячейки.
   Как правило, внутри уже заданной вручную области типа Таблица программа корректно определяет строки, столбцы и ячейки. Чтобы убедиться, что это так, щелкните кнопкой мыши на любой из ячеек таблицы. Ячейка внутри области будет выделена – подсвечена более ярким синим цветом.

На панели свойств области установите переключатель **Применить к:** в положение **Выделенным ячейкам**. В этом случае все действия, выполняемые на панели свойств области, будут относиться только к выделенной ячейке.

Выделяя по очереди размеченные программой ячейки, убедитесь, что они совпадают с ячейками таблицы на изображении и что им верно присвоены та-

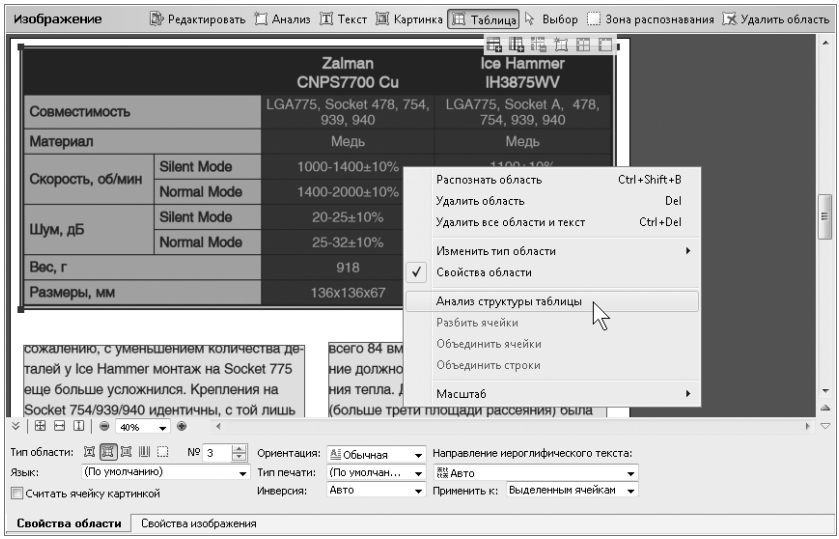


Рис. 5.19 ▾ Анализ структуры таблицы, обозначенной вручную

кие свойства, как ориентация, тип печати и инверсия. В примере на рис. 5.18–5.19 ячейки верхних двух рядов и двух правых колонок содержат светлый текст на темном фоне.

- Для таких ячеек в раскрывающемся списке **Инверсия** следует выбрать значение **Инвертирован**.
- Когда в ячейке текст расположен вертикально, выберите в раскрывающемся списке **Ориентация** правильное направление текста.
- Если ячейка содержит рисунок или формулу, установите флажок **Считать ячейку картинкой**. Содержимое этой ячейки будет передано в выходной документ без распознавания, как рисунок.

Если программа не смогла правильно выявить все ячейки на изображении таблицы, откорректируйте разбивку вручную. При необходимости перетащите мышью автоматически намеченные разделительные линии так, чтобы они совпали с сеткой таблицы.

При необходимости разделительные линии можно добавлять или удалять вручную. Щелкните на области, которую вы хотите изменить. Над левым верхним углом области появится полупрозрачная панель с кнопками-пиктограммами (рис. 5.20).

- Чтобы добавить вертикальную линию, разделяющую столбцы, нажмите кнопку **Вертикальный разделитель** и проведите линию внутри области.

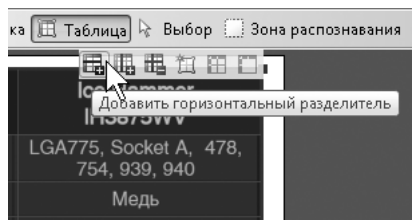


Рис. 5.20 ▾ Всплывающая панель области типа **Таблица**

- Для добавления горизонтального разделителя между рядами (строками) таблицы нажмите кнопку **Горизонтальный разделитель** и нарисуйте дополнительную линию внутри области.
- Для удаления любой разделительной линии нажмите кнопку **Удалить разделитель**. Указатель мыши примет вид косого крестика. Щелкните кнопкой мыши на линии, которую надо удалить.

ПРИМЕЧАНИЕ

Три названные функции похожи на те, которые мы рассмотрели в редакторе изображений. Разница в том, что в редакторе изображений наносимые линии делят все изображение на страницы, а в данном случае они делят область типа Таблица на отдельные ячейки.

- Кнопка **Анализ структуры таблицы** на всплывающей панели инструментов позволяет выполнить автоматический анализ структуры таблицы.

Кроме того, ячейки таблицы можно разделять или объединять. Две кнопки в правой части всплывающей панели инструментов становятся активны, когда выделены ячейки, к которым могут быть применены эти действия.

- Чтобы объединить две или более ячеек в одну, выделите мышью эти ячейки, а затем нажмите кнопку **Объединить ячейки**.
- Чтобы разбить выделенную ячейку на две, нажмите кнопку **Разбить ячейки**. Эту операцию можно применить только к ранее объединенным ячейкам.

На аккуратность автоматического анализа и распознавания таблиц влияет настройка, задаваемая на вкладке **Распознать** диалогового окна **Опции** (см. рис. 6.6). По умолчанию переключатель **Режим распознавание** установлен в положение **Тщательное распознавание**.

- Когда переключатель **Режим распознавание** установлен в положение **Тщательное распознавание**, качество распознавания таблиц, в том числе таблиц без линий сетки или с цветными ячейками, выше.
- Когда переключатель **Режим распознавание** установлен в положение **Быстрое распознавание**, анализ и распознавание документов выполняются несколько быстрее. Однако в таком режиме возрастает вероятность ошибок при работе с документами, содержащими сложное форматирование.

При распознавании документов, содержащих таблицы, постарайтесь тщательно разметить таблицу на этапе анализа. Как правило, это гораздо проще, чем впоследствии править неправильно распознанный текст.

Резюме

Обработка изображений предшествует их анализу и распознаванию. В ходе обработки устраняются некоторые проблемы, возникшие при получении изоб-

ражений, например перекос и искажение прямолинейности строк. Вместе с тем при любой возможности лучше сразу исключать такие проблемы еще при сканировании или фотографировании оригиналов.

От удачной разбивки изображения на области зависит, насколько успешно будет распознано содержимое документа. Программа FineReader почти всегда правильно анализирует изображения и выделяет на них связный текст, таблицы и иллюстрации. Коррективы в автоматическую разбивку стоит вносить лишь при необходимости.

Шаблон областей – сохраненная схема деления изображения на области. Впоследствии шаблон загружается в программу и применяется к изображениям других оригиналов с таким же расположением текста, таблиц и иллюстраций. Использование шаблонов позволяет ускорить и автоматизировать процесс распознавания документов на стандартных бланках.

Глава 6 Распознавание текстов

В последней версии программы реализованы широкие функциональные возможности по распознаванию текстов разных уровней сложности. В этой главе мы узнаем, каким образом можно «научить» FineReader распознавать незнакомые символы или некачественные тексты, создавать пользовательские языки и работать со словарями.

Применение пользовательского эталона

Далеко не всегда процесс распознавания текстов проходит гладко. Особенно это касается документов, которые выполнены декоративным или другим нетрадиционным шрифтом, содержат специфические символы (например, в формулах) или просто плохого качества. В таких случаях FineReader сталкивается с затруднениями.

Общие правила работы с пользовательскими эталонами

В программе реализован механизм, который позволяет объяснить ей, как надо распознавать нестандартные или некачественные тексты. Для этого нужно создать и обучить специальный пользовательский эталон, в котором для каждого сомнительного или непонятного символа определяется его усредненное точечное изображение и название.

ВНИМАНИЕ

Пользовательский эталон создается на начальном этапе распознавания документа и в дальнейшем используется для распознавания основного объема текста. Созданный эталон можно сохранить и впоследствии использовать для работы с другими документами. Кроме этого, в программе имеются также встроенные эталоны.

Далее перечислим несколько правил, которые следует учитывать при распознавании текстов с помощью пользовательских эталонов.

Созданные эталоны впоследствии можно применять для распознавания только тех документов, в которых разрешение, шрифт и его размер совпадают с документом, на основании которого данный эталон был создан. При несоблюдении этого правила результат распознавания может быть непредсказуемым.

Программа не различает некоторые символы и сопоставляет их с каким-то одним символом. Характерный пример – апострофы: правый (`) и левый (´) в программе не идентифицируются и ассоциируются с прямым апострофом ('). Поэтому в распознанном документе никогда не отобразится ни правый, ни левый апостроф: вместо них будет вставлен прямой, причем даже в том случае, когда в процессе обучения эталона были указаны именно они.

Применять пользовательские эталоны для распознавания имеет смысл лишь тогда, когда документ содержит декоративные или нестандартные символы либо когда нужно распознать большое количество текста плохого качества. В других случаях это может оказаться нецелесообразно.

В некоторых случаях программа FineReader делает вывод о сопоставлении изображения тому или иному символу на основании общего анализа текста. В частности, так она может определить, какому символу сопоставить изображение «кружок» – нулю или букве «о», исходя из того, какие символы находятся поблизости (цифры или буквы).

Чтобы войти в режим работы с эталонами, выполните в главном меню команду **Сервис ➢ Редактор эталонов** либо нажмите комбинацию клавиш **Ctrl+Shift+A**. В результате на экране откроется окно, изображенное на рис. 6.1.

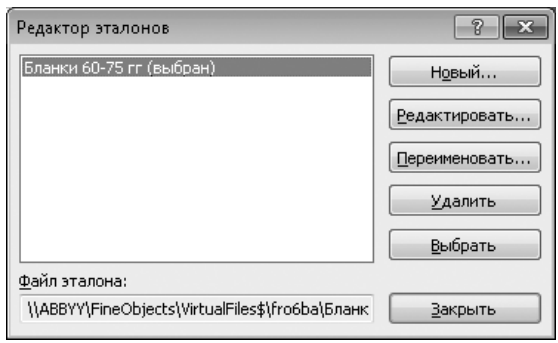


Рис. 6.1 ▾ Редактор эталонов

В данном окне представлен перечень имеющихся в программе эталонов. С помощью кнопки **Выбрать** осуществляется выбор эталона для распознавания текущего документа, который соответствующим образом помечается в списке.

Чтобы добавить в программу новый эталон, нажмите кнопку **Новый**. В результате откроется окно, которое показано на рис. 6.2.

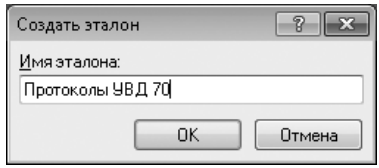


Рис. 6.2 ▾ Ввод нового эталона

В данном окне нужно с клавиатуры ввести произвольное имя эталона и нажать кнопку **ОК**. После этого вновь созданный эталон отобразится в окне редактора (см. рис. 6.1).

С помощью кнопки **Редактировать** осуществляется переход в режим редактирования эталона. Отметим, что для новых эталонов это не имеет смысла: вначале нужно обучить эталон на основании какого-то документа, и лишь после этого его можно будет как-то изменять. Поэтому более подробно порядок редактирования эталонов мы рассмотрим ниже – после того, как освоим технику их обучения и использования в процессе распознавания.

Если потребуется переименовать какой-то эталон, выделите его в списке щелчком мыши и нажмите кнопку **Переименовать**, после чего в открывшемся окне (см. рис. 6.2) введите требуемое имя и нажмите **ОК**.

Для удаления ненужных эталонов используйте кнопку **Удалить**, после чего утвердительно ответьте на появившийся запрос программы.

Пример обучения и использования эталона

Далее на конкретном примере рассмотрим порядок обучения и применения пользовательских эталонов.

Предположим, что нам нужно распознать документ, фрагмент которого показан на рис. 6.3.

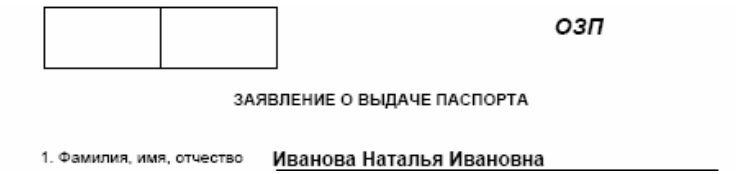


Рис. 6.3 ▾ Фрагмент документа для распознавания

Вначале попробуем распознать его обычным способом – без применения пользовательских эталонов. Для этого выполним команду главного меню **Файл ➤ Открыть PDF/изображение** (эта команда вызывается также нажатием комбинации клавиш **Ctrl+O**) и в открывшемся окне укажем путь к требуемому файлу, после чего нажмем кнопку **Открыть**. Начнется процесс распознавания, информация о ходе которого будет отображаться в открывшемся окне (рис. 6.4).

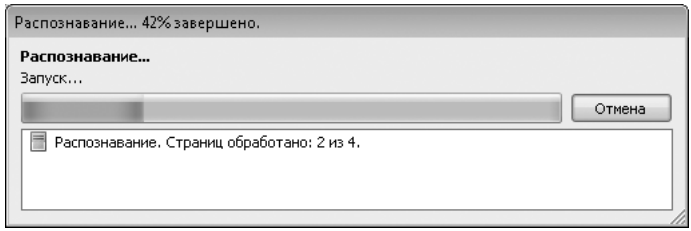


Рис. 6.4 ▾ Процесс распознавания документа

Через какое-то время (в зависимости от скорости работы компьютера) отобразится рабочий интерфейс программы. Результат распознавания будет представлен в окне **Текст** (рис. 6.5).

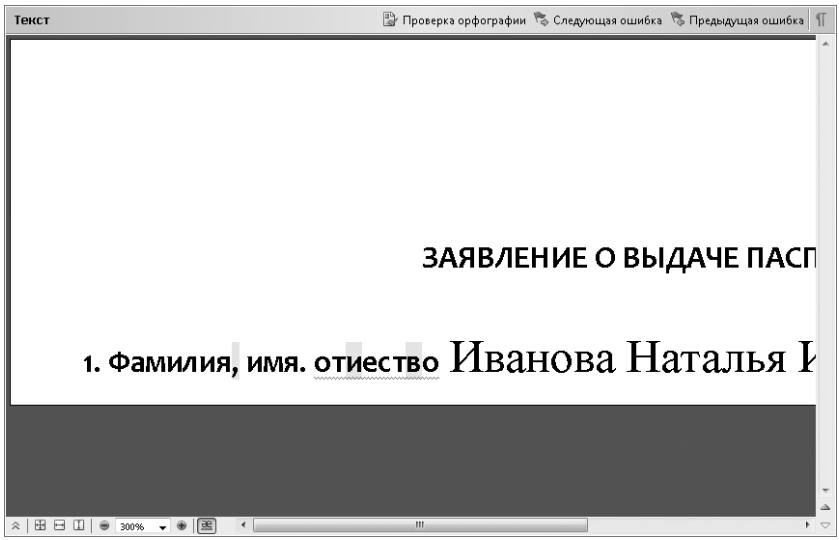


Рис. 6.5 ▾ Результат распознавания документа

Как видно на рисунке, текст документа распознан некорректно: одна из запятых распознана как точка, есть ошибка в слове **Отчество**. Кроме того, программа неуверенно распознала некоторые символы (об этом свидетельствует их цветовое выделение).

Чтобы решить проблему, используем механизм распознавания с обучением. Для этого вначале войдем в режим настройки программы, выполнив в главном меню команду **Сервис ➢ Опции** (эта команда вызывается также нажатием **Ctrl+Shift+O**). В открывшемся окне перейдем на вкладку **Распознать** и установим переключатель **Обучение** в положение **Распознавание с обучением** (рис. 6.6).

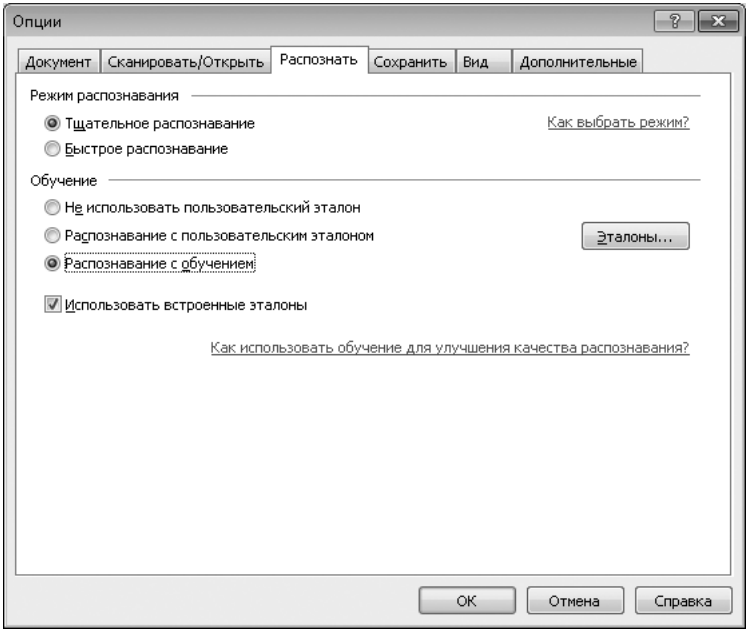


Рис. 6.6 ▾ Включение режима распознавания с обучением

В результате станет доступным флажок **Использовать встроенные эталоны**. Если он установлен (значение по умолчанию), то для распознавания документа будут применяться не только пользовательские, но и встроенные эталоны.

Теперь нужно выбрать эталон для обучения и последующего использования. Для этого нажмем кнопку **Эталоны** – в результате на экране откроется уже знакомое нам окно **Редактор эталонов** (см. рис. 6.1). Чтобы выбрать в нем имеющийся эталон, нужно выделить его щелчком мыши, нажать кнопку **Выбрать**, а затем – кнопку **Закрыть** (причем дважды – в окне редактора эталонов и в режиме настройки).

ПРИМЕЧАНИЕ

*После выбора эталона и нажатия кнопки **Закрыть** в редакторе эталонов кнопка **ОК** в режиме настройки программы также будет называться **Закрыть**.*

Как мы уже отмечали выше, применять уже имеющиеся эталоны для распознавания новых документов можно лишь при соблюдении определенных условий. В противном случае придется создать для обучения новый эталон (см. рис. 6.2). В этом случае он автоматически выбирается для работы с документом.

Теперь возвращаемся в рабочий интерфейс, последовательно закрыв редактор эталонов и окно настройки программы. Запускаем процесс распознавания, нажав на главной панели инструментов кнопку **Распознать** (рис. 6.7).

Рис. 6.7 ▾ Кнопка **Распознать**

Отметим, что также для этого можно воспользоваться соответствующими командами меню **Документ** и **Страница**, расположенными в главном меню программы.

В самом начале распознавания на экране откроется окно, информирующее о ходе процесса (см. рис. 6.4). Когда программа встретит сомнительный символ, отобразится окно **Ручное обучение эталона** (рис. 6.8).

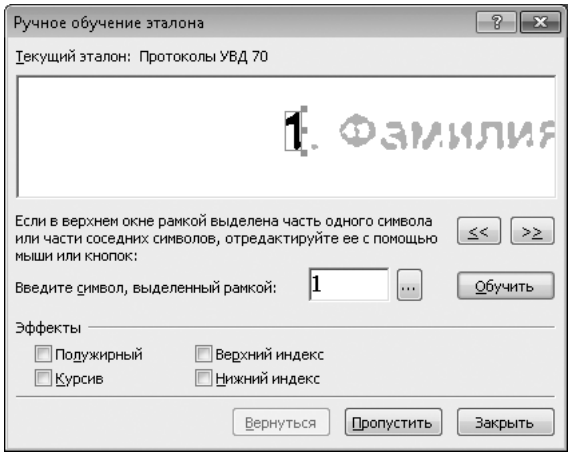


Рис. 6.8 ▾ Режим ручного обучения

В данном окне осуществляется ручное обучение эталона. Смысл данной операции заключается в том, чтобы четко указать программе, каким образом здесь и далее следует распознавать незнакомые ей символы.

На рисунке видно, что первым таким символом в нашем документе является цифра **1**, которую программа не смогла уверенно распознать (см. рис. 6.5). В верхней части окна он выделен рамкой (см. рис. 6.8).

ПРИМЕЧАНИЕ

Иногда бывает так, что рамка выделяет не один, а сразу два незнакомых символа, то есть программа воспринимает их как один, хотя распознать их необходимо по отдельности. Случается и обратное – когда рамка делит символ пополам там, где это не нужно. В подобных ситуациях нужно подкорректировать размеры рамки с помощью расположенных справа кнопок со стрелками или перемещая

границы рамки мышью. Если сочетания двух или трех символов не удается разделить из-за особенностей их начертания, обучите эталон этим комбинациям.

В поле **Введите символ, выделенный рамкой** нужно указать символ, которым должно распознаваться незнакомое программе изображение. Другими словами, вместо неизвестного символа FineReader при распознавании вставит тот, который будет указан в данном поле. Чтобы выбрать требуемый символ, нажмем расположенную справа от поля **Введите символ, выделенный рамкой** кнопку выбора – в результате на экране откроется окно, изображенное на рис. 6.9.

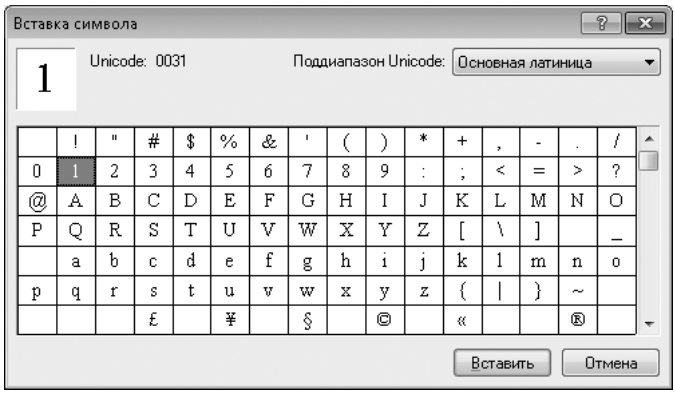


Рис. 6.9 ▾ Выбор символа для распознавания

В данном окне содержится библиотека символов, которые можно использовать в процессе распознавания документов. В нашем случае следует выбрать цифру **1**; для этого выделим ее щелчком мыши и нажмем кнопку **Вставить**. В результате она отобразится в поле **Введите символ, выделенный рамкой** (см. рис. 6.8), чтобы зафиксировать в эталоне замену, нажмем кнопку **Обучить**.

СОВЕТ

При обучении шаблона вы можете назначать символам дополнительные эффекты: включать полужирное или курсивное начертание, а также применять верхний или нижний индекс. Для этого достаточно установить соответствующие флажки, расположенные внизу окна в области ***Эффекты*** *(см. рис. 6.8).*

Сразу после этого рамка в верхней части окна **Ручное обучение эталона** автоматически переместится на следующий непонятный программе символ. В нашем случае это буква **Ф**, являющаяся первой в слове **Фамилия** (рис. 6.10).

Как видно на рисунке, в данном случае программа идентифицирует прописную букву **Ф** со строчной буквой **о**, что является ошибкой. Чтобы исправить ее, введем в поле **Введите символ, выделенный рамкой** прописную букву **Ф** и нажимаем кнопку **Обучить**.

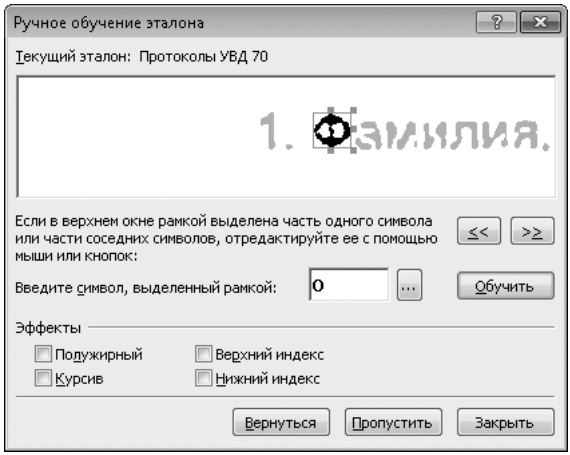


Рис. 6.10 ▾ Следующий этап ручного обучения

Аналогичным образом обучим FineReader распознавать все остальные незнакомые символы в данном документе. Напомним, что по умолчанию в программе все неправильно или неуверенно распознанные символы выделяются цветом, но этот режим можно отключить с помощью команды главного меню **Вид ➢ Окно Изображение/Текст ➢ Выделять неуверенно распознанные символы**. Если какой-то символ, в правильности которого FineReader сомневается, распознан все же верно – при обучении его можно проигнорировать, нажав кнопку **Пропустить** (см. рис. 6.10). Для возврата к предыдущему символу используйте кнопку **Вернуться**.

ВНИМАНИЕ

*При нажатии кнопки **Вернуться** рамка переместится на предыдущую позицию, при этом последняя настроенная аналогия (то есть пара «изображение – символ») будет автоматически удалена из данного шаблона. Помните, что данная кнопка функционирует лишь в пределах одного слова.*

После того как обучение эталона завершено, нажмите кнопку **Закрыть**. При этом программа выдаст запрос относительно сохранения выполненных в эталоне изменений (рис. 6.11).

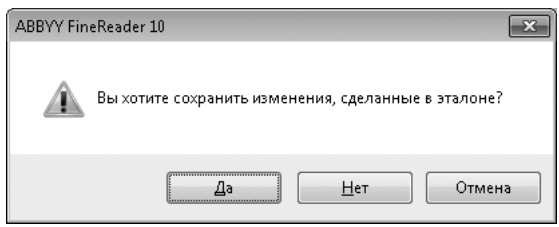


Рис. 6.11 ▾ Запрос на сохранение изменений в эталоне

Чтобы впоследствии применять данный пользовательский эталон с учетом последних изменений, нажмите кнопку **Да**. При нажатии **Нет** изменения будут утрачены, но и в первом, и во втором случае начнется процесс распознавания документа с применением эталона. Чтобы отказаться от немедленного распознавания и возврата в режим обучения, нажмите кнопку **Отмена**.

Если все сделано правильно, то в результате распознавания текста с применением пользовательского эталона наш фрагмент документа будет выглядеть так, как показано на рис. 6.12.

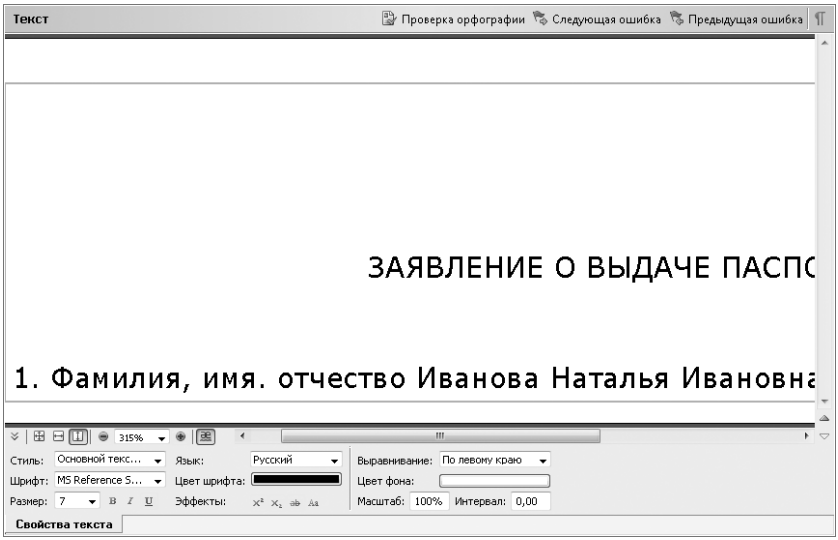


Рис. 6.12 ▾ Результат корректного распознавания текста

После распознавания документа тексту можно придать более эргономичный вид, используя для этого инструменты форматирования и оформления, знакомые каждому пользователю редактора Microsoft Word. Чтобы панель инструментов форматирования отобразилась в нижней части окна **Текст** (рис. 6.12), нажмите кнопку **Показать свойства текста**, расположенную в нижнем левом углу этого окна.

Чтобы впоследствии применить обученный пользовательский эталон для распознавания другого документа, нужно в режиме настройки (см. рис. 6.6) установить переключатель в положение **Распознавание с пользовательским эталоном**, после чего выбрать его в редакторе эталонов (см. рис. 6.1).

Редактирование пользовательских эталонов

Как мы уже отмечали ранее, любой созданный пользователем эталон впоследствии можно отредактировать. Для этого в редакторе эталонов (см. рис. 6.1) нужно выделить его щелчком мыши и нажать кнопку **Редактировать**. В результате на экране откроется окно, изображенное на рис. 6.13.

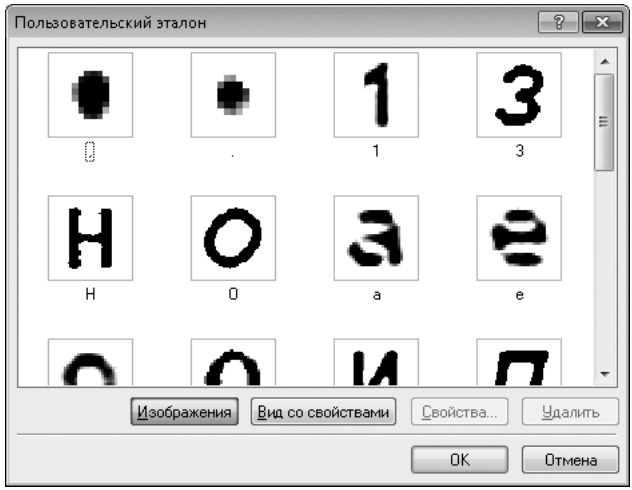


Рис. 6.13 ▾ Редактирование эталона, режим **Изображения**

Окно **Символы пользовательского эталона** имеет два представления: **Изображения** (см. рис. 6.13) и **Вид со свойствами** (рис. 6.14). Переключение между ними осуществляется с помощью соответствующих кнопок, расположенных внизу окна.

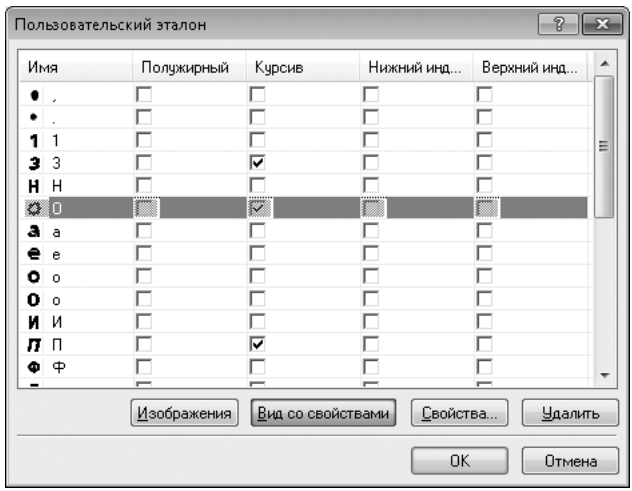


Рис. 6.14 ▾ Редактирование эталона, режим **Вид со свойствами**

И в первом, и во втором случае в окне представлен перечень аналогий, назначенных данному эталону в процессе обучения. В режиме **Изображения** эти аналогии можно только просматривать или удалить, а в режиме **Вид со свой-**

ствами можно также редактировать их свойства путем установки или снятия соответствующих флажков. Отметим, что просматривать и редактировать свойства выбранных позиций можно также в окне (рис. 6.15), открываемом нажатием кнопки **Свойства**.

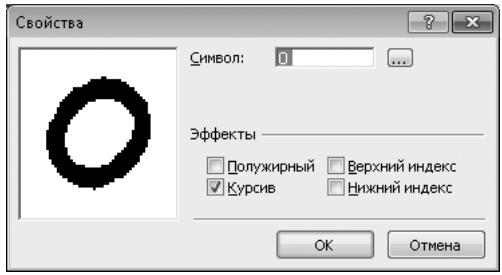


Рис. 6.15 ▾ Окно **Свойства**

В данном режиме можно выполнять те же действия, что и при обучении эталона: выбирать символ для замены непонятного или нестандартного изображения, а также применять к нему эффекты. Выполненные изменения вступают в силу после нажатия кнопки **ОК**.

Чтобы удалить ненужный элемент из эталона, выделите его щелчком мыши (это можно делать как в режиме **Изображения**, так и **Вид со свойствами**) и нажмите кнопку **Удалить**. При этом программа выдаст дополнительный запрос на подтверждение данной операции.

ПРИМЕЧАНИЕ

Обратите внимание – добавлять новые символы в пользовательский эталон в режиме редактирования невозможно. Здесь вы можете лишь просматривать его содержимое и редактировать свойства имеющихся символов. Добавление же новых символов возможно только в процессе обучения эталона.

Отметим, что каждый пользовательский эталон может включать в себя до 1000 символов. При этом вы можете обучать эталон как символам, так и лигатурам (лигатура – сочетание двух или даже трех символов, которые неделимы из-за особенностей начертания и потому назначаются в виде комбинаций; работа с ними ведется так же, как и с отдельными символами). Но слишком увлекаться лигатурами не рекомендуется – может пострадать качество распознавания.

Таким образом, с помощью пользовательских эталонов мы сможем распознать практически любой нестандартный или плохо читаемый документ. Но как быть с документами, составленными не на русском, а на другом языке, либо еще сложнее – содержащими текст сразу на нескольких языках? Ответ на этот вопрос вы найдете в следующем разделе.

Распознавание многоязычных документов

Здесь мы расскажем о том, каким образом в программе FineReader осуществляется распознавание многоязычных документов, а также текстов, составленных с применением редких и нестандартных языков.

Пример распознавания двуязычного документа

Задача распознавания документов, содержащих текст на двух языках, возникает достаточно часто. Например, слова и предложения на иностранных языках, как правило, присутствуют в научно-технической литературе, в прайс-листах или сопроводительной документации к импортным товарам. С распознаванием документов, в которых встречаются слова на русском, английском, немецком, французском и испанском языках, программа FineReader 10 успешно справляется при настройках, принятых по умолчанию.

Настройка языков распознавания может потребоваться, если язык документа не входит в число пяти названных языков. Предположим, что нам нужно распознать документ pdf-формата, составленный на двух языках: русском и литовском (рис. 6.16).

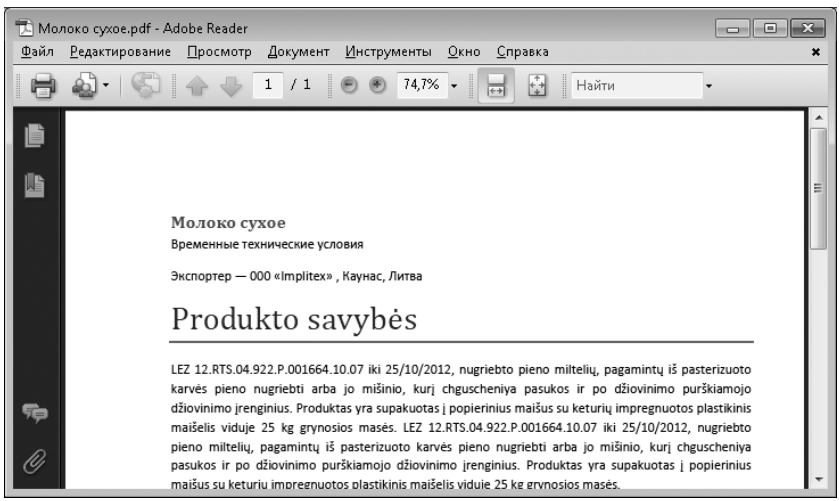


Рис. 6.16 ▾ Документ для распознавания

Попробуем вначале сделать это с автоматическим выбором языка – такова установка, принятая по умолчанию. В раскрывающемся списке выбора языка, находящемся в верхней части окна **Страницы**, в таком случае отображается значение **Автовыбор**. Раскрывающийся список выбора языка присутствует и на панели быстрого доступа (по умолчанию она скрыта).

Выполним распознавание документа, открыв его с помощью команды главного меню **Файл ➢ Открыть PDF/Изображение** либо нажав комбинацию клавиш **Ctrl+O**. Полученный результат вряд ли можно признать удовлетворительным (рис. 6.17).

Рис. 6.17 ▾ Некорректное распознавание двуязычного текста

Как видно на рисунке, FineReader отлично распознал все символы русского алфавита, цифры и знаки препинания. Однако программа не смогла правильно распознать текст на иностранном языке. Это вполне объяснимо: литовский не входит в число языков, которые программа по умолчанию определяет автоматически.

Теперь изменим языковые настройки. В окне **Страницы** в поле **Язык документа** из раскрывающегося списка выберем значение **Выбор языков**. Откроется окно **Редактор языков**. Установите в нем переключатель в положение **Указать языки распознавания вручную** и флажки напротив тех языков, которые используются в распознаваемом документе (рис. 6.18).

Нажмите кнопку **OK**. Окно редактора языков закроется, а в раскрывающихся списках выбора языка в окне **Страницы** и на панели быстрого доступа вы увидите значение **Литовский, Русский**. Вновь запустите распознавание документа, и в результате текст на обоих языках распознается корректно (рис. 6.19).

Положительные изменения налицо: FineReader успешно распознал символы не только русского, но и литовского алфавита. Отсюда вывод: если вы собираетесь распознать текст, составленный на иностранном языке, не забудьте перед распознаванием выбрать этот язык в раскрывающемся списке в окне

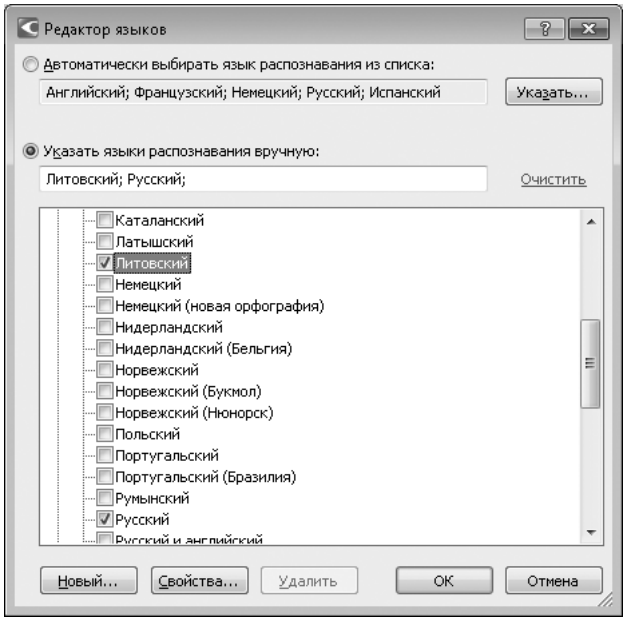


Рис. 6.18 ▾ Редактор языков: выбор языков вручную

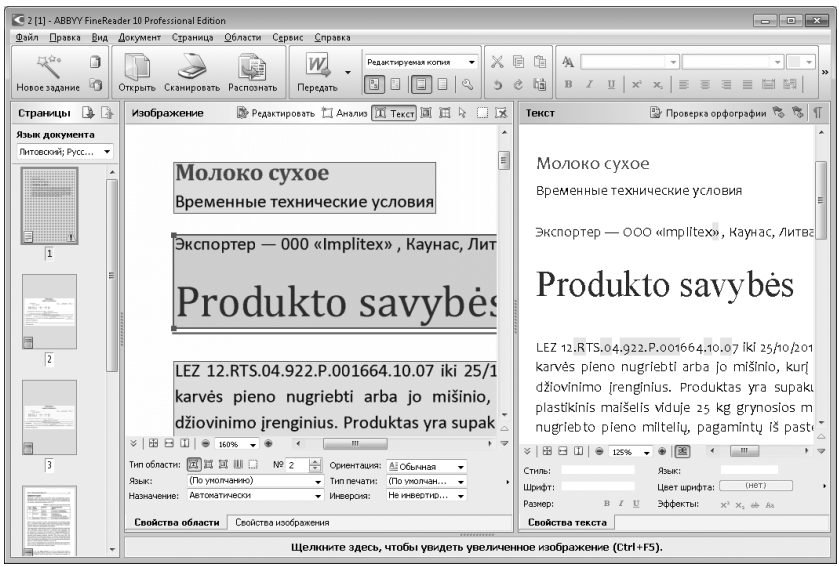


Рис. 6.19 ▾ Корректное распознавание двуязычного документа

Страницы или в окне **Новое задание**, если вы используете для распознавания встроенные сценарии. Если же требуется распознать многоязычный документ – выберите в раскрывающемся списке группу языков, например **Русский и английский**, либо укажите необходимые языки в окне **Редактор языков**.

По умолчанию программа содержит множество языков, а также несколько сформированных языковых групп, в которые включены наиболее часто используемые языки. Но как быть, если необходимо распознать документ, составленный на каком-то редко употребляемом языке?

В таком случае придется самостоятельно выполнить соответствующую настройку программы, сформировав в ней пользовательский язык и, при необходимости, включив его в соответствующую группу. О том, как это делать, читайте далее.

Выбор языка для распознавания документа

Выше мы рассмотрели один из методов выбора языка для распознавания документа – когда в окне **Страницы** (см. рис. 6.17) вначале выбрали значение **Русский**, а затем – **Литовский; Русский**. Но кроме конкретного указания языка для распознавания в данном окне можно выбрать также значения **Авто** или **Выбор языков**.

Чтобы программа автоматически выбирала язык распознавания документа из заранее подготовленного списка, нужно выбрать значение **Авто**. Но предварительно этот список следует подготовить. Необходимые действия выполняются в окне **Редактор языков**, которое можно вызвать одним из трех способов: выбрать в окне **Страницы** из раскрывающегося списка значение **Выбор языков**, выполнить команду главного меню **Сервис ➢ Редактор языков** либо нажать комбинацию клавиш **Ctrl+Shift+L**. В результате любого из перечисленных действий на экране отобразится окно, изображенное на рис. 6.20.

В данном окне нужно переключатель установить в положение **Автоматически выбирать язык распознавания из списка** и нажать расположенную справа кнопку **Указать**. В результате на экране откроется окно, которое показано на рис. 6.21.

В данном окне путем установки соответствующих флажков укажите языки, из числа которых программа должна будет автоматически выбирать язык для распознавания документа, и нажмите кнопку **ОК**.

Вы можете также не предоставлять программе право выбора языка распознавания, а указать один или несколько языков, которые она должна использовать. Для этого в окне **Страницы** в поле **Языки документа** выберите из раскрывающегося списка значение **Выбор языков**. В открывшемся редакторе языков (см. рис. 6.20) установите переключатель в положение **Указать языки распознавания вручную** и в расположенном ниже списке флажками отметьте один или несколько языков, которые программа должна использовать, после чего нажмите **ОК**.

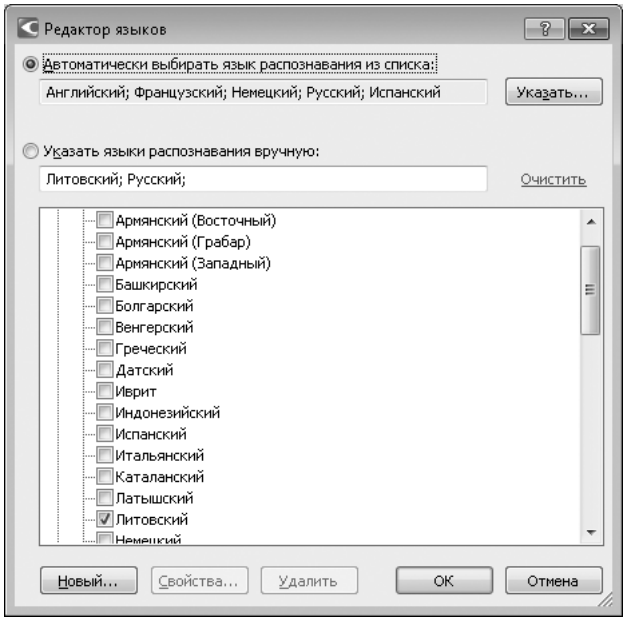


Рис. 6.20 ▾ Редактор языков

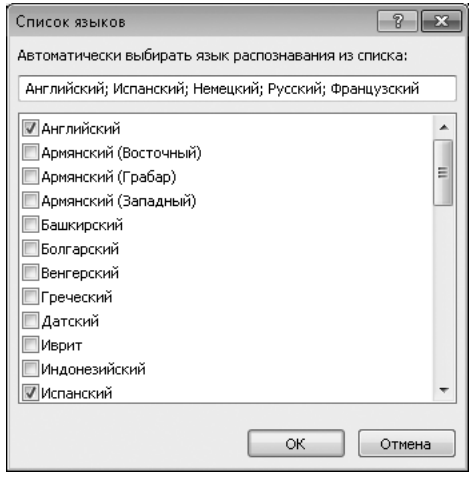


Рис. 6.21 ▾ Выбор языков для включения в список

Создание группы языков

Если вы часто используете для распознавания документов один и тот же набор языков, который каждый раз указываете в режиме ручного выбора, имеет смысл объединить их в группу. Удобство очевидно: это избавит вас от необхо-

димости каждый раз вручную выбирать языки – достаточно будет один раз указать группу, в которую они объединены.

Для решения данной задачи нажмите в редакторе языков (см. рис. 6.20) кнопку **Новый** – в результате на экране откроется окно, изображенное на рис. 6.22.

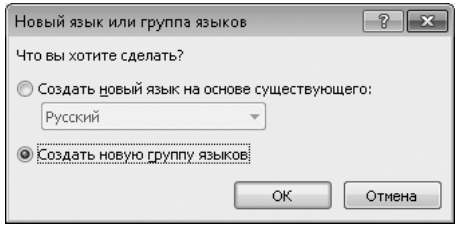


Рис. 6.22 ▾ Выбор режима создания группы

Из этого окна осуществляется переход в режим создания либо группы, либо пользовательского языка. Поскольку нам нужно сформировать группу языков, установим переключатель **Что вы хотите сделать?** в положение **Создать новую группу языков** и нажмем **ОК**. В результате выполненных действий откроется окно, изображенное на рис. 6.23.

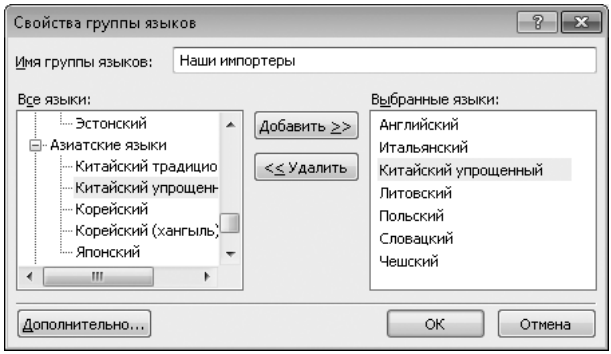


Рис. 6.23 ▾ Формирование группы языков

В первую очередь в поле **Имя группы языков** нужно с клавиатуры ввести имя создаваемой группы, под которым она впоследствии будет отображаться в интерфейсах и списках выбора. После этого в левой части окна в списке **Все языки** нужно выбрать категорию, в которой находится требуемый язык. Отметим, что все языки, с которыми работает FineReader, объединены в следующие категории:

- Языки со словарной поддержкой. Данная категория включает в себя самые употребительные языки: английский, русский, французский, итальянский, немецкий, и т. д. Для этих языков в программе ABBYY FineReader предусмотрена проверка распознанного текста (нахождение неуверенно распознанных слов и слов с орфографическими ошибками).
- Азиатские языки. В данную категорию входят китайский упрощенный, китайский традиционный, корейский и японский языки.
- Дополнительные языки. Здесь содержатся языки, которые используются намного реже: армянский, белорусский, цыганский, ацтекский, дакота

и др. Большинство из них является родными языками народов, общин, племен и т. п.

- Искусственные языки. Данная категория включает в себя несколько искусственно созданных языков, наиболее известным из которых является язык эсперанто.
- Формальные языки. Возможности программы предусматривают распознавание текстов, содержащих простые химические формулы, а также написанных с применением языков программирования (попросту говоря – программных кодов и иных подобных документов): Java, Basic, Pascal, Fortran и др.
- Пользовательские языки. В данную категорию включены языки, созданные пользователями (поэтому эта категория по умолчанию пуста). О том, как это делать, мы расскажем чуть ниже. Также сюда включаются пользовательские группы языков.

Чтобы добавить язык в группу, нужно открыть соответствующую категорию (щелкнув мышью на значке «плюс»), выделить его в списке щелчком мыши и нажать кнопку **Добавить**. Сразу после этого язык отобразится в списке **Выбранные языки**, который находится в правой части окна.

Чтобы удалить язык из группы, нужно выделить его щелчком мыши в списке выбранных языков и нажать кнопку **Удалить**.

При необходимости можно выполнить более тонкую настройку группы языков. Для этого нажмите кнопку **Дополнительно** – в результате откроется окно, которое представлено на рис. 6.24.

ПРИМЕЧАНИЕ

Переход в данный режим возможен лишь при одновременном соблюдении двух условий: группе должно быть присвоено имя, и в списке выбранных языков должна находиться хотя бы одна позиция.

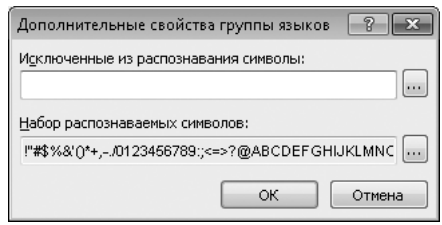


Рис. 6.24 ▾ Дополнительная настройка группы языков

Здесь можно указать символы, которые в процессе распознавания будут игнорироваться. Это бывает целесообразно, например когда из какого-либо языка, включенного в группу, при распознавании используются не все, а лишь несколько символов, а также в иных случаях.

Чтобы указать игнорируемые символы, нажмите кнопку выбора справа от поля **Исключенные из распознавания символы**. При этом откроется окно, изображенное на рис. 6.25.

Работа в данном режиме ведется следующим образом: вначале в поле **Набор символов** следует указать категорию, из которой будут выбираться запрещенные для распознавания символы, а затем щелчком мыши отметить их в открыв-

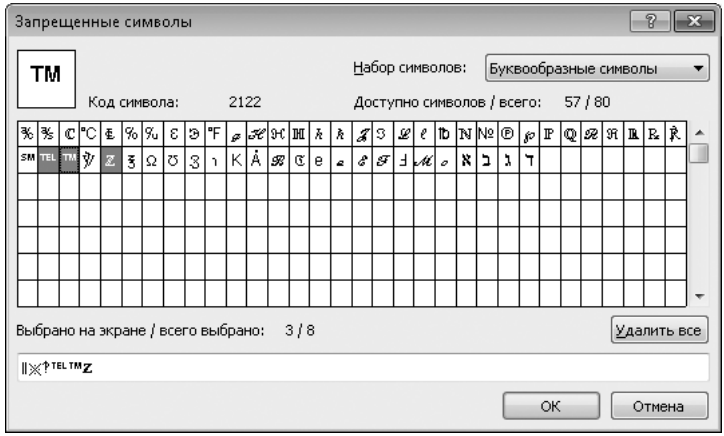


Рис. 6.25 ▾ Выбор запрещенных для распознавания символов

шемся перечне. Выбранные символы автоматически отображаются в нижней части окна. Справа от названия данного поля через разделитель показано число выбранных символов в данной категории и общее число выбранных символов во всех категориях.

Чтобы отказаться от выбора символа, повторно щелкните на нем мышью. Чтобы быстро снять пометки со всех символов, нажмите кнопку **Удалить все**.

Завершается создание группы языков нажатием в окне свойств (см. рис. 6.23) кнопки **ОК**. С помощью кнопки **Отмена** осуществляется выход из данного режима без сохранения выполненных изменений.

Новая группа будет добавлена в редакторе языков в категорию **Пользовательские языки** (рис. 6.26).

Впоследствии вы можете просмотреть и, при необходимости, отредактировать содержимое и свойства созданной группы: для этого выделите ее в редакторе языков щелчком мыши и нажмите кнопку **Свойства**. В результате откроется уже знакомое нам окно **Свойства группы языков** (см. рис. 6.23).

Чтобы удалить пользовательскую группу языков, выделите ее щелчком мыши и нажмите кнопку **Удалить**. При этом программа выдаст дополнительный запрос на подтверждение данной операции.

ПРИМЕЧАНИЕ

*Подобным образом можно удалить только пользовательские группы и языки. В других случаях кнопка **Удалить** будет недоступна.*

Созданную группу языков мы можем увидеть в раскрывающемся списке окна **Страницы** (рис. 6.27).

Подобным образом вы можете создавать любые пользовательские группы языков в зависимости от своих потребностей.

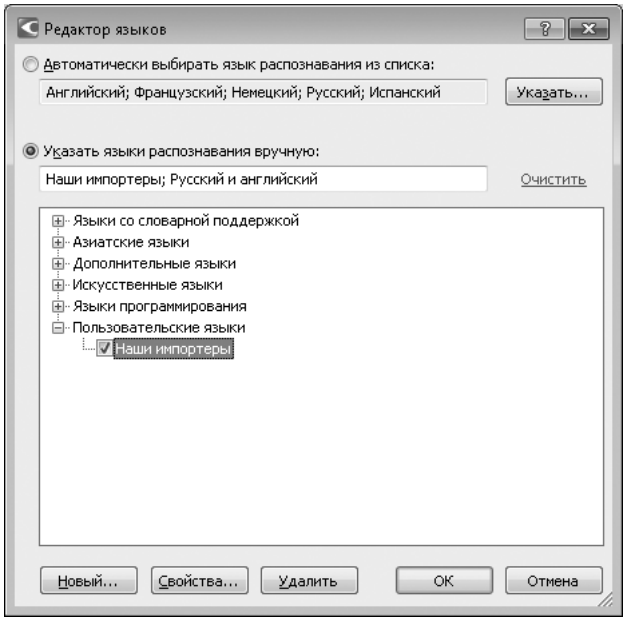


Рис. 6.26 ▾ Новая группа в редакторе языков

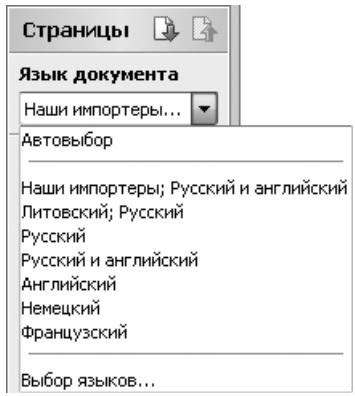


Рис. 6.27 ▾ Выбор пользовательской группы языков

Создание пользовательского языка

Некоторые документы могут оказаться настолько сложными или специфичными, что для их распознавания не хватит ни возможностей пользовательских эталонов, ни имеющихся в программе языков. В частности, этим отличаются тексты, имеющие много неестественных конструкций (например, содержа-

щих артикулы): при их распознавании намного повышается вероятность возникновения ошибок.

Чтобы решить эту проблему, необходимо создать пользовательский язык, который будет применяться при распознавании таких документов, и учитывать все их специфические особенности. Для этого в редакторе языков (см. рис. 6.20) нужно нажать кнопку **Новый**, затем в открывшемся окне **Новый язык или группа языков** нужно установить переключатель **Что вы хотите сделать?** в положение **Создать новый язык на основе существующего** (рис. 6.28).

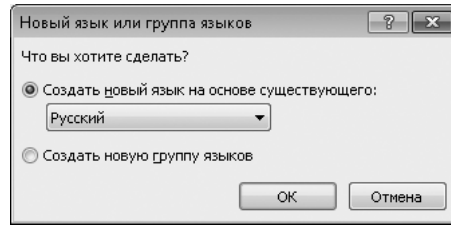


Рис. 6.28 ▾ Создание пользовательского языка на основании существующего

В результате для редактирования откроется расположенный ниже раскрывающийся список; в нем нужно указать язык, на основании которого будет создан новый язык. Например, вы можете создать пользовательский язык на основе языка программирования, либо на основе английского языка и т. д. Мы для примера возьмем в качестве языка-основания русский язык.

После нажатия в данном окне кнопки **ОК** на экране отобразится окно, которое показано на рис. 6.29.

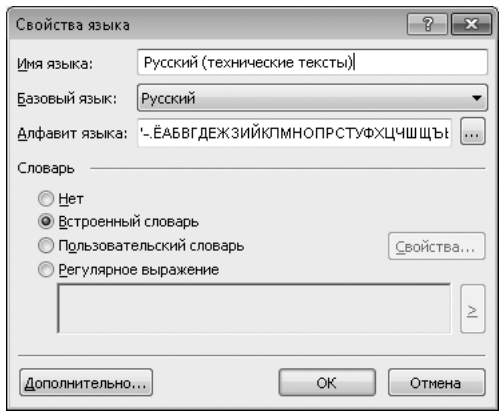


Рис. 6.29 ▾ Окно свойств пользовательского языка

В данном окне осуществляются настройка и определение свойств нового языка. По умолчанию ему присваивается имя языка-основания с добавленным впереди словом **Копия**, но вы можете с клавиатуры ввести имя, указывающее на конкретные особенности создаваемого языка (см. рис. 6.29). В поле **Базовый язык** вы можете выбрать в качестве основания другой язык.

В поле **Алфавит языка** представлен перечень символов алфавита, который будет использоваться в созданном языке. Поскольку в данный момент он полностью совпадает с языком-основанием, его следует отредактировать. Для перехода в режим редактирования нажмите кнопку выбора – в результате откроется окно **Алфавит** (рис. 6.30).

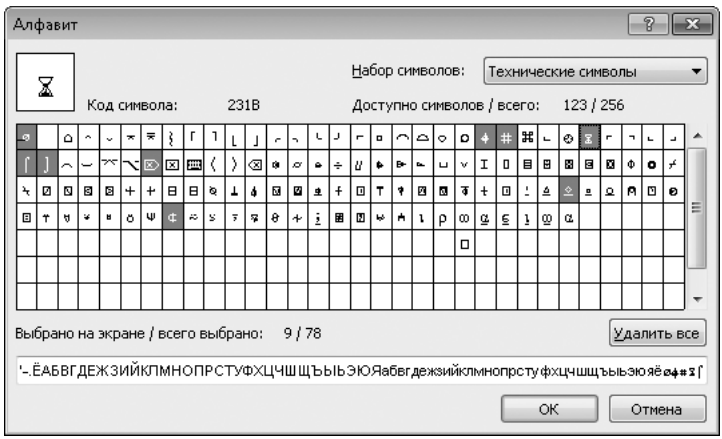


Рис. 6.30 ▾ Настройка алфавита

Принцип работы в данном режиме такой же, как и в окне **Запрещенные символы** (см. рис. 6.25). Мышью помечайте символы, которые нужно включить в новый алфавит, и снимайте пометки с тех, которые нужно из него удалить. Завершив настройку алфавита, нажмите кнопку **ОК**. Окно **Алфавит** закроется, а вы вернетесь к диалогу **Свойства языка** (см. рис. 6.29).

В диалоге **Свойства языка** укажите, как программа должна использовать словари при распознавании текста на этом языке. Если переключатель установить в положение **Нет**, то ни один словарь к данному языку подключен не будет. Если включить положение **Встроенный**, то при распознавании будет применяться словарь, поставляемый с программой. Чтобы создать пользовательский словарь или указать путь к файлу словаря, выберите положение **Пользовательский словарь** и нажмите расположенную справа кнопку **Свойства**. В результате откроется окно настройки и редактирования словарей, с которым мы познакомимся ниже.

Для распознавания текста можно использовать регулярные выражения – для этого нужно установить переключатель в положение **Регулярные выражения** и ввести выражение в расположенном ниже поле. В следующем разделе мы на конкретном примере покажем, как это осуществляется на практике.

Завершается создание пользовательского языка нажатием кнопки **ОК**. После этого новый язык появится в категории **Пользовательские языки** редактора языков (см. рис. 6.26), а также в раскрывающемся списке окна **Страницы** (см. рис. 6.27).

Пример распознавания текста с помощью регулярных выражений

Регулярные выражения представляют собой систему синтаксического разбора текстовых фрагментов, которая базируется на предварительно заданном шаблоне или сформированном наборе правил. Они используются в самых разных направлениях IT-технологий: программирование, веб-проектирование, техническое документирование и т. д., – а также весьма удобны для распознавания формализованных текстов (списки, бланки и т. п.).

Механизм регулярных выражений основан на применении условных обозначений в виде отдельных символов или их наборов. Вот пример простейшего условного выражения: `д.м`. Здесь в качестве условного обозначения выступает символ «точка», находящийся между буквами `д` и `м`. Такое регулярное выражение допускает использование слов «дом», «дым», «дам» и т. д. Вот еще один пример регулярного выражения, в котором используется уже другой разделитель: `л(a|y)па` – оно допускает использование слов «лапа» и «лупа» (здесь символ `|` является разделителем).

ПРИМЕЧАНИЕ

В данном разделе мы не будем приводить описание синтаксиса регулярных выражений – для этого есть специальная литература. Здесь мы на конкретном примере покажем, каким образом их можно использовать для распознавания текстов.

Ввод текста регулярного выражения осуществляется в окне свойств языка в поле, которое открывается для редактирования после того, как переключа-

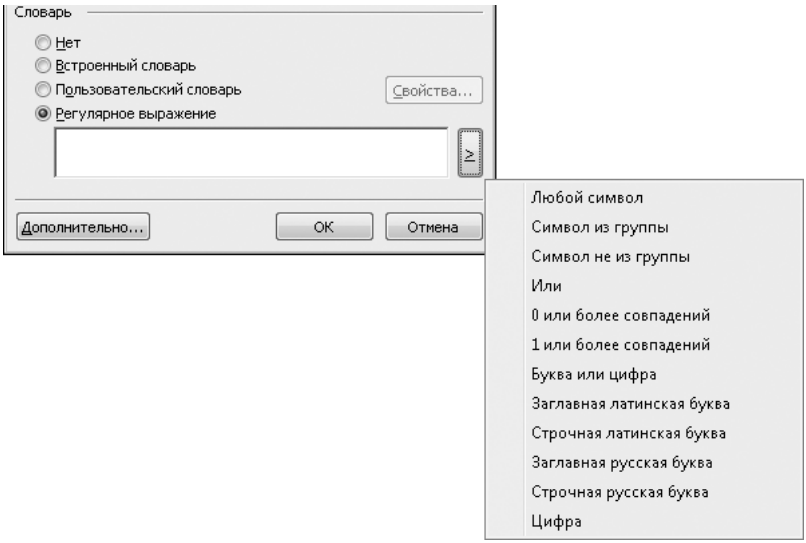


Рис. 6.31 ▾ Настройка регулярных выражений

тель **Словарь** установлен в положение **Регулярные выражения**. Это можно делать вручную, но намного удобнее использовать меню, открываемое нажатием кнопки со стрелкой (рис. 6.31).

При активизации команд данного меню в поле вставляется соответствующий символ или набор символов, которые используются в качестве условного обозначения.

Предположим, что нам нужно распознать хранящийся в формате TIFF текст, представляющий собой перечень дат (рис. 6.32).

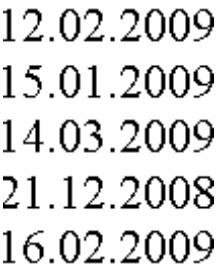


Рис. 6.32 ▾ Документ для распознавания

Теперь на основании русского языка создадим пользовательский язык Date (о том, как это делать, см. предыдущий раздел). В окне свойств языка установим переключатель **Словарь** в положение **Регулярное выражение**, а из поля **Алфавит языка** удалим все символы, оставив какой-нибудь один (например, точку), рис. 6.33.

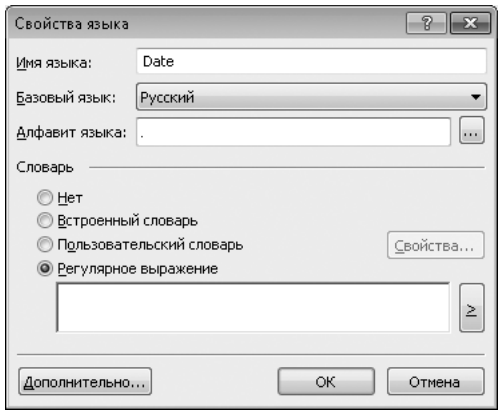


Рис. 6.33 ▾ Настройка языка для распознавания формализованного текста

Теперь выполним настройку регулярного выражения, с помощью которого будет произведено распознавание.

В поле регулярного выражения введем следующий текст:

```
((|0)[1-9])|([1|2][0-9])|(30)|(31)\.      ((|0)[1-
9])|(10)|(11)|(12)\.((19)[0-9][0-9])|([0-9][0-9])
```

Напомним, что делать это можно как вручную, так и с помощью соответствующих команд меню (см. рис. 6.31).

ПРИМЕЧАНИЕ

Обратите внимание – перед точкой мы ставим «обратный слеш» \ потому, что точка в данном случае играет роль разделителя между датой-месяцем и месяцем-годом (то есть является полноценным, а не служебным символом).

Теперь нажмем кнопку **Дополнительно** и проверим, чтобы в окне **Дополнительные свойства языка** был снят флажок **Наличие в тексте арабских и римских цифр, аббревиатур** (рис. 6.34).

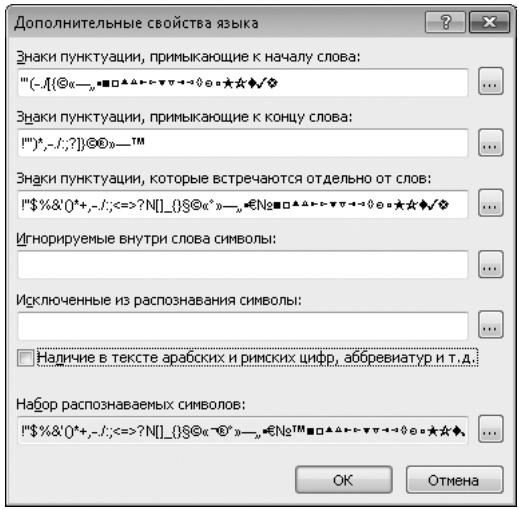


Рис. 6.34 ▾ Настройка дополнительных параметров

Нажимаем в данном окне кнопку **ОК**. Вновь открывшееся окно свойств языка в данный момент будет выглядеть, как показано на рис. 6.35 (регулярное выражение сформировано).

После нажатия в данном окне кнопки **ОК** программа выдаст запрос по поводу добавления в алфавит символов регулярного выражения (рис. 6.36).

На этот запрос отвечаем утвердительно. Нажимаем в редакторе языков кнопку **ОК** и выбираем в окне **Страницы** из раскрывающегося списка имя созданного языка – **Date** (рис. 6.37).

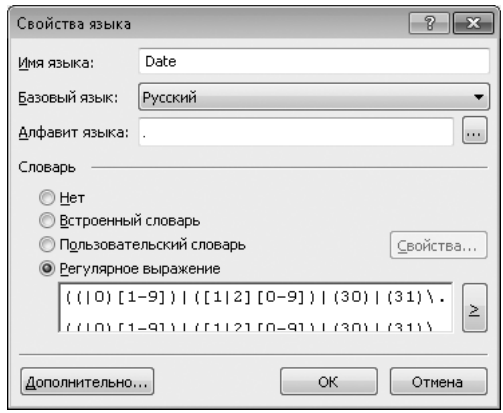


Рис. 6.35 ▾ Ввод регулярного выражения

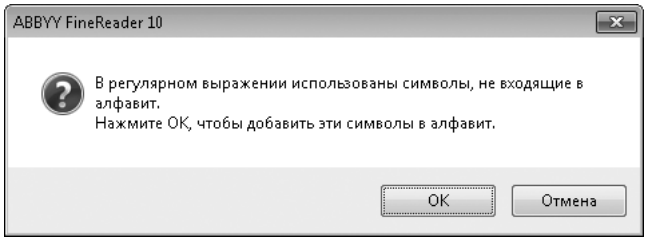


Рис. 6.36 ▾ Запрос программы

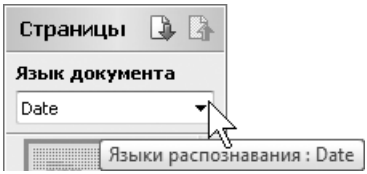


Рис. 6.37 ▾ Выбор языка с регулярным выражением

Теперь открываем документ (см. рис. 6.32) с помощью команды главного меню **Файл ➢ Открыть PDF/Изображение** или нажатием комбинации клавиш **Ctrl+O**. Результат распознавания показан на рис. 6.38.

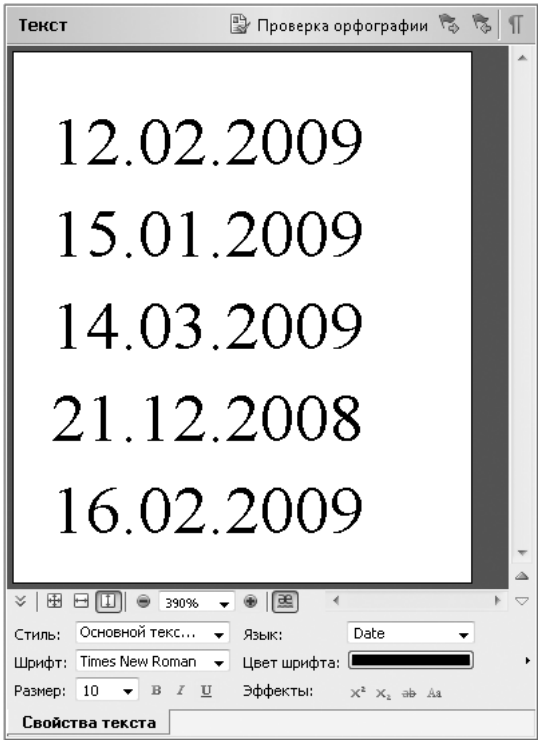


Рис. 6.38 ▾ Результат распознавания с помощью регулярных выражений

Как видно на рисунке, текст нашего документа распознан корректно.

Использование словарей

Для распознавания текстов FineReader использует множество словарей, каждый из которых соответствует определенному языку (для русского языка – свой словарь, для английского – свой и т. д.). Чтобы повысить качество распознавания, иногда бывает необходимо добавлять в словари новые слова, причем делать это можно по-разному.

Общие правила работы со словарями

Чтобы проверить неуверенно распознанные программой слова, удобно применять специальный режим (рис. 6.39), который вызывается с помощью команды главного меню **Сервис ➢ Проверка** либо нажатием комбинации клавиш **Ctrl+F7**. Здесь мы остановимся на нем вкратце, а более подробно рассмотрим в следующей главе.

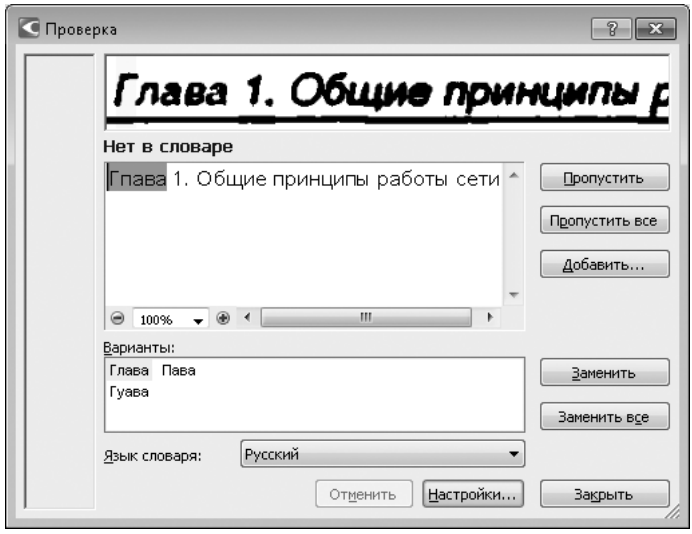


Рис. 6.39 ▾ Проверка неуверенно распознанных символов

Слова, отсутствующие в словаре, подчеркиваются в окне **Текст** красной волнистой линией. В окне **Проверка** такие слова выделяются розовым фоном. Возможны два случая, в которых программа сообщает о том, что слова нет в словаре:

- если неправильно распознаны отдельные символы, происходит то, что при наборе текста мы обычно называем опечаткой или грамматической

ошибкой. Например, на рис. 6.39 в слове *Глава* буква *л* была распознана как буква *п*. В списке **Варианты** отображаются похожие слова, присутствующие в словаре. В таком случае выберите из предлагаемых вариантов написания правильный и нажмите кнопку **Заменить**. Неправильно распознанное слово будет заменено на предложенное из словаря, а проверка продолжится;

- если вы видите, что слово было распознано верно, но в словаре оно действительно отсутствует, его нужно добавить в словарь. Например, в словаре может не быть некоторых специальных терминов, имен собственных, каких-либо редко употребляемых слов или сокращений. Для этого нажмите в окне **Проверка** кнопку **Добавить**. Отобразится окно, изображенное на рис. 6.40.

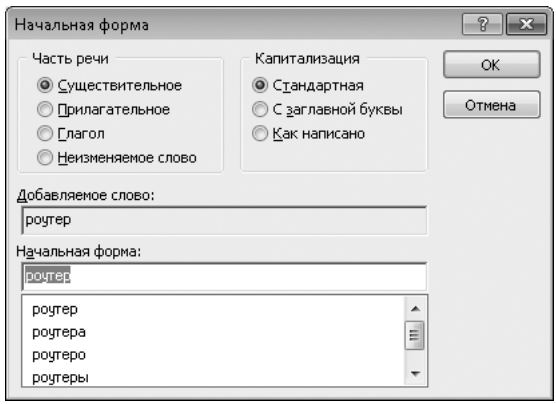


Рис. 6.40 ▾ Настройка параметров добавляемого слова

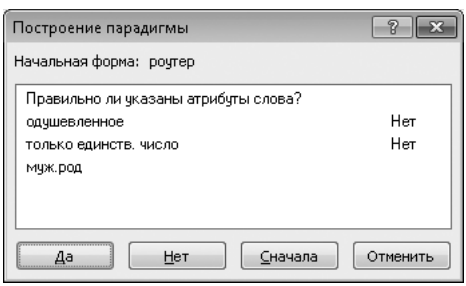


Рис. 6.41 ▾ Настройка парадигмы

В данном окне с помощью соответствующих переключателей нужно указать часть речи, к которой относится данное слово, а также капитализацию (возможность написания слова как с прописной, так и со строчной буквы). В поле **Начальная форма** отображается слово так, как оно распознано программой, но это значение можно отредактировать с клавиатуры.

После нажатия в данном окне кнопки **ОК** программа предложит построить парадигму добавляемого в словарь слова. В данном окне нужно ответить на вопросы программы нажатием кнопок **Да** или **Нет** (рис. 6.41).

С помощью кнопки **Сначала** можно вернуться к предыдущему вопросу, если, например, вы ответили на него ошибочно. После настройки парадигмы слово будет добавлено в словарь, а на экране вновь отобразится окно **Проверка** (см. рис. 6.39).

Можно добавлять новые слова в словари и иначе. Вначале откроем список словарей (рис. 6.42) с помощью команды главного меню **Сервис ➢ Просмотр словарей** или нажатием комбинации клавиш **Ctrl+Alt+D**.

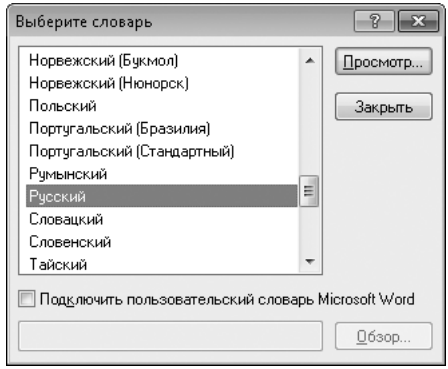


Рис. 6.42 ▾ Список словарей

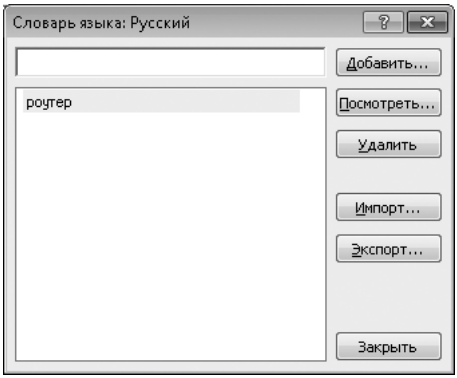


Рис. 6.43 ▾ Редактирование словаря

В данном окне выделим щелчком мыши словарь, который следует дополнить, и нажмем кнопку **Просмотр** – в результате на экране откроется окно, изображенное на рис. 6.43.

Здесь для каждого словаря показываются лишь те слова, которые были добавлены в него пользователем (именно поэтому на рисунке для русского языка показано лишь одно слово). Чтобы добавить в словарь слово, введите его в верхнем поле с клавиатуры и нажмите кнопку **Добавить**. Если такое слово уже имеется в словаре, программа выдаст соответствующее информационное сообщение (рис. 6.44).

Если в данном окне нажать кнопку **Добавить**, откроется окно настройки начальной формы слова (см. рис. 6.40), при нажатии кнопки **Парадигма** отобразится окно редактирования парадигмы.

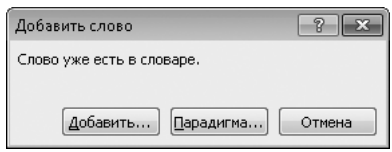


Рис. 6.44 ▾ Информация о наличии в словаре добавляемого слова

Чтобы удалить из словаря добавленное ранее слово, выделите его щелчком мыши и нажмите кнопку **Удалить**, после чего подтвердите удаление.

Вы можете создавать в программе пользовательские словари – но только при условии предварительного создания пользовательского языка: ведь, как мы уже отмечали ранее, каждый словарь «привязан» к соответствующему языку. В окне редактирования языка (см. рис. 6.29) нужно установить переключа-

тель **Словарь** в положение **Пользовательский словарь** и нажать кнопку **Свойства** – откроется окно редактирования словаря (см. рис. 6.43).

Для быстрого заполнения нового словаря удобно импортировать слова из внешнего источника (из файла формата txt или dic), в котором они разделены пробелами или иными символами, отсутствующими в алфавите данного языка. Рассмотрим на конкретном примере, как это делается.

Предположим, что хотим импортировать слова в словарь из внешнего текстового файла **Словарь-сети.txt**. Для этого в режиме редактирования словаря (см. рис. 6.43) нажмем кнопку **Импорт** и в открывшемся окне укажем путь к нашему файлу (рис. 6.45).

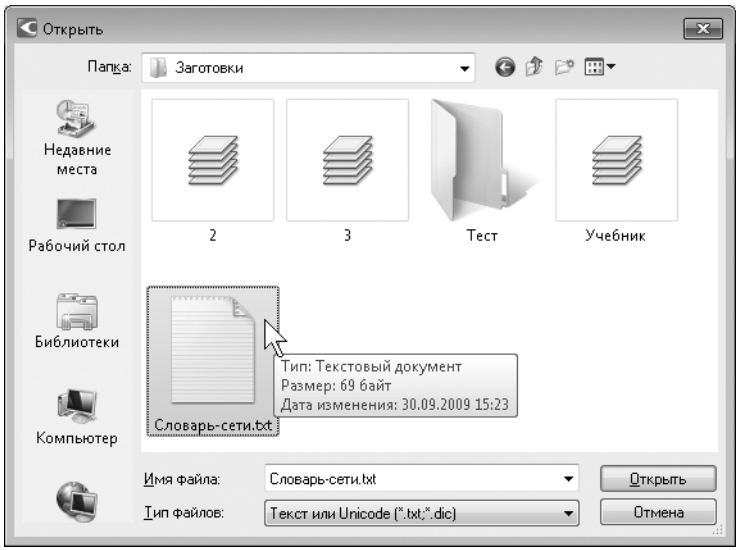


Рис. 6.45 ▾ Импорт данных из внешнего файла

После нажатия в данном окне кнопки **Открыть** в словарь будут добавлены слова из файла-источника (рис. 6.46).

Очевидно, что механизм импорта позволяет практически моментально заполнить новый словарь, избавляя пользователя от необходимости делать это вручную. Подобным образом можно импортировать данные из словарей предыдущих версий программы.

Любое из добавленных слов вы можете редактировать или удалять, предваритель-

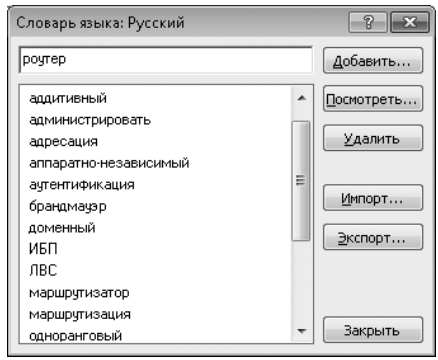


Рис. 6.46 ▾ Результат импорта данных в словарь

но выделив его щелчком мыши. После заполнения словаря нажмите в данном окне кнопку **Закрыть**.

С помощью кнопки **Экспорт** вы можете экспортировать содержимое словаря во внешний файл. При нажатии кнопки отобразится окно **Сохранить как** (рис. 6.47).

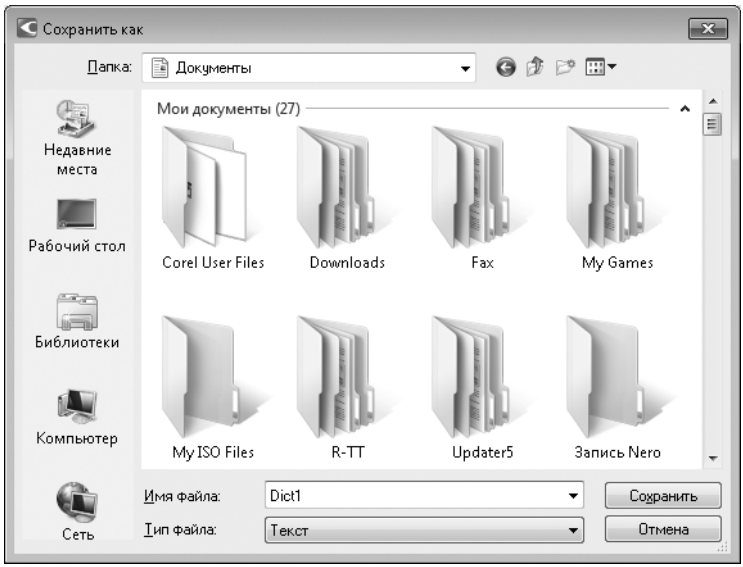


Рис. 6.47 ▾ Экспорт словаря во внешний файл

В данном окне в поле **Имя файла** введите с клавиатуры имя файла и нажмите кнопку **Сохранить**. Впоследствии эти данные вы сможете импортировать в другой словарь.

Пример редактирования и применения пользовательского словаря

В данном разделе мы на конкретном примере проиллюстрируем, как можно повысить качество распознавания с помощью пользовательских словарей. К этому приему уместно прибегнуть в том случае, если в ряде однотипных по содержанию документов используется довольно ограниченный словарный запас и важно обеспечить однозначное и безошибочное распознавание каждого слова. В качестве примеров можно привести официальные документы, инструкции, техническую документацию.

Сформируем новый язык на основе русского языка так, как рассказано выше, в разделе «Создание пользовательского языка». При этом в окне свойств языка установим переключатель **Словарь** в положение **Пользовательский словарь** (рис. 6.48).

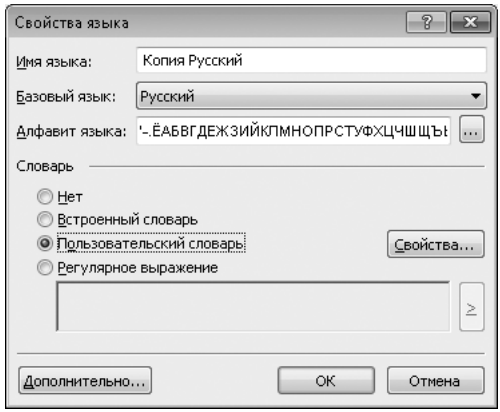


Рис. 6.48 ▾ Включение режима пользовательского словаря

Тем самым мы даем понять программе, что для распознавания документа следует применять не встроенный словарь и не регулярные выражения, а пользовательский словарь данного языка. В данный момент мы не будем заполнять наш пользовательский словарь и оставим его пустым.

Теперь попробуем с помощью созданного языка распознать текст документа.

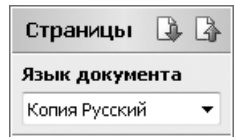


Рис. 6.49 ▾ Выбор языка для распознавания документа

Не забываем, что перед началом распознавания следует выбрать язык, который будет для этого применяться. В нашем примере пользовательский язык называется **Копия Русский**, и именно его следует выбрать в раскрывающемся списке окна **Страницы** (рис. 6.49).

Теперь можно запустить процесс распознавания с помощью команды главного меню **Файл ➢ Открыть PDF/Изображение** или нажатием комбинации клавиш **Ctrl+O**. Через некоторое время в окне **Текст** отобразится результат распознавания (рис. 6.50).

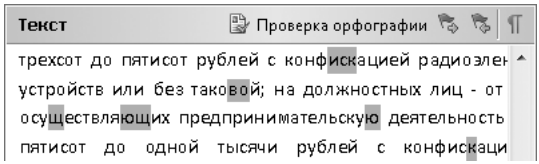


Рис. 6.50 ▾ Некорректное распознавание

Как видно на рисунке, многие символы в тексте распознаны неуверенно. Программе знакомы символы русского алфавита (ни один из них не заменен квадратиком или чем-то еще), потому что они заданы в свойствах языка (см. рис. 6.48, поле **Алфавит языка**), но вот как правильно ими распорядиться – она

не знает: встроенный словарь отключен (см. рис. 6.48), а пользовательский мы не заполнили.

Следовательно, нужно устранить этот недочет. Открываем окно редактирования языка (см. рис. 6.48) и нажимаем кнопку **Свойства**. Затем в открывшемся окне добавляем в словарь слова из нашего документа, используя кнопку **Добавить** (рис. 6.51).

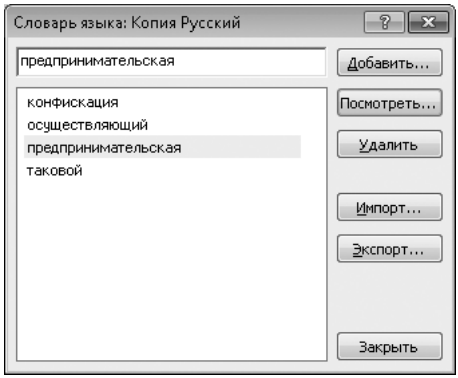


Рис. 6.51 ▾ Добавление слов в пользовательский словарь

Нажимаем в данном окне кнопку **Закрыть**, затем в окне свойств языка – кнопку **ОК** и повторно распознаем документ – опять же с помощью пользовательского языка **Копия Русский** (см. рис. 6.48). Результат распознавания показан на рис. 6.52.

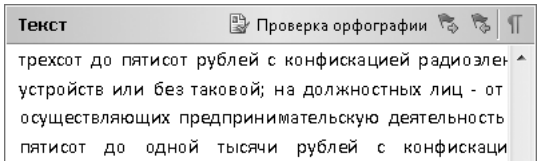


Рис. 6.52 ▾ Корректное распознавание текста

Как видно на рисунке, сейчас наш документ распознан полностью корректно. Это стало возможным благодаря тому, что мы добавили незнакомые программе слова в пользовательский словарь, тем самым указав ей, каким образом следует распорядиться известными ей символами русского алфавита.

Резюме

Итак, подведем краткие итоги.

Поскольку корректно и правильно тексты распознаются далеко не всегда (особенно это касается специфических и нестандартных документов, состав-

ленных с применением нетрадиционных символов), возникает необходимость научить программу работе с такими документами. Для этого в программе FineReader реализована возможность создания и обучения пользовательских эталонов, распознавания многоязычных документов, применения пользовательских языков и словарей.

Знаний, полученных в данной главе, вполне достаточно для того, чтобы научиться распознавать любые документы, независимо от их содержимого и уровня сложности.

Но даже в случае успешного распознавания текста его иногда приходится дополнительно проверять, корректировать, дополнительно обрабатывать (в том числе и средствами Microsoft Word). О том, как это делать, мы расскажем в следующей главе.

Глава 7 Проверка и корректировка распознанного документа

Нередко распознанный документ приходится «доводить до ума»: заменить пару-тройку символов, поправить «слетевшее» или создать новое форматирование, где-то что-то подкорректировать и т. д. В этой главе мы расскажем о том, как производятся окончательная обработка и оформление распознанных текстов.

Проверка и корректировка документа в программе FineReader

Проверка и корректировка распознанных текстов в программе FineReader осуществляется двумя взаимодополняющими способами: непосредственно в рабочей области окна **Текст** и в диалоговом окне **Проверка**. Рассмотрим порядок работы в каждом из них.

Пример корректировки документа в окне Текст

Предположим, что нам нужно распознать хранящийся в формате tiff документ, текст которого представлен на рис. 7.1.

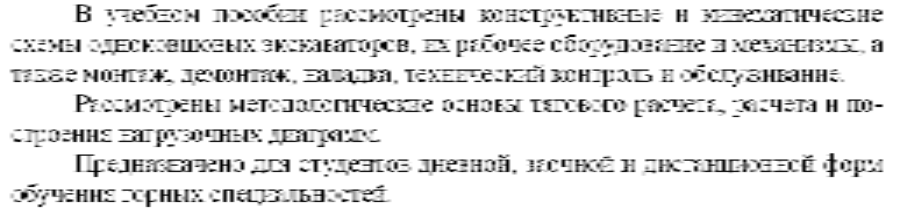


Рис. 7.1 ▾ Пример текста для распознавания

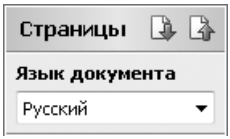


Рис. 7.2 ▾ Выбор языка для распознавания документа

Распознаем его с применением русского языка: текст несложен для распознавания, не содержит иностранных слов и специальных символов. Выберем соответствующее значение в раскрывающемся списке окна **Страницы** (рис. 7.2).

После этого запустим процесс распознавания с помощью команды главного меню **Файл ➢ Открыть PDF/Изображение** или нажатием комбинации клавиш **Ctrl+O**. Результат распознавания (напомним, что он отображается в окне **Текст**, которое по умолчанию находится в правой части рабочего интерфейса программы) представлен на рис. 7.3.

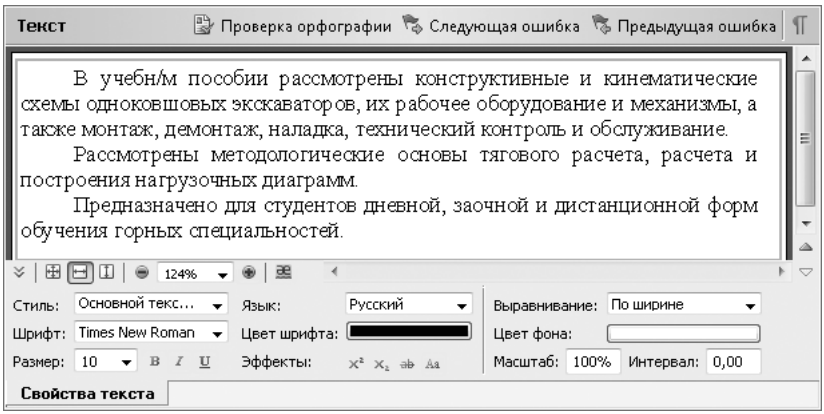


Рис. 7.3 ▾ Результат распознавания текста

Как видно на рисунке, текст распознан, в общем-то, корректно; есть лишь одно замечание: в слове *учебном* почему-то не определился один символ. Вот с этого мы и начнем корректировку распознанного документа в окне **Текст**.

А кроме этого, выполним еще следующие действия:

- оформим некоторые фрагменты текста полужирным, курсивным и подчеркнутым начертанием;
- применим к одному текстовому фрагменту шрифт, отличный от остального текста;
- изменим размер шрифта выделенного фрагмента текста;
- вставим в документ гиперссылку на внешний файл, хранящийся на жестком диске.

Как мы уже знаем, для устранения ошибок и повышения качества распознавания текста можно применять обучаемые пользовательские эталоны, а также редактировать словари. Но в нашем случае это не имеет смысла: ведь в распознанном тексте имеется всего одна неточность, и намного проще исправить ее вручную, чем затевать процедуру создания и обучения пользовательского эталона или внесения дополнений в словарь.

Итак, устанавливаем курсор на слово *учебном* после буквы «*н*», нажимаем клавишу **Delete** (чтобы удалить ошибочно распознанный символ) и вместо него вводим с клавиатуры букву «о». Подобная процедура наверняка знакома тем, кто имеет хотя бы минимальный опыт работы с текстовыми документами.

Первую задачу мы решили. Теперь научимся применять разные виды начертания текста.

Отметим, что для форматирования распознанного текста, а также выполнения ряда иных операций (отмены последнего действия, работы с буфером обмена и др.) предназначены кнопки главной панели инструментов, а также панель **Свойства текста**, которая расположена внизу окна **Текст** (см. рис. 7.3).

ВНИМАНИЕ

Учтите, что доступность кнопок инструментальной панели, а также инструментов, содержащихся в панели ***Свойства текста****, зависит от того, какой тип документа выбран в данный момент. Например, для документа Microsoft Word доступны все инструменты, а для текстового документа (формат txt) – лишь малая часть из них. Тип документа указывается в раскрывающемся списке, который находится слева вверху окна* ***Текст****. В этом разделе мы работаем с документом типа Microsoft Word (см. рис. 7.3).*

Выделим в тексте слова **конструктивные** и **кинематические** и нажмем на инструментальной панели кнопку [B] либо комбинацию клавиш **Ctrl+B**. В результате к этим словам будет применено полужирное начертание. С учетом исправленной ранее ошибки в данный момент наш текст будет выглядеть так, как показано на рис. 7.4.

Теперь выделим в тексте словосочетание **одноковшовых экскаваторов**, нажмем на инструментальной панели кнопку [I] (или комбинацию клавиш **Ctrl+I**), затем выделим словосочетание **методологические основы** и нажмем

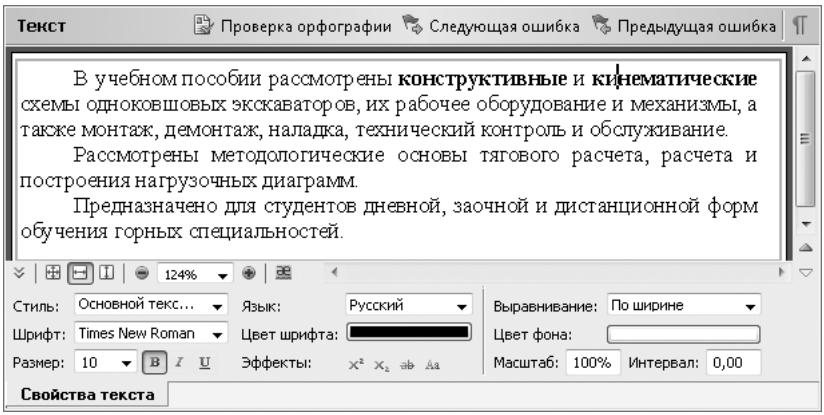


Рис. 7.4 ▾ Выделение слов полужирным шрифтом

на панели инструментов кнопку U (или комбинацию клавиш **Ctrl+U**). Результат выполненных действий показан на рис. 7.5.

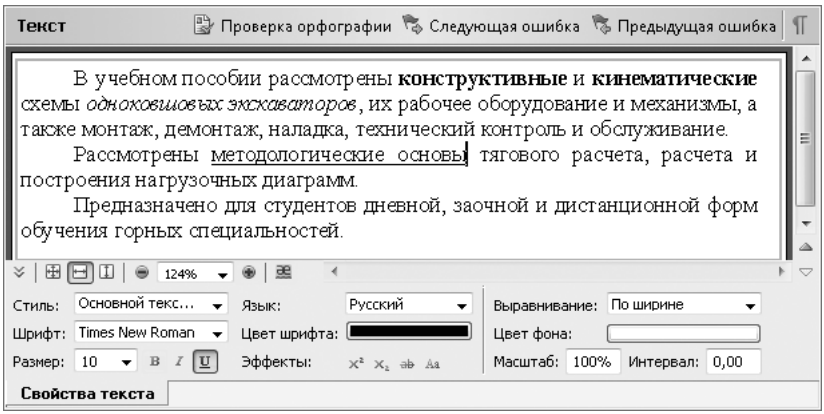


Рис. 7.5 ▾ Выделение текста курсивом и подчеркиванием

Как видно на рисунке, к первому словосочетанию применилось курсивное начертание, а ко второму – подчеркнутое.

Следующая задача – применение к текстовому фрагменту шрифта, отличающегося от остального текста. Пусть в нашем примере это будет второй абзац. Выделим его и в раскрывающемся списке, который расположен на инструментальной панели, выберем значение **Arial** (рис. 7.6).

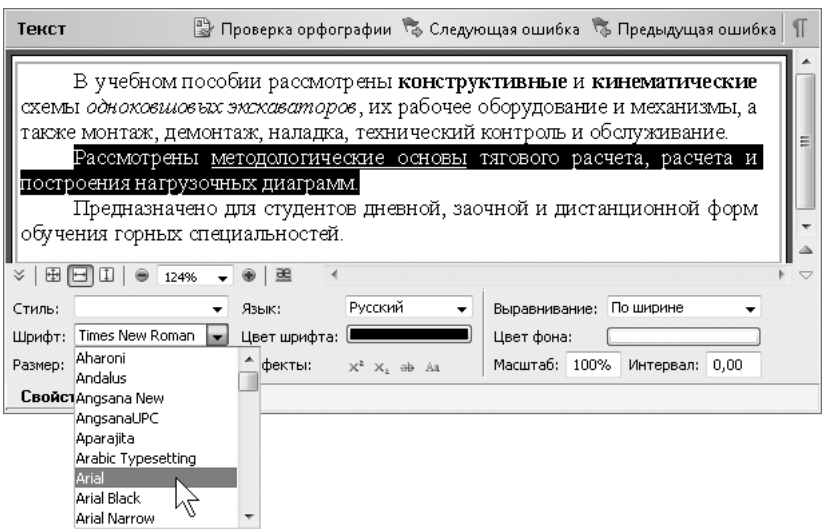


Рис. 7.6 ▾ Выбор шрифта

Отметим, что аналогичным образом можно выбрать шрифт и на панели **Свойства текста**. В любом случае результат получится таким, как показано на рис. 7.7.

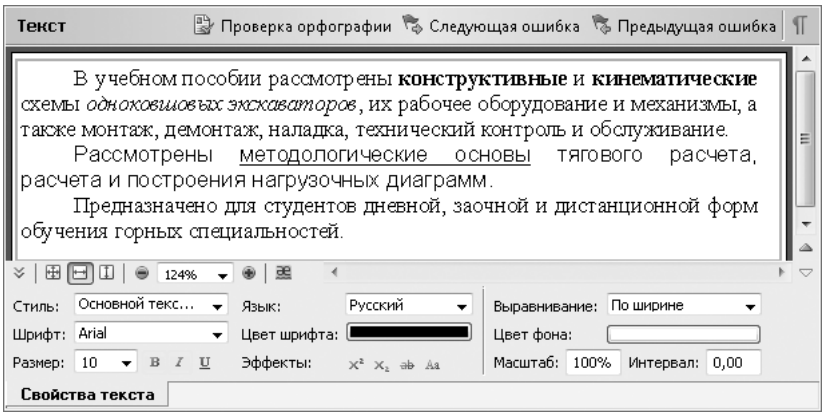


Рис. 7.7 ▾ Применение шрифта Arial к фрагменту текста

Очевидно, что необходимо уменьшить размер выбранного шрифта: с учетом его специфики текст стал смотреться неэргономично. Для этого опять выделяем второй абзац и в раскрывающемся списке, который находится в инструментальной панели справа от поля выбора шрифта, выбираем значение **9** (рис. 7.8).

Рис. 7.8 ▾ Выбор размера шрифта

Таким образом, мы решили следующую задачу из нашего списка – изменение размера шрифта выделенного фрагмента.

Как видно на рисунке, после уменьшения размера шрифта второй абзац стал выглядеть намного эргономичнее.

Следующая наша задача – вставить в распознанный документ гиперссылку на внешний файл. Для примера возьмем обычный текстовый файл, назовем его `Учебник.txt` и сохраним в папке `dokument`. Текст этого документа может быть совершенно произвольным; если принимать во внимание специфику нашего примера, то в подобных случаях можно делать ссылку из текста аннотации на файл с учебником.

Гиперссылкой в нашем примере будет являться словосочетание **учебном пособии**. Выделите его, щелкните на выделенном правой кнопкой мыши и в открывшемся контекстном меню выберите команду **Гиперссылка** (рис. 7.9).

Для вставки гиперссылки можно также нажать кнопку **Редактировать гиперссылку** на главной панели инструментов. В обоих случаях на экране отобразится окно **Редактирование гиперссылки**, изображенное на рис. 7.10.

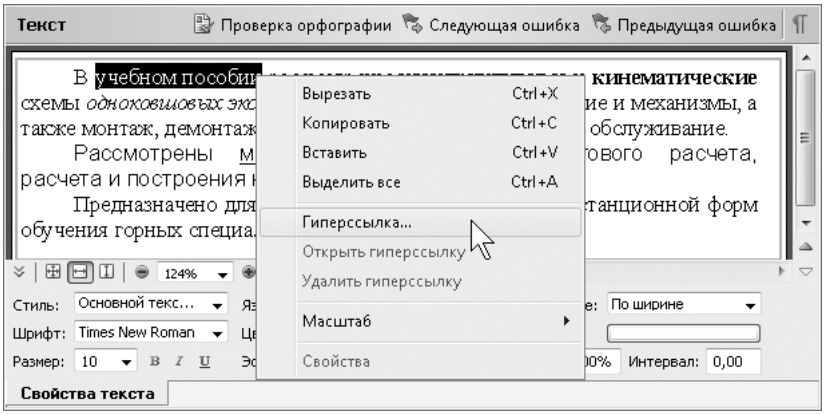


Рис. 7.9 ▾ Размер шрифта изменен

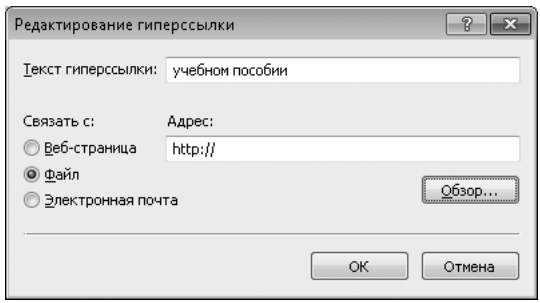


Рис. 7.10 ▾ Редактирование гиперссылки

В данном режиме выполняются все действия по созданию и редактированию гиперссылок в окне **Текст**.

Как видно на рисунке, в окне **Редактирование гиперссылки** автоматически заполнено поле **Текст гиперссылки**. В нем по умолчанию предлагается использовать значение, соответствующее выделенному в тексте фрагменту (при отсутствии выделенного фрагмента в данном поле отобразится слово, на котором был установлен курсор). При необходимости вы можете с клавиатуры изменить это значение.

Далее необходимо указать, на какой объект должна указывать наша гиперссылка. Поскольку мы планируем сделать гиперссылку на текстовый файл, хранящийся на жестком диске компьютера, то установим переключатель **Связать с** в положение **Файл**. При этом станет доступной находящаяся справа кнопка **Обзор**, при нажатии которой на экране откроется окно, изображенное на рис. 7.11.

В данном окне нужно выбрать каталог и указать требуемый файл. Щелчком мыши выделите файл, к которому должна вести гиперссылка, и нажмите кноп-

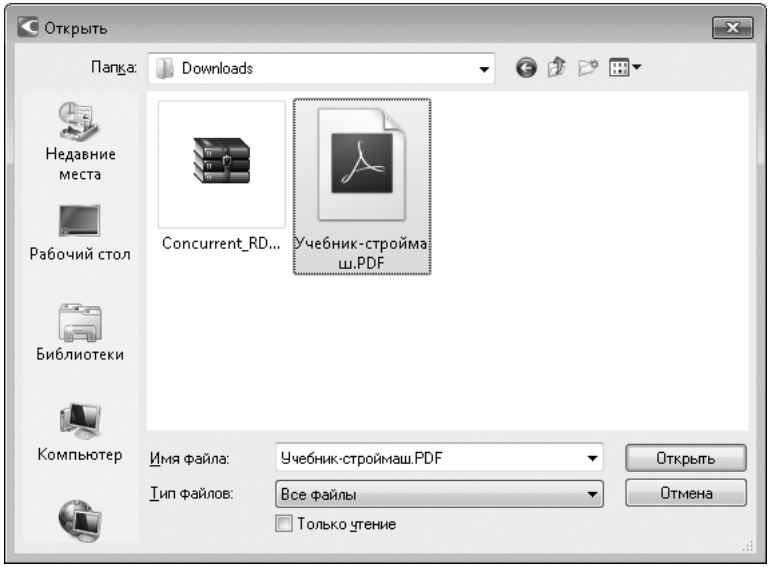


Рис. 7.11 ▾ Выбор внешнего файла для гиперссылки

ку **Открыть**. В результате выполненных действий в окне **Редактирование гиперссылки** будет заполнено поле **Адрес** (рис. 7.12).

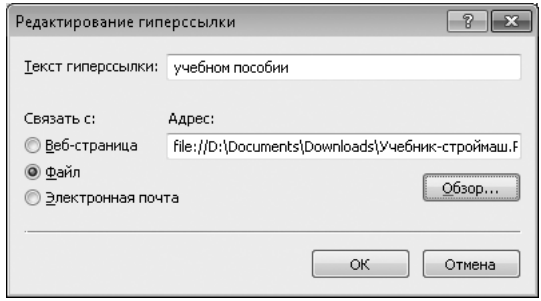


Рис. 7.12 ▾ Ввод адреса гиперссылки

Теперь нам осталось лишь нажать в данном окне кнопку **ОК** – и гиперссылка готова. Если вы наведете указатель мыши на гиперссылку, то увидите всплывающую подсказку (рис. 7.13). В ней приведен адрес, к которому ведет гиперссылка, и говорится, что для перехода по ссылке нужно нажать клавишу **Ctrl** и щелкнуть кнопкой мыши по ссылке.

Чтобы перейти по гиперссылке, щелкните на ней мышью, удерживая нажатой клавишу **Ctrl**. В нашем примере на экране должен открыться документ, путь к файлу которого мы указали в настройках гиперссылки (см. рис. 7.11 и 7.12).

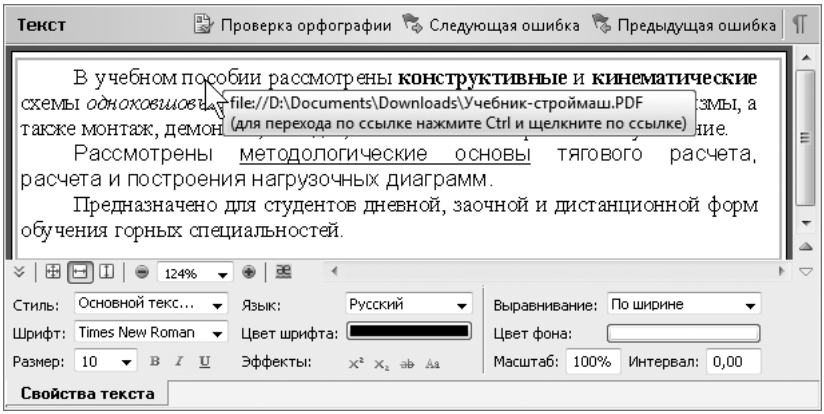


Рис. 7.13 ▾ Гиперссылка в окне **Текст**

Подобным образом осуществляется вставка в документ гиперссылок, связанных с веб-страницами в Интернете. В этом случае нужно в настройках гиперссылки (см. рис. 7.12) установить переключатель **Связать с** в положение **Веб-страница** и в расположенном справа поле **Адрес** ввести адрес (URL) соответствующей веб-страницы, например **http://www.abbyy.ru**.

Чтобы создать ссылку для отправки электронного письма по указанному адресу, установите переключатель **Связать с** в положение **Электронная почта** и в расположенном справа поле **Адрес** введите адрес E-mail, например **mailto:support@abbyy.ru**. Отметим, что адреса можно как вводить с клавиатуры, так и вставлять из буфера обмена.

Когда распознанный документ будет сохранен в одном из поддерживаемых форматов, а сохраненный файл затем открыт в соответствующей программе, в нем также будут присутствовать гиперссылки, созданные в программе FineReader. Это позволяет непосредственно из документа открывать другие файлы, переходить на веб-страницы или отправлять сообщения на адрес электронной почты, указанный в качестве ссылки. Обычно гиперссылки принято использовать в документах HTML (веб-страницах) и в документах PDF. Иногда гиперссылки вставляют и в документы Microsoft Office – все приложения этого пакета поддерживают переходы по ссылкам.

После того как обработка распознанного текста завершена, документ следует сохранить в отдельном файле с помощью кнопки **Сохранить**. О том, как это делать, мы расскажем ниже, в главе 8 «Сохранение распознанного документа».

В следующем же разделе мы расскажем о том, каким образом осуществляются корректировка и проверка распознанного текста в диалоге **Проверка**.

Пример проверки и корректировки текста в диалоге Проверка

В программе реализована возможность проверки и, при необходимости, корректировки неуверенно распознанного текста в автоматическом режиме. Рассмотрим на конкретном примере, как это делается.

Предположим, что нам нужно распознать оригинал, который изображен на рис. 7.14.

В данном учебном пособии рассказано о диференциалах, суппортах, шкворнях, карданах, ступицах, ШРУСах и коленвалах. Схемы обозначены как сх., таблицы – как табл., рисунки – как рис.

Рис. 7.14 ▾ Пример текста для распознавания

Как видно на рисунке, данный текст содержит несколько специфических терминов, а также сокращений. Обратите внимание – в самом оригинале в одном слове допущена ошибка: *диференциал*.

Запустим процесс распознавания с помощью команды главного меню **Файл ➤ Открыть PDF/Изображение** или нажатием комбинации клавиш **Ctrl+O**. Результат распознавания представлен на рис. 7.15.

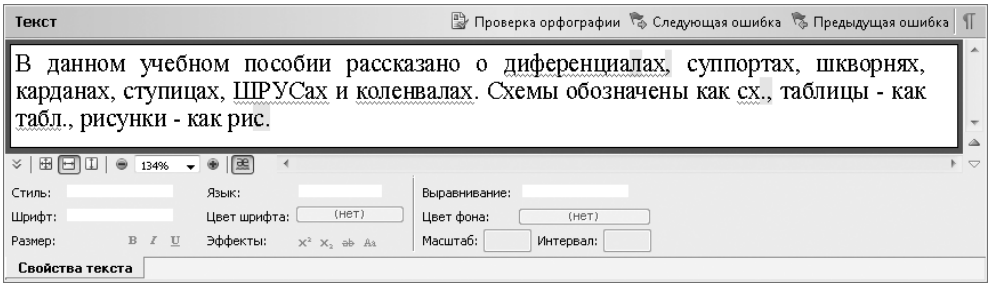


Рис. 7.15 ▾ Результат распознавания документа

В целом текст распознан правильно. Однако в нем присутствуют неуверенно распознанные символы, а некоторые слова подчеркнуты, поскольку отсутствуют в словаре. Поможем программе разобраться с этими затруднениями.

Для перехода в режим автоматической проверки документа выполните команду главного меню **Сервис ➤ Проверка** (эта команда вызывается также нажатием комбинации клавиш **Ctrl+F7**). В результате откроется диалог **Проверка** (рис. 7.16).

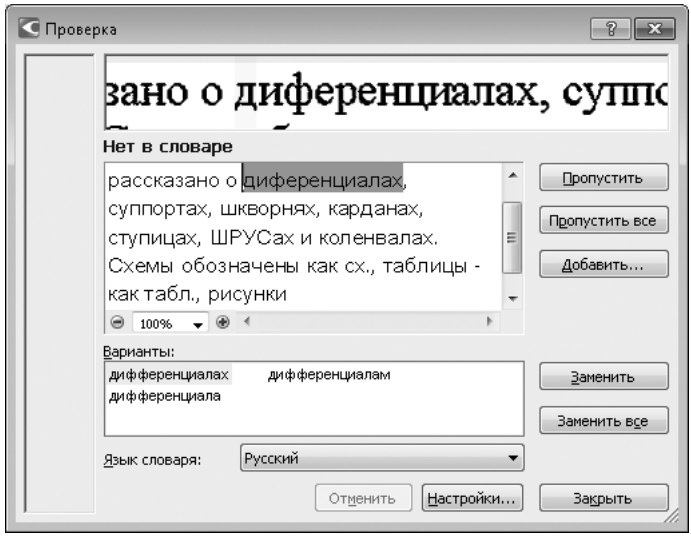


Рис. 7.16 ▾ Диалог **Проверка**

Структура данного окна такова: в его верхней части содержится аналог окна **Крупный план**, которое по умолчанию находится внизу рабочего интерфейса FineReader под окном **Текст** (кстати, управление отображением окна **Крупный план** осуществляется с помощью комбинации клавиш **Ctrl+F5**, а также команд подменю **Вид ➢ Окно Крупный план**).

Сразу под ним отображается тип обнаруженной ошибки или некорректности (на рис. 7.20 тип ошибки – **Нет в словаре**). Далее следует распознанный текст, в котором выделено слово с ошибкой. Внизу окна в поле **Варианты** представлен перечень вариантов исправления обнаруженной ошибки (например, предлагаются имеющиеся в словаре слова, которые по написанию похожи на обнаруженное ошибочное слово).

На рисунке видно, что в распознанном тексте выделено слово, которое на самом деле содержит ошибку, – **диференциалах**. Программе оно незнакомо, но в словаре есть несколько похожих слов, которые FineReader предлагает в качестве замены: **дифференциалах**, **дифференциалам** и **дифференциала**. Очевидно, что подходящий вариант замены в нашем примере – слово **дифференциалах**. Поэтому выделим его в поле **Варианты** щелчком мыши и нажмем кнопку **Заменить**, которая находится в правой части окна. Результат выполненных действий показан на рис. 7.17.

Обратите внимание: в окне **Текст** слово заменено на правильный вариант – **дифференциалах**. В то же время в диалоговом окне **Проверка** выполнен переход к следующему неуверенно распознанному слову – **ШРУСах**. Ни один из предложенных вариантов замены в данном случае нам не подходит по той простой причине, что это слово отсутствует в словаре. Программа, предлагая варианты замены, полагает, что слово написано с ошибкой (как в предыдущем

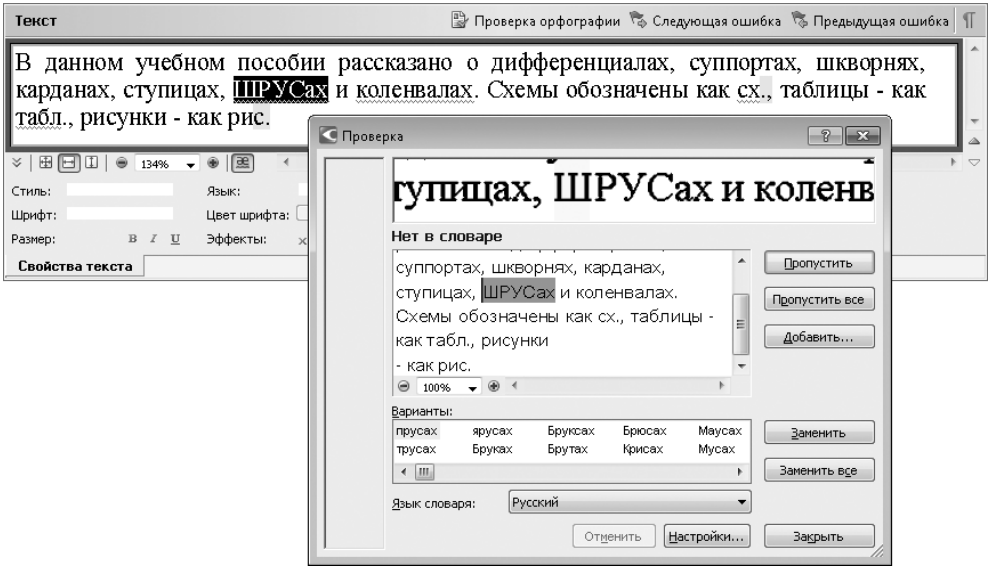


Рис. 7.17 ▾ Результат замены ошибочно написанного слова

случае), но мы-то знаем, что это не так. Следовательно, незнакомое программе слово необходимо добавить в словарь, чтобы в будущем оно распознавалось уверенно и корректно. Для этого нажмем кнопку **Добавить** – в результате на экране откроется окно **Начальная форма**. В данном окне определим значения параметров так, как это показано на рис. 7.18.

Обратите внимание: в поле **Начальная форма** необходимо правильно указать начальную форму слова: в именительном падеже и единственном числе. В нашем случае по умолчанию предложено значение *ШРУСах* (то есть так, как слово встретилось в распознанном тексте), и его нужно поправить.

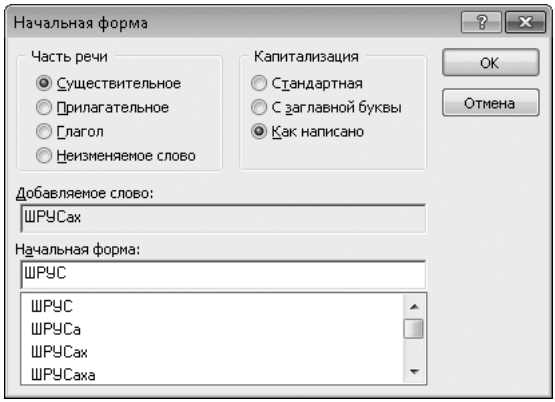


Рис. 7.18 ▾ Добавление слова в словарь

После нажатия в данном окне кнопки **ОК** на экране отобразится окно **Построение парадигмы** (рис. 7.19).

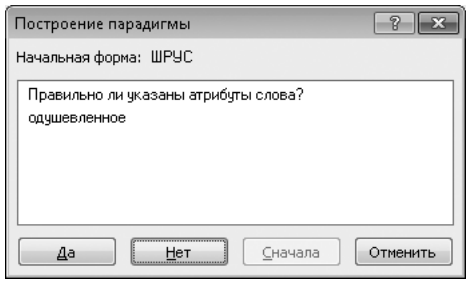


Рис. 7.19 ▾ Окно **Построение парадигмы**

Поскольку предложенный в данном окне атрибут указан неверно, нажимаем кнопку **Нет**. При этом произойдет переход к следующему этапу добавления слова (рис. 7.20).

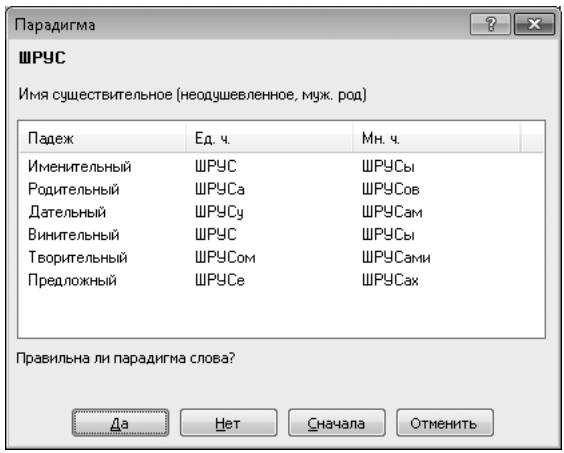


Рис. 7.20 ▾ Парадигма добавляемого слова

После нажатия в данном окне кнопки **Да** новое слово будет добавлено в словарь. Если оно впоследствии встретится в этом или другом документе, FineReader уверенно его распознает.

После добавления слова в словарь на экране вновь отобразится диалог **Проверка**, в котором будет выделено следующее неуверенно распознанное слово – **коленвалах** (рис. 7.21).

Как видно на рисунке, предложенный вариант замены нас не устраивает: предлагается заменить неуверенно распознанное слово двумя известными программе словами. Решить проблему можно двумя способами: либо добавить сло-

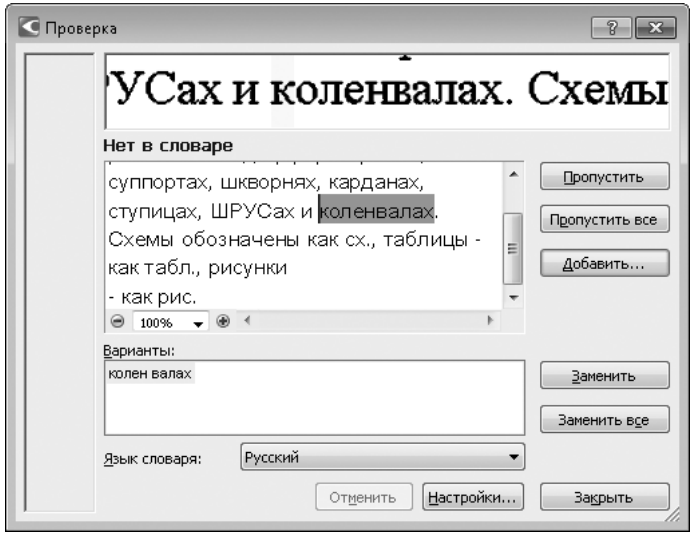


Рис. 7.21 ▾ Очередное неуверенно распознанное слово

во в словарь так, как мы это сделали в предыдущем случае, либо просто игнорировать его, нажав в диалоге **Проверка** кнопку **Пропустить**. Пропустить слово разумнее в том случае, если оно встречается в документе всего один-два раза и добавлять в словарь его нецелесообразно.

После нажатия кнопки **Пропустить** программа перейдет к следующему проблемному фрагменту распознанного текста (рис. 7.22).

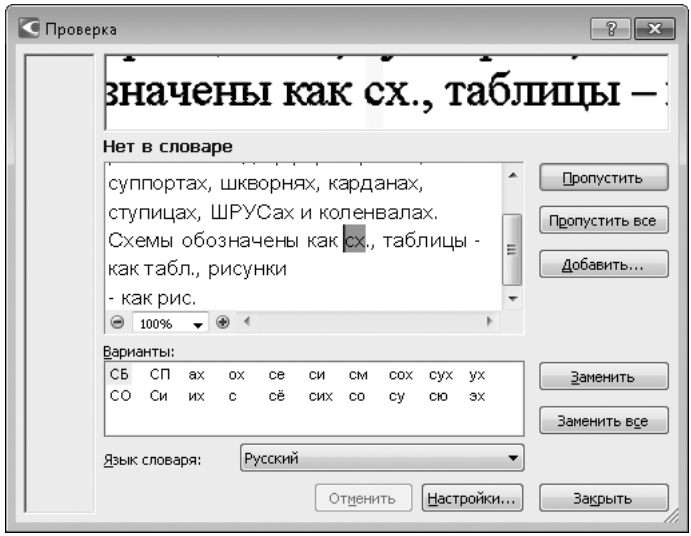


Рис. 7.22 ▾ Следующий проблемный фрагмент

Очевидно, что в данном случае программе незнакомо сокращение (**сх**) слова **схема**. Можно добавить его в словарь, но удобнее – пропустить, что мы и сделаем.

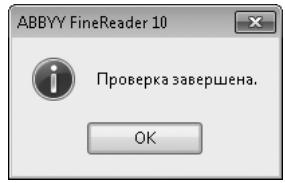


Рис. 7.23 ▾ Сообщение о завершении проверки

Как оказалось, этот проблемный фрагмент в нашем примере являлся последним. Поэтому после нажатия кнопки **Пропустить** проверка будет завершена, о чем на экране отобразится соответствующее информационное сообщение (рис. 7.23).

После нажатия в данном окне кнопки **ОК** диалог **Проверка** автоматически закроется. Теперь наш документ будет выглядеть в окне **Текст** так, как показано на рис. 7.24.

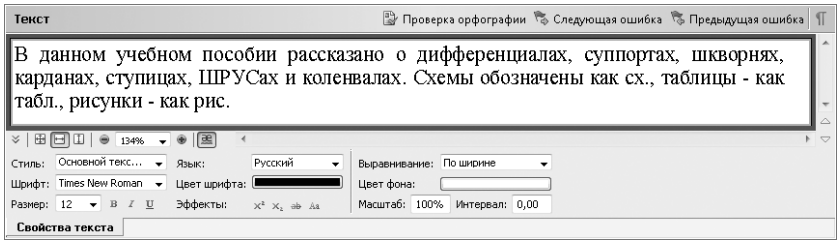


Рис. 7.24 ▾ Результат проверки и корректировки текста

Обратите внимание: из документа исчезли все подчеркивания и выделения: это свидетельствует о том, что FineReader уверен в правильности и корректности распознавания.

В диалоге **Проверка** есть еще несколько инструментов, которые в нашем примере оказались не задействованы. Кратко остановимся на каждом из них.

С помощью кнопки **Заменить все** можно быстро заменить все сомнительные слова предложенными по умолчанию вариантами замены. Чтобы применять данный метод, нужно быть уверенным в том, что FineReader по умолчанию предложит действительно подходящие варианты (учитывая падеж, род, число и др.). Кнопка **Пропустить все** предназначена для пропуска текущей и всех последующих аналогичных некорректностей.

Если вы неверно исправили какую-то ошибку и перешли к следующей, то всегда можете вернуться назад с помощью кнопки **Отменить**. Причем возвращаться можно не только на одну, но и на несколько позиций назад, нажав данную кнопку соответствующее число раз.

В поле **Язык словаря** вы можете заменить язык словаря, выбрав его из раскрывающегося списка.

С помощью кнопки **Настройки** осуществляется переход в режим настройки параметров проверки текста. При этом на экране открывается окно, изображенное на рис. 7.25.

На данном рисунке представлены настройки, которые используются в программе по умолчанию; именно при такой конфигурации мы рассматривали

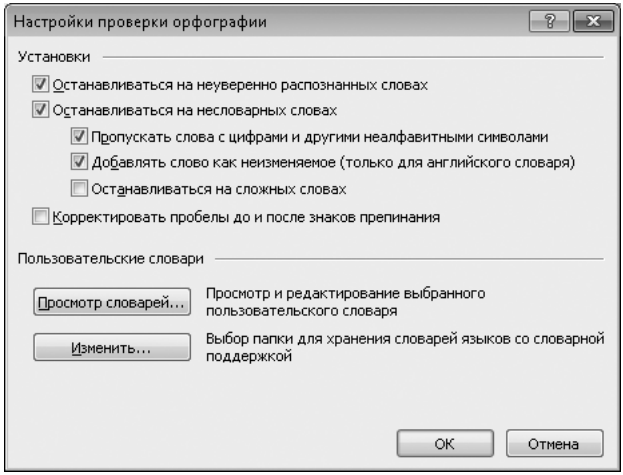


Рис. 7.25 ▾ Настройка параметров проверки

приведенный выше пример. Если вы снимете, например, флажок **Останавливаться на неуверенно распознанных словах**, то в диалоге **Проверка** неуверенно распознанные слова будут пропускаться. При установленном флажке **Останавливаться на несловарных словах** программа будет выделять слова, отсутствующие в словаре, даже если все символы в этом слове распознаны уверенно.

В некоторых случаях полезно включить опцию **Останавливаться на сложных словах**. Это особенно актуально для тех текстов, в которых встречается большое количество сложных и незнакомых слов.

Перед проверкой некоторых текстов целесообразно отключить параметр **Пропускать слова с цифрами и другими неалфавитными символами**. Характерный пример – адреса электронной почты: они могут содержать цифры, и в любом из них присутствует символ @. Добавлять в словарь адреса не имеет никакого смысла, равно как и тратить время на их проверку.

Также заслуживает внимания параметр **Корректировать пробелы до и после знаков препинания**. По правилам правописания знаки препинания пишутся слитно с предшествующим словом, а после знака препинания ставится пробел. В ходе распознавания в текст могут быть вставлены лишние пробелы перед знаками препинания и, наоборот, пропущены пробелы после них. При установленном данном флажке в процессе проверки текста FineReader будет акцентировать ваше внимание на этих недостатках. Снимать данный флажок целесообразно при распознавании текстов, имеющих большое количество специфических формул, обозначений, нестандартных текстовых конструкций и т. п.

С помощью кнопки **Просмотр словарей** вы можете перейти в режим просмотра и редактирования какого-либо пользовательского словаря. При нажатии кнопки открывается окно **Выберите словарь**, которое вызывается также с помощью команды главного меню **Сервис ➢ Просмотр словарей** или нажа-

тием комбинации клавиш **Ctrl+Alt+D**. Порядок работы в данном режиме мы рассмотрели в предыдущей главе.

Кнопка **Изменить** предназначена для замены каталога, в котором хранятся словари. При ее нажатии открывается окно **Обзор папок** (рис. 7.26).

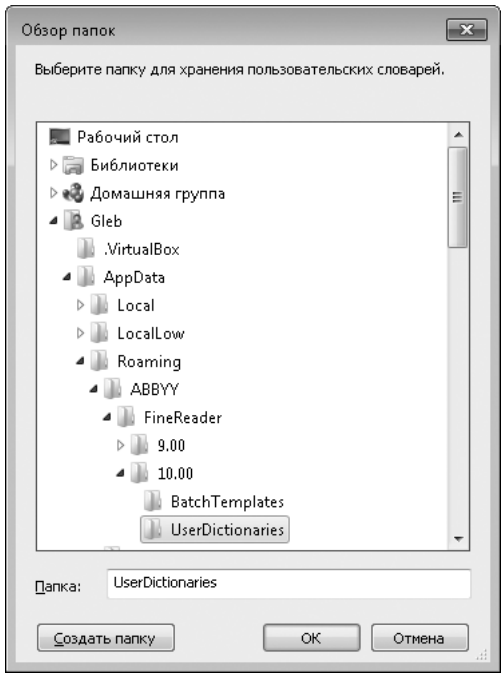


Рис. 7.26 ▾ Выбор каталога для хранения словарей

В данном окне нужно щелчком мыши указать требуемый каталог и нажать кнопку **ОК**. При необходимости вы можете создать для этого новый каталог, нажав кнопку **Создать папку**.

Все изменения, выполненные в режиме настройки параметров проверки текста, вступают в силу только после нажатия кнопки **ОК** или клавиши **Enter**. Кнопка **Отмена** предназначена для выхода из данного режима без сохранения изменений.

Использование стилей

Для оформления текста, распознанного программой и отображающегося в окне **Текст**, в FineReader реализована возможность применения стилей. В данном случае стиль – это набор определенных правил форматирования и оформления, объединенных под одним названием.

Общие сведения о стилях в FineReader

Отличие стиля от тех способов форматирования, которые мы рассмотрели выше, заключается в следующем. Например, нам нужно применить к выделенному фрагменту текста одновременно полужирное и курсивное начертания, при этом сам шрифт также должен отличаться. Если действовать так, как в предыдущем разделе, то придется вначале нажимать на инструментальной панели кнопку **B**, затем кнопку *I*, а после этого из раскрывающегося списка выбрать требуемый тип шрифта.

Механизм стилей позволяет решить эту задачу одним действием – выбором предварительно настроенного стиля из соответствующего раскрывающегося списка на панели **Свойства текста** (рис. 7.27).

Вы можете самостоятельно создавать, редактировать и удалять стили оформления. В следующем разделе мы на конкретном примере продемонстрируем, каким образом в FineReader осуществляется оформление распознанных текстов с помощью стилей.

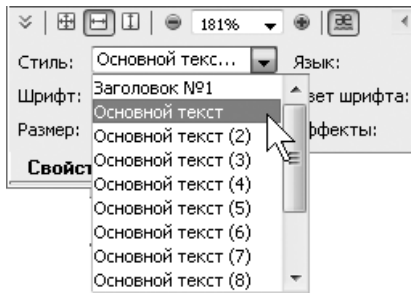


Рис. 7.27 ▾ Выбор стиля для оформления текста

Пример создания и применения пользовательского стиля

Чтобы перейти в режим работы с пользовательскими стилями, выполните команду главного меню **Сервис** ➢ **Редактор стилей**. При этом на экране отобразится окно, которое показано на рис. 7.28.

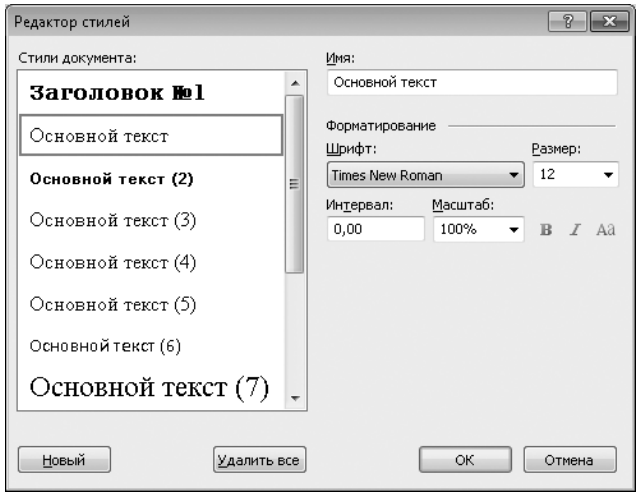


Рис. 7.28 ▾ Редактор стилей

В данном режиме осуществляются ввод, редактирование и удаление стилей. В левой части окна в поле **Стили документа** отображается перечень созданных ранее стилей.

Чтобы создать новый стиль, нажмем кнопку **Новый**. В поле **Имя** предлагается название созданного стиля – **CustomStyle1** (рис. 7.29). При желании вы можете ввести в это поле произвольное название.

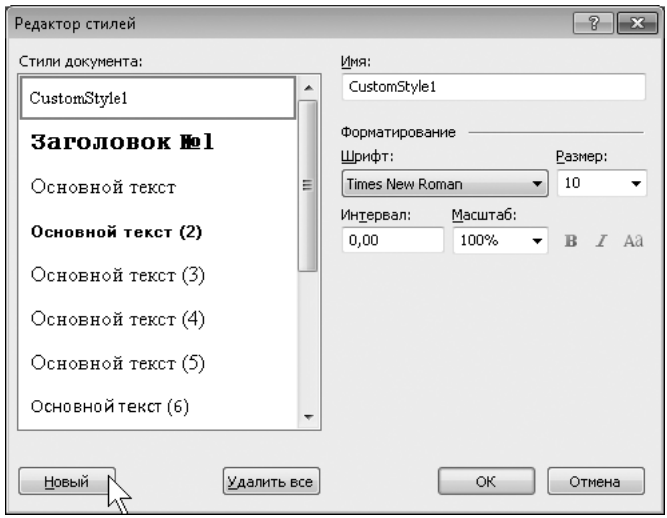


Рис. 7.29 ▾ Создание нового стиля

Теперь следует задать свойства стиля. В раскрывающемся списке **Шрифт** выберем тип шрифта – **Tahoma**, затем в поле размер укажем его размер – **14** и применим полужирное и курсивное начертания, нажав кнопки соответственно **B** и *I* (при подведении к этим кнопкам указателя мыши отображаются всплывающие подсказки). В полях **Интервал** и **Масштаб** оставим значения, предложенные программой по умолчанию. В конечном итоге настройки шрифта должны выглядеть так, как показано на рис. 7.30.

Теперь нажимаем в диалоге **Редактор стилей** кнопку **ОК**. Диалог закроется, а созданный стиль будет сохранен.

Предположим, что нам нужно применить созданный стиль к первому абзацу. Выделите этот фрагмент в окне **Текст** (рис. 7.31).

В раскрывающемся списке панели **Свойства текста** выбираем значение **CustomStyle1** (рис. 7.32).

Результат выполненных действий показан на рис. 7.33.

Таким образом, одним лишь действием мы применили к текстовому фрагменту сразу несколько действий: выделение полужирным начертанием, выделение курсивным начертанием, изменение вида шрифта, а также его размера.

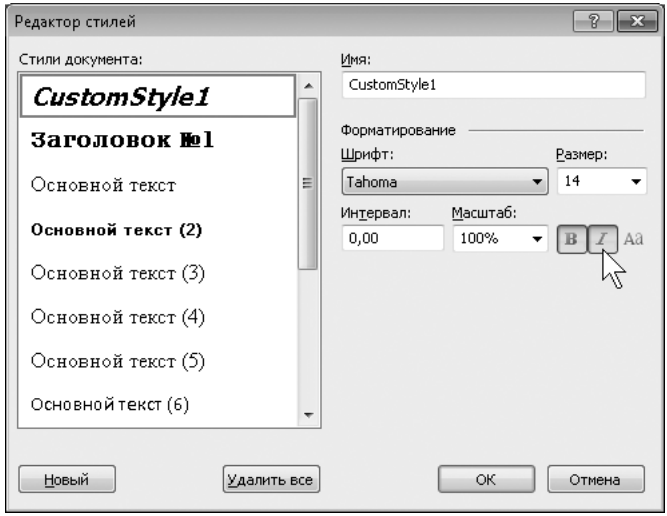


Рис. 7.30 ▾ Настройка параметров стиля

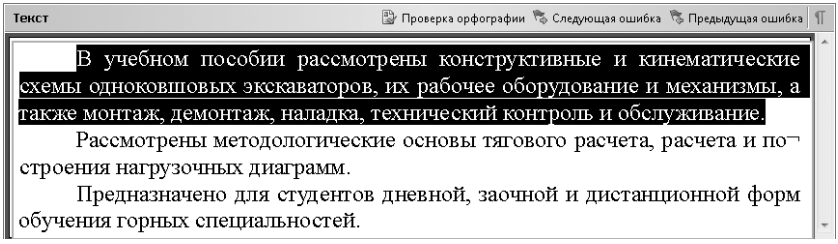


Рис. 7.31 ▾ Текст выделен для применения стиля

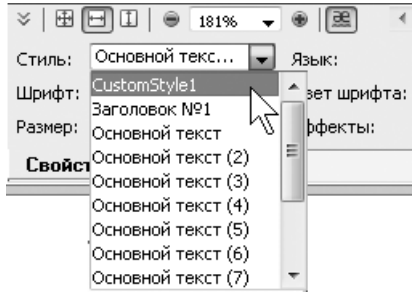


Рис. 7.32 ▾ Выбор пользовательского стиля

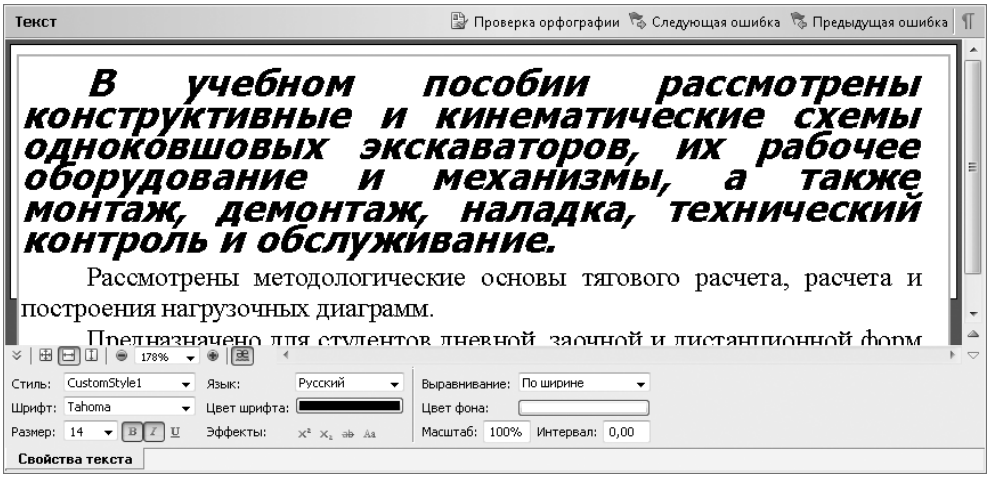


Рис. 7.33 ▾ Результат применения пользовательского стиля

Окончательная обработка распознанного документа в программе Microsoft Word

Несмотря на широкие функциональные возможности по обработке распознанных текстов, реализованные в FineReader, некоторые правки удобнее выполнять в программе Word. Это объясняется просто: ведь Word является специализированным текстовым редактором, в то время как главной задачей FineReader является распознавание текстов, а последующая обработка рассматривается лишь как дополнительная опция.

Пример избавления импортированного в Word документа от стилей FineReader

В предыдущем разделе мы научились создавать и применять к распознанному тексту пользовательские стили и смогли убедиться в том, что это очень удобно и практично. Однако бывает так, что при сохранении документа в Word это форматирование становится ненужным: например, из FineReader текст был выведен на печать, и стили стали не нужны, поскольку последующая обработка документа будет производиться в Word. Чтобы в документе Word не использовались стили, примененные к этому документу в FineReader стилей, достаточно выбрать соответствующий способ сохранения или передачи.

Для передачи или сохранения распознанного документа служит кнопка **Передать/Сохранить** на главной панели инструментов. Нажатие на треугольную стрелку справа от этой кнопки открывает список доступных действий: сохранения документа в один из поддерживаемых форматов или передачи его в другие приложения. Эти действия мы подробно рассмотрим в следующей гла-

ве, а пока остановимся только на передаче документа в Microsoft Word или сохранении его в формате документа Word.

В обоих случаях в раскрывающемся списке, расположенном правее кнопки **Передать/сохранить**, вы можете выбрать один из четырех способов сохранения или передачи. Чтобы текст документа был передан или сохранен в документ Microsoft Word без стилей и форматирования, использованных в FineReader, из раскрывающегося списка нужно выбрать значение **Простой текст** (рис. 7.34).

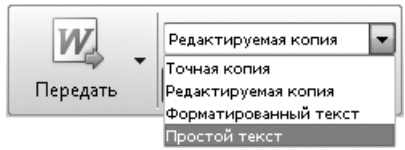


Рис. 7.34 ▾ Выбор способа сохранения документа

На рис. 7.35 показано, как будет выглядеть в окне Word переданный таким образом текст, к которому мы ранее применяли пользовательский стиль (см. рис. 7.33).

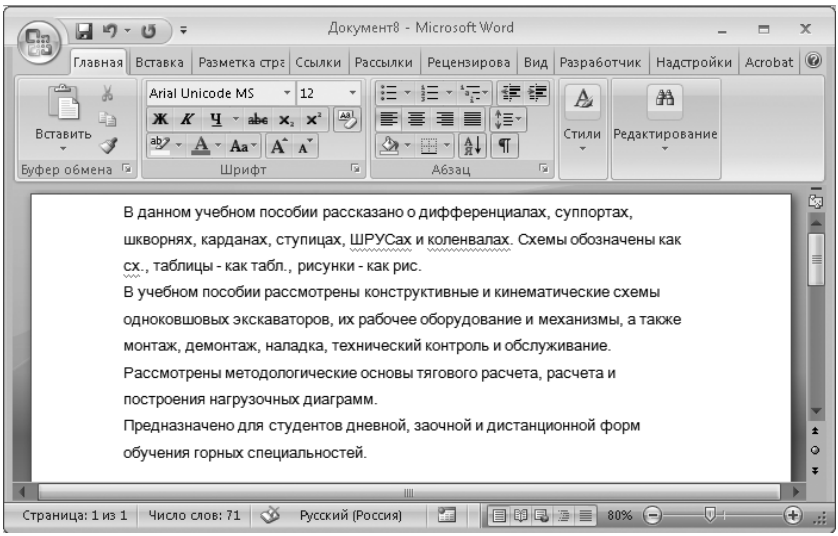


Рис. 7.35 ▾ Сохранение документа без форматирования

Как видно на рисунке, в программе Word ко всему тексту, независимо от того, как он выглядел в FineReader, применен один и тот же стиль, принятый в Microsoft Word по умолчанию. В Microsoft Word этот стиль носит название **Обычный**.

Однако такой способ не всегда является оптимальным. Бывают случаи, когда импортировать текст в Word нужно именно в том виде, как он выглядел в окне **Текст** программы FineReader, и только после этого отказываться от пользовательских стилей FineReader.

ПРИМЕЧАНИЕ

Пользовательские стили, созданные в FineReader, автоматически экспортируются в документ Word, за исключением случаев, когда в качестве способа сохранения выбрано значение ***Простой текст*** *(см. рис. 7.34).*

Рассмотрим, каким образом можно решить эту задачу на примере уже знакомого нам документа.

Вначале выберем способ сохранения документа **Редактируемая копия**, чтобы экспортировать его вместе с пользовательскими стилями, а также убедимся в том, что в качестве типа документа выбрано значение **Документ Microsoft Word** (рис. 7.36).

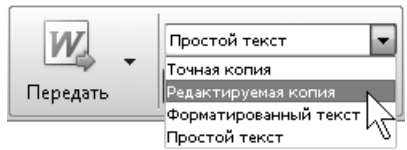


Рис. 7.36 ▾ Выбор способа сохранения **Редактируемая копия**

Затем нажимаем кнопку **Сохранить** – в результате на экране откроется окно, изображенное на рис. 7.37.

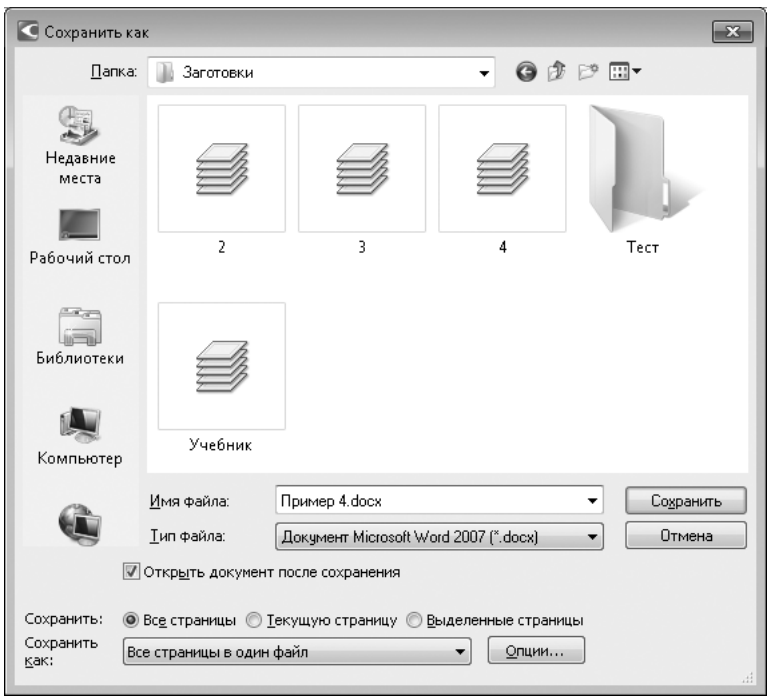


Рис. 7.37 ▾ Сохранение документа

В данном окне по обычным правилам указываем путь для сохранения и имя файла, после чего нажимаем кнопку **Сохранить**. Файл будет сохранен в указанном месте. Если в диалоге сохранения установлен флажок **Открыть документ**

после сохранения, на экране откроется окно Word, в котором будет представлен сохраненный текст (рис. 7.38).

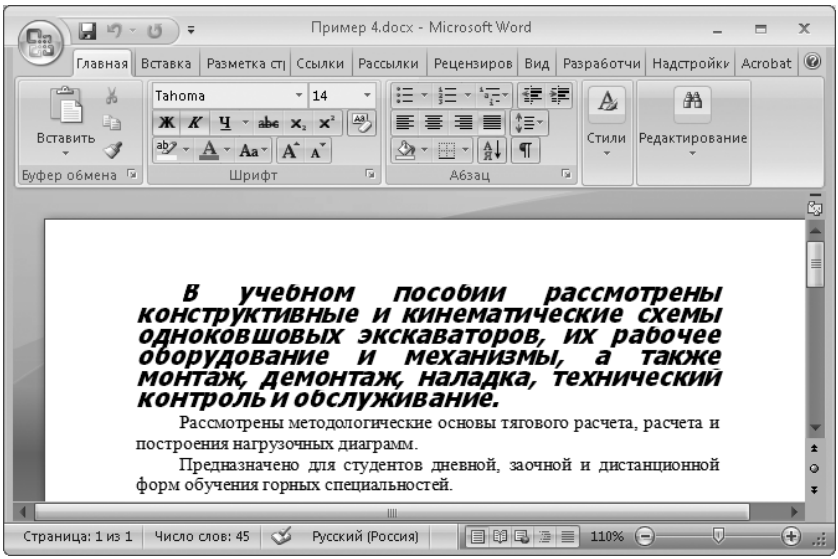


Рис. 7.38 ▾ Сохранение документа с форматированием

Теперь сравните рис. 7.35 и 7.38: отличается не только форматирование текста, но и примененный к нему стиль. Обратите внимание: на рис. 7.38 используется стиль **CustomStyle1**, а не **Обычный**, как на рис. 7.35. Этот стиль был импортирован из FineReader. Если мы установим курсор на первом абзаце текста, то в окне стилей (рис. 7.39) отобразится значение **CustomStyle1**, то есть имя пользовательского стиля, созданного в FineReader. Этот стиль можно применить в Word для оформления других фрагментов текста.

В Microsoft Word существуют три вида стилей:

- символьный стиль, определяющий форматирование символов: вид, размер и начертание шрифта. Такой стиль применяется к отдельным символам или словам и не влияет на форматирование абзаца. Другими словами, если вы уста-

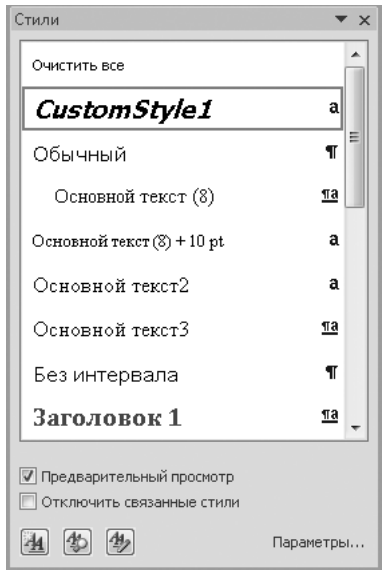


Рис. 7.39 ▾ Несколько стилей импортировано из FineReader

новите курсор внутри слова или выделите фрагмент текста и примените такой стиль, изменятся шрифт и начертание только символов этого слова или выделенного фрагмента; границы, отступы и выравнивание абзаца, в котором находится этот фрагмент, не изменятся;

- стиль абзаца, определяющий форматирование абзаца: отступы, границы, выравнивание, положение на странице. На форматирование символов эти стили не влияют. Таким стилем, например, является стиль **Обычный** в Microsoft Word;
- стили, сочетающие в себе символьный стиль и стиль абзаца. К таковым относится большинство встроенных стилей Microsoft Word. Если установить курсор в любом месте абзаца и применить такой стиль, изменится и форматирование всех символов (шрифт, начертание) внутри этого абзаца, и форматирование самого абзаца: границы, выравнивание и т. п.

Стили, создаваемые и используемые программой FineReader 10, являются смешанными, то есть сочетают в себе символьный стиль и стиль абзаца. Поэтому вы сможете свободно применять к любому фрагменту или слову импортированного текста «родные» стили Microsoft Word.

Если стили, импортированные из FineReader, больше не нужны и вы хотите удалить их, то это можно сделать двумя способами. Первый заключается в том, чтобы вначале полностью убрать все форматирование (либо со всего текста, либо с предварительно выделенного фрагмента). Чтобы убрать форматирование, выделите текст, а затем в окне стилей (см. рис. 7.39) выберите значение **Очистить все**. В результате текст будет отформатирован так, как это принято в Microsoft Word по умолчанию (рис. 7.40).

К абзацам применился стиль **Обычный**, а форматирование символов соответствует используемому в Microsoft Word по умолчанию (так называемый *основной шрифт абзаца*).

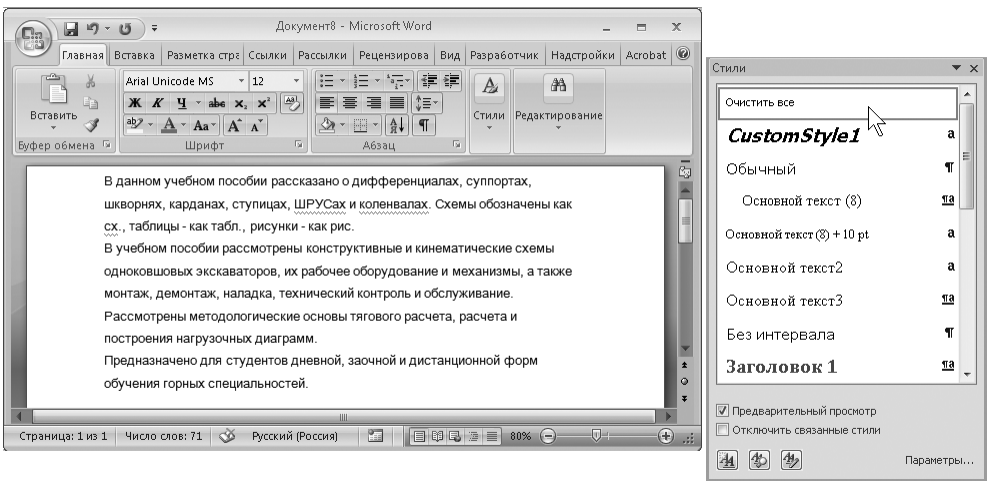


Рис. 7.40 ▾ Очистка формата

Другой вариант – полностью удалить «чужие» (то есть импортированные из FineReader) стили. В таком случае ко всем фрагментам, которые ранее были отформатированы с использованием удаленных стилей, будет применено форматирование, принятое в Microsoft Word по умолчанию. Рассмотрим, как это делается на практике.

В окне стилей (см. рис. 7.39) щелкните правой кнопкой мыши на названии стиля. Откроется контекстное меню стиля. Иначе вы можете выбрать стиль щелчком левой кнопки мыши и нажать на треугольную стрелку справа от названия стиля – откроется то же самое меню. Выберите в нем команду **Удалить (имя_стиля)** (рис. 7.41).

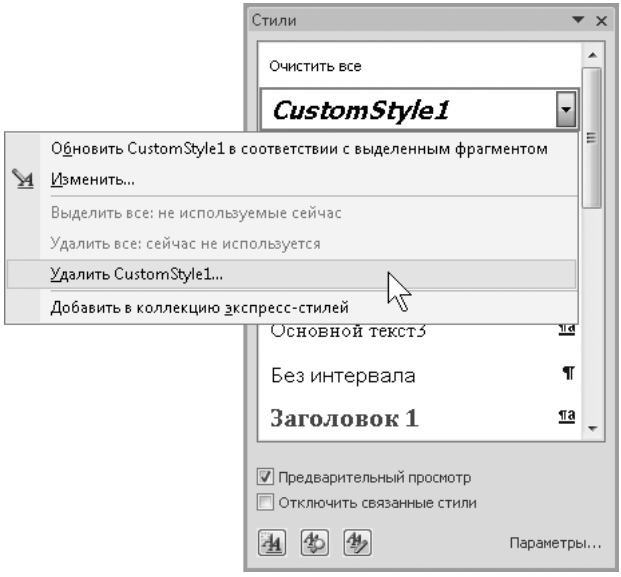


Рис. 7.41 ▾ Стили и форматирование

В ответ на появившийся запрос подтвердите, что вы хотите удалить стиль (рис. 7.42). Стиль, импортированный вместе с документом из FineReader, будет удален.

ПРИМЕЧАНИЕ

Удалять таким способом встроенные стили Word не получится – для них в контекстном меню команда ***Удалить*** *неактивна.*

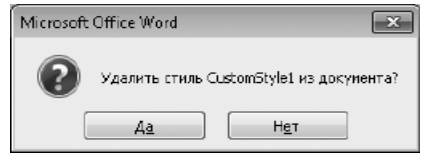


Рис. 7.42 ▾ Удаление импортированного стиля

Выполним эту операцию поочередно для всех стилей, импортированных из Fine-

Reader. По мере удаления стилей наш текст будет соответствующим образом изменяться. В конечном итоге он примет вид, как показано на рис. 7.43, а все импортированные списки исчезнут из меню стилей редактора Word.

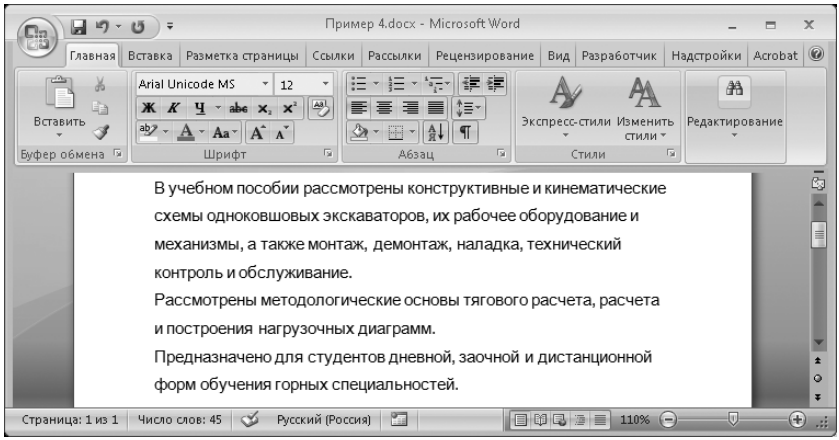


Рис. 7.43 ▾ Результат удаления импортированных стилей

Как видно на рисунке, наш текст полностью освобожден от любого форматирования – к нему применен стиль **Обычный**, который в Microsoft Word используется по умолчанию.

Пример обработки распознанного документа в Word

В этом разделе мы на конкретном примере покажем, как в редакторе Word выполняется обработка распознанных в FineReader документов. При этом мы будем рассматривать действия, которые невозможно выполнить в окне **Текст** программы FineReader, а именно удаление двойных и многократных пробелов, вставка символа, создание маркированного списка и настройка интервалов между абзацами.

Сохраняем наш документ в Word, выбрав режим **Редактируемая копия**. В окне Word выделим весь текст и уберем форматирование, выбрав в меню **Стили** пункт **Очистить формат**. Распознанный текст, с которым мы будем работать, передан в Microsoft Word для окончательной правки (рис. 7.44).

Очевидно, что текст распознан корректно. Единственное замечание – неравномерные промежутки между словами. Можно предположить, что это связано с наличием следующих подряд пробелов. Этот недостаток можно устранить и в окне **Текст**, удаляя лишние пробелы вручную, но намного удобнее делать это средствами Word.

Чтобы убедиться в существовании в распознанном тексте лишних пробелов, используем функцию отображения скрытых непечатаемых символов. Для этого включим в Microsoft Word режим отображения скрытых символов, нажав

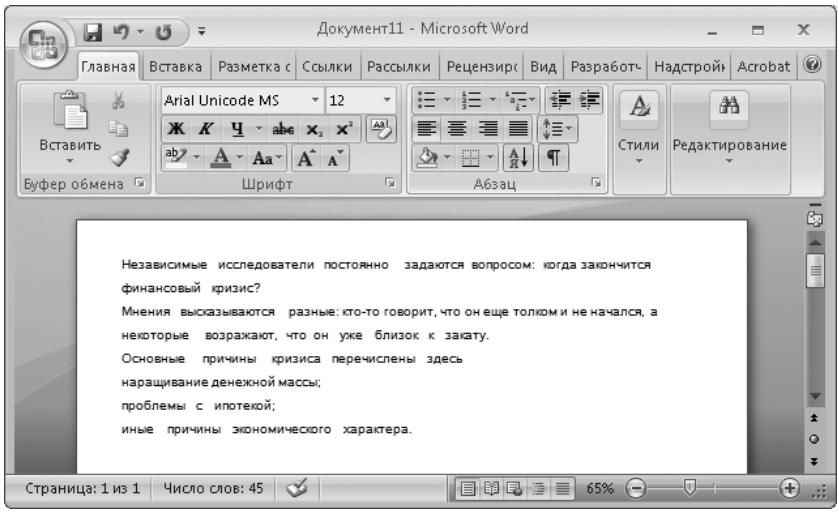


Рис. 7.44 ▾ Распознанный текст передан в Microsoft Word

на панели инструментов **Главная** в группе **Абзац** кнопку **Отобразить все знаки** (рис. 7.45). Название кнопки отображается в виде всплывающей подсказки при подведении к ней указателя мыши. В результате мы увидим на экране непечатаемые символы: пробелы, символы конца абзаца и др.

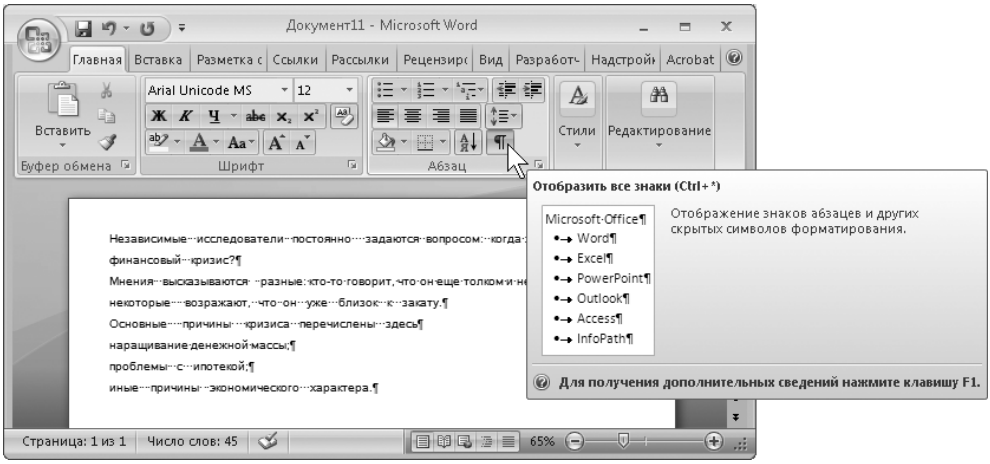


Рис. 7.45 ▾ Отображение скрытых символов

Наше предположение подтвердилось. Лишние пробелы могут появиться в распознанном тексте, поскольку программа FineReader при распознавании пробелов ориентируется на усредненную ширину символов в тексте. Если

в оригинале промежутки между словами значительно превышают эту величину (это следствие принудительного выравнивания текста по ширине), при распознавании в таких местах могут быть помещены два и более пробела.

Теперь начнем удалять из текста лишние пробелы. Можно делать это и вручную, клавишами **Del** или **Backspace**, но эффективнее обратиться к функции поиска и замены. Вызовите диалог поиска и замены: меню **Главная ➢ Редактирование ➢ Заменить**. Откроется диалог **Найти и заменить** (рис. 7.46).

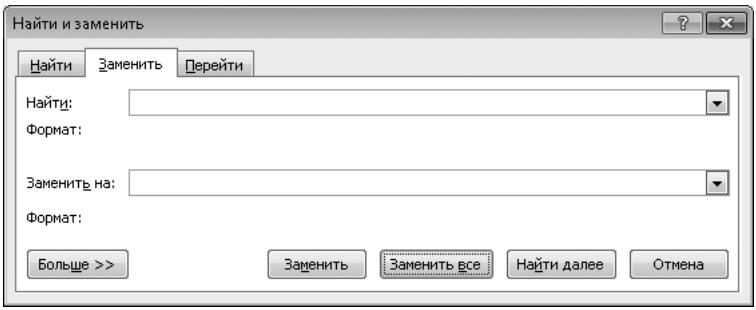


Рис. 7.46 ▾ Диалог **Найти и заменить**

Введите в поле **Найти:** два пробела подряд, а в поле **Заменить на:** один пробел. Нажмите кнопку **Заменить все**. В результате каждые два идущих подряд пробела будут заменены на один.

Программа Microsoft Word сообщит вам о том, сколько замен произведено в документе. Закройте сообщение и вновь нажмите в диалоге **Найти и заменить** кнопку **Заменить все**. Повторяйте эту операцию до тех пор, пока вы не увидите в очередном сообщении, что во всем документе произведено 0 замен. Иначе говоря, в документе не осталось двойных пробелов.

Удалив пробелы, отключим отображение скрытого текста и вставим в текст (после слова **здесь**) символ из коллекции Word. Для этого в главном меню программы выберем команду **Вставка ➢ Символы ➢ Символ**, затем в открывшемся окне на вкладке **Символы** в поле **Шрифт** выберем значение **Symbol** и щелчком мыши укажем в списке символ, как показано на рис. 7.47.

Затем нажмем кнопку **Вставить** – сразу после этого кнопка **Отмена** будет называться **Закрыть**. Нажимаем эту кнопку – результат выполненных действий показан на рис. 7.48.

Стрелка вниз, которую мы вставили, указывает на находящееся ниже перечисление, что логично вытекает из контекста документа. Для наглядности это перечисление оформим в виде маркированного списка. Для этого выделим весь текст, начиная со слова **наращивание**, и нажмем в инструментальной панели **Главная ➢ Абзац** кнопку **Маркеры**. Результат показан на рис. 7.49.

И последнее, что нам осталось сделать, – это настроить интервалы между абзацами. Это повысит наглядность и читабельность текста, сделает его более

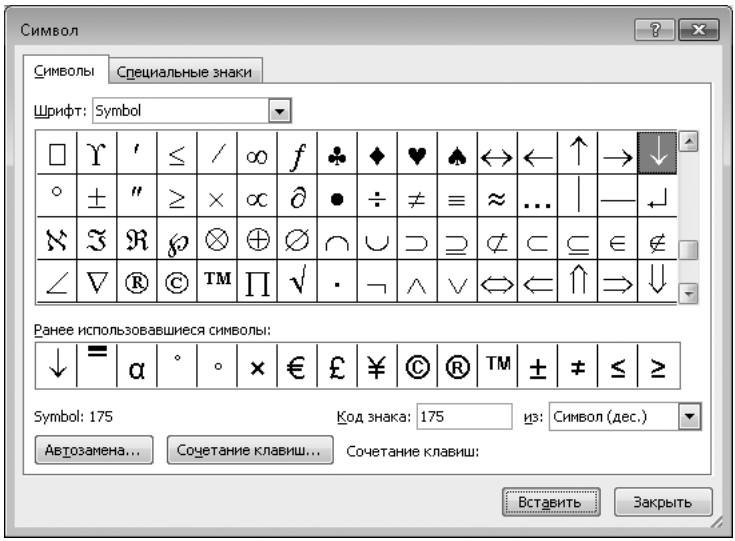


Рис. 7.47 ▾ Выбор символа для вставки в документ

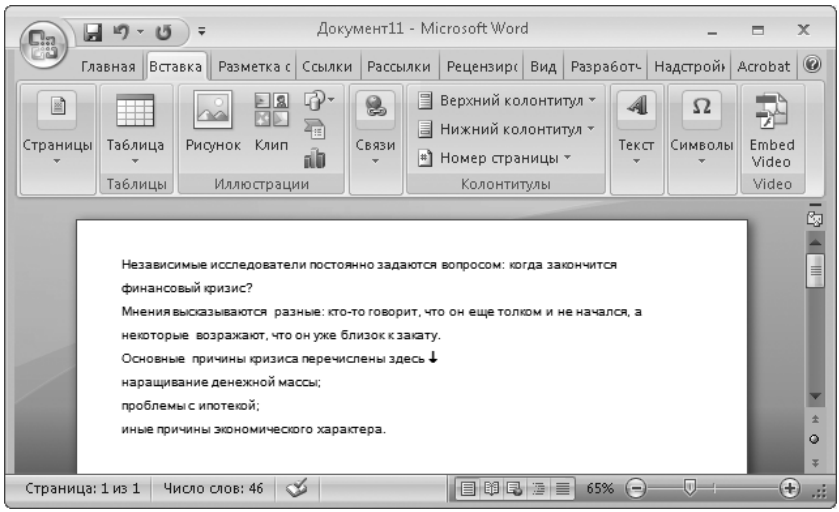


Рис. 7.48 ▾ Вставка символа

эргономичным. Для этого выделим весь текст (сочетание клавиш **Ctrl+A**), выполним команду главного меню **Формат ➢ Абзац** и в открывшемся окне на вкладке **Отступы и интервалы** в группе **Интервал** в поле **после:** введем значение **3 пт**. Остальные параметры оставим без изменений (рис. 7.50).

После нажатия в данном окне кнопки **ОК** наш текст примет вид, как показано на рис. 7.51.

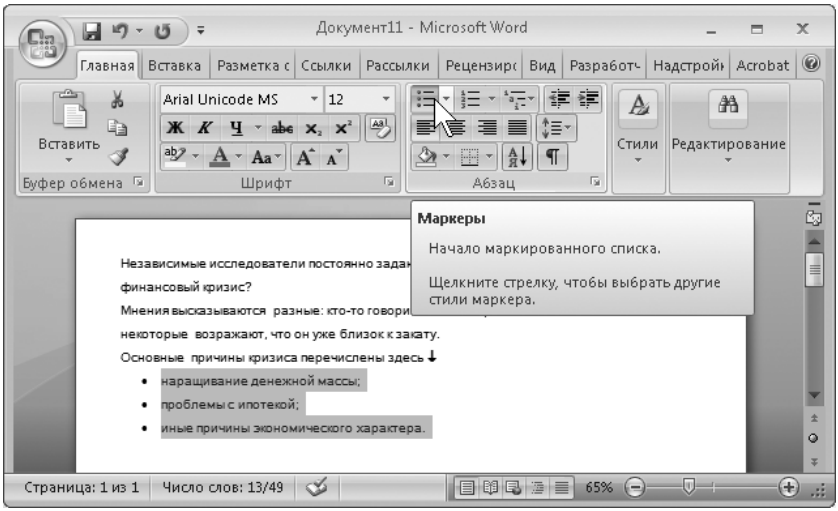


Рис. 7.49 ▾ Создание маркированного списка

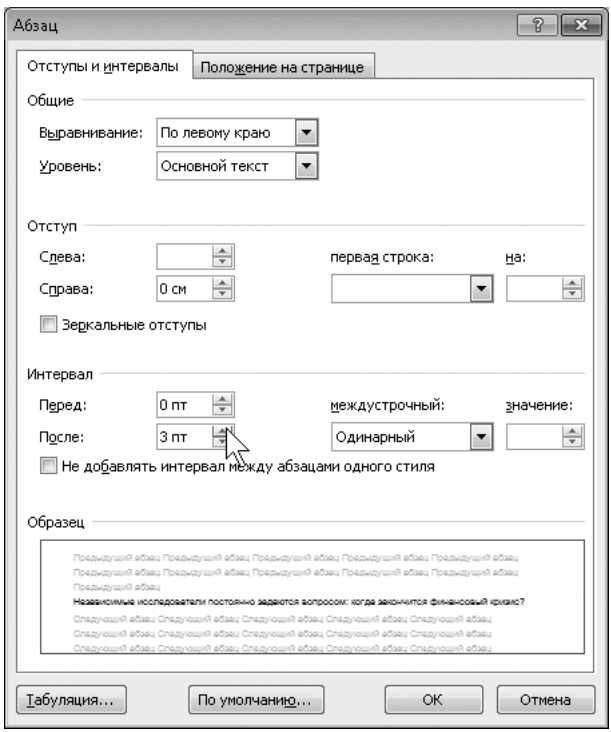


Рис. 7.50 ▾ Настройка интервалов между абзацами

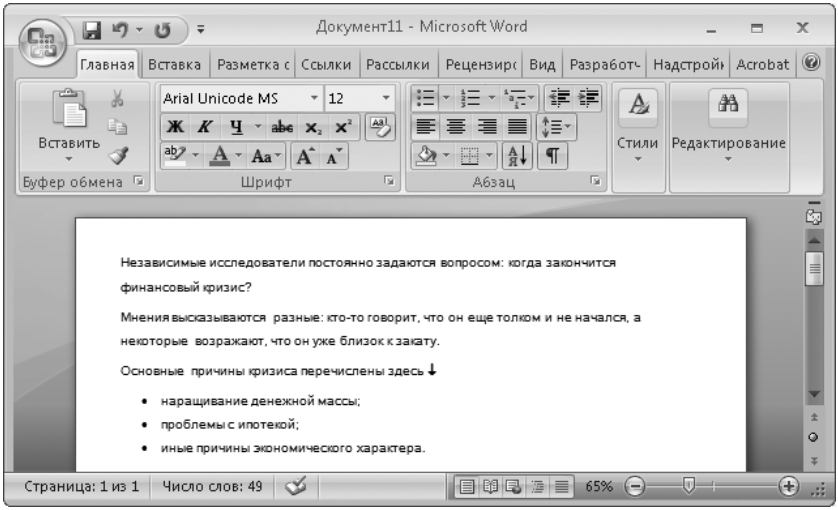


Рис. 7.51 ▾ Результат обработки документа в редакторе Word

Очевидно, что в окне **Текст** нам не удалось бы оформить текст так, как мы это сделали в Word, поскольку соответствующие инструменты в программе FineReader не предусмотрены.

Корректировка таблиц

Если распознанный документ содержит одну или несколько таблиц, вы можете откорректировать ее в окне **Текст**. Здесь мы расскажем о том, какие действия можно выполнять с таблицами в программе FineReader.

Настройка панели инструментов для работы с таблицами

С любой распознанной таблицей вы можете в окне **Текст** выполнять следующие действия: объединение ячеек, объединение строк и разбиение ранее объединенных ячеек, а также удалять содержимое ячеек. Последняя операция выполняется нажатием клавиши **Delete** (предварительно следует выделить все ячейки, которые предполагается очистить), а первые три осуществляются с помощью соответствующих кнопок инструментальной панели.

Однако по умолчанию отображение этих кнопок отключено, поэтому нам нужно добавить их на панель инструментов **Быстрый доступ**. Для этого выполним команду главного меню **Сервис** ➢ **Настройка**, в открывшемся окне перейдем на вкладку **Панели инструментов** и выберем в поле **Категории** значение **Правка** (рис. 7.52).

Затем в левой части окна в поле **Команды** щелчком мыши выделим значение **Разбить ячейки** и нажмем кнопку со стрелками вправо, которая находится

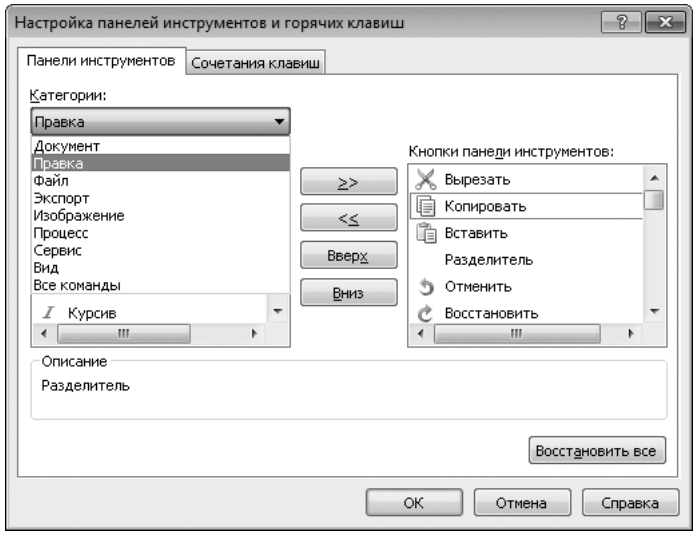


Рис. 7.52 ▾ Режим настройки панелей инструментов

в центральной части окна. Далее проделаем то же самое с позициями **Объединить ячейки** и **Объединить строки таблицы** (рис. 7.53). Если все сделано правильно, то выбранные позиции переместятся в поле **Кнопки панели инструментов**.

Закройте диалог **Настройка панелей инструментов и горящих клавиш**, нажав в нем кнопку **OK**. При необходимости сделайте панель **Быстрый доступ**

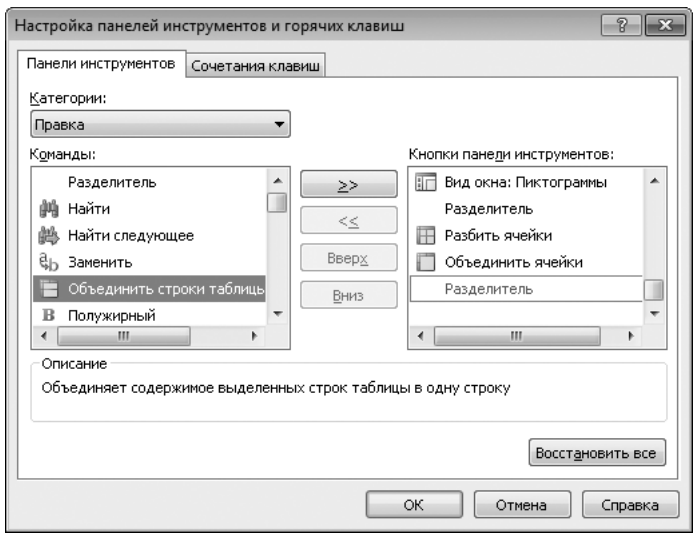


Рис. 7.53 ▾ Помещаем кнопку на панель инструментов

видимой: выберите команду меню **Вид ➢ Панели инструментов ➢ Быстрый доступ.** Инструментальная панель **Быстрый доступ** будет выглядеть так, как показано на рис. 7.54 (на ней появятся три новые кнопки).

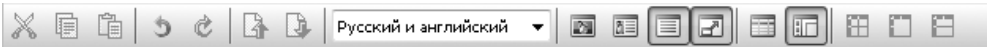


Рис. 7.54 ▾ Панель **Быстрый доступ** с добавленными кнопками

На рисунке видно, что эти кнопки пока неактивны (притушены). Но как только в таблице будут выделены хотя бы две ячейки, то кнопки **Объединить ячейки** и **Объединить строки** сразу станут активными. Кнопка **Разбить ячейки** становится доступной тогда, когда курсор находится в ячейке, которая образовалась в результате объединения других ячеек.

Пример корректировки таблицы

Корректировка таблиц может потребоваться, если в распознанном документе расположение ячеек таблицы отличается от того, которое вы хотели бы видеть. Например, в выходном документе объединены ячейки, которые по логике построения таблицы объединять не надо, либо, наоборот, отдельные ячейки разбиты на несколько частей.

Как правило, корректировку таблиц целесообразно выполнять еще на этапе анализа, в окне **Изображение**. Такие операции рассмотрены в предыдущей главе. Можно вносить изменения и в уже распознанные таблицы, наряду с проверкой орфографии и исправлением ошибок. Здесь мы на конкретном примере продемонстрируем, какие действия можно выполнять с распознанными таблицами в окне **Текст**.

Предположим, что мы распознали таблицу, изображенную на рис. 7.55. Хотя в целом таблица распознана правильно, желательно для наглядности объединить две ячейки в заголовке таблицы, а для удобства редактирования – объединить некоторые строки в самой таблице.

В окне **Текст** выделите в таблице две ячейки – так, как показано на рис. 7.56. На панели инструментов **Быстрый доступ** стала активна кнопка **Объединить ячейки**. Нажмем кнопку **Объединить ячейки** – выделенные ячейки объединятся.

В результате две выделенные ячейки объединены в одну. Теперь вернем таблице исходный вид: для этого установим курсор в образовавшуюся большую ячейку и нажмем кнопку **Разбить ячейки**. Ячейка будет разделена на две, которые существовали до выполнения операции объединения.

ПРИМЕЧАНИЕ

*Команда **Разбить ячейки** применима только к ячейкам, образовавшимся в результате слияния ячеек с помощью команды **Объединить ячейки**.*

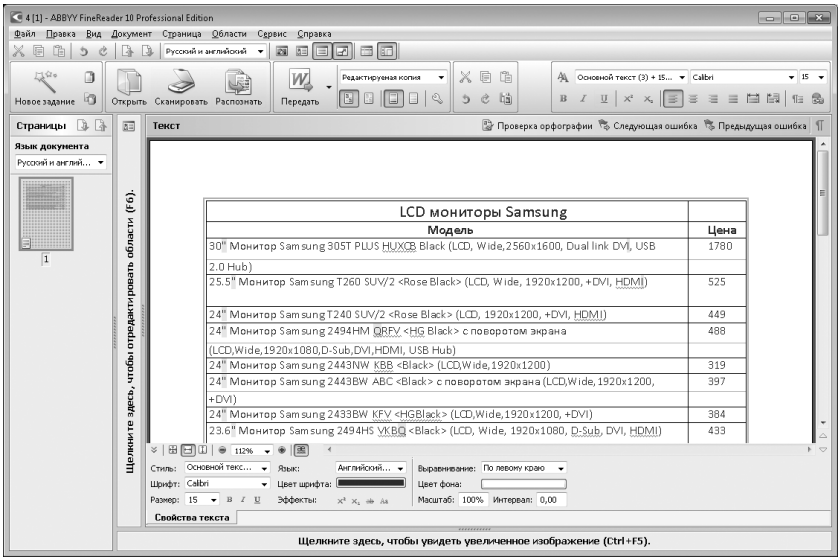
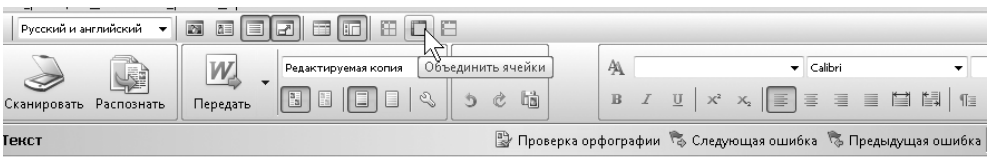


Рис. 7.55 ▾ Распознанная таблица для обработки



LCD мониторы Samsung	
Модель	**Цена**
30" Монитор Samsung 305T PLUS HUXCB Black (LCD, Wide,2560x1600, Dual link DVI, USB 2.0 Hub)	1780
25.5" Монитор Samsung T260 SUV/2 <Rose Black> (LCD, Wide, 1920x1200, +DVI, HDMI)	525

Рис. 7.56 ▾ Объединение ячеек

Чтобы объединить строки таблицы, выделите эти строки и нажмите кнопку **Объединить строки** (рис. 7.57).

В результате объединяются только ячейки, расположенные одна под другой: из двух или более выделенных строк таблицы образуется одна. В то же время деление на столбцы сохраняется.

Помните, что объединенные строки, в отличие от ячеек, не разъединяются с помощью кнопки **Разбить ячейки**. В данном случае для возврата таблицы к предыдущему состоянию используйте на инструментальной панели кнопку **Отменить** или команду главного меню **Правка ➢ Отменить**, которая вызывается также нажатием комбинации клавиш **Ctrl+Z**.

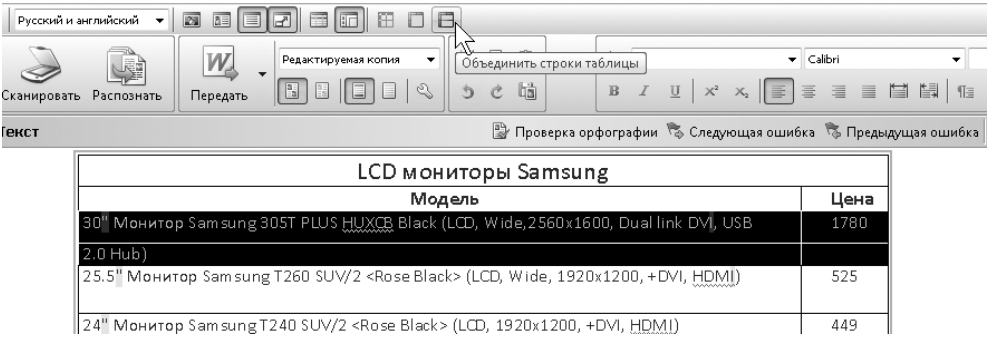


Рис. 7.57 ▾ Объединение строк таблицы

Резюме

Распознать документ – это иногда лишь половина дела. Порой для того, чтобы он принял приемлемый вид, приходится дополнительно с ним поработать: проверить корректность распознавания, что-то удалить, что-то добавить, отформатировать и т. д.

Прочитав эту главу, вы получили необходимые знания, которые позволят вам выполнять все требуемые действия по проверке и редактированию распознанных текстов. Теперь вы знаете, что нужно делать, если какие-то слова или символы распознаны неуверенно, как красиво отформатировать документ буквально несколькими щелчками мыши и что делать, если возможностей FineReader недостаточно для приведения документа к требуемому виду.

Как вам уже известно, из FineReader текст можно импортировать в Word для последующего редактирования или просто для хранения в указанном формате. Однако это далеко не единственный вариант сохранения распознанных документов, реализованный в программе, и об этом мы подробно поговорим в следующей главе, которая так и называется – «Сохранение распознанного документа».

Глава 8 Сохранение распознанного документа

После того как распознанный документ отобразился в окне **Текст**, вы можете вывести его на печать с помощью команд подменю **Файл ➤ Печать**. Напомним, что для этого также можно использовать комбинации клавиш **Ctrl+Alt+P** (документ будет распечатан как изображение) или **Ctrl+P** (как текст).

Однако иногда возникает необходимость сохранить распознанный документ в отдельный файл, для того чтобы с ним можно было работать и в будущем, или передать его в какое-либо приложение. В программе FineReader реализованы широкие функциональные возможности по сохранению распознанных документов и передаче их в другие приложения. О том, как это делается, мы расскажем в данной главе.

Чтобы сохранить документ в файл или передать его в приложение, вы можете воспользоваться одной из команд меню **Файл** либо кнопкой на главной панели инструментов. Поскольку вариантов сохранения и передачи много, в меню **Файл** команды организованы в виде групп, или вложенных меню (рис. 8.1).

На главной панели инструментов расположена кнопка **Передать/Сохранить**. Эта кнопка настраиваемая: при нажатии на стрелку справа от кнопки открывается меню (рис. 8.2). В нем вы можете выбрать действие, которое будет выполняться при нажатии кнопки. В зависимости от выбранного действия изменяются название кнопки и изображенная на ней пиктограмма.

Кроме того, правее кнопки расположен раскрывающийся список для выбора конкретного способа сохранения оформления (рис. 7.34). Для каждого из форматов или приложений, в которые будет сохраняться (передаваться) распознанный документ, доступен определенный набор способов. Например, до-

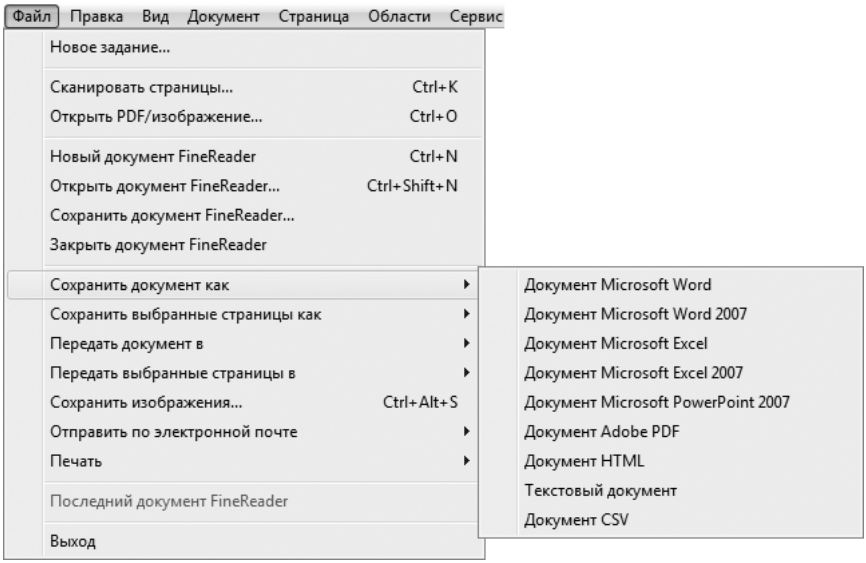


Рис. 8.1 ▾ Сохранение/передача документа с помощью меню

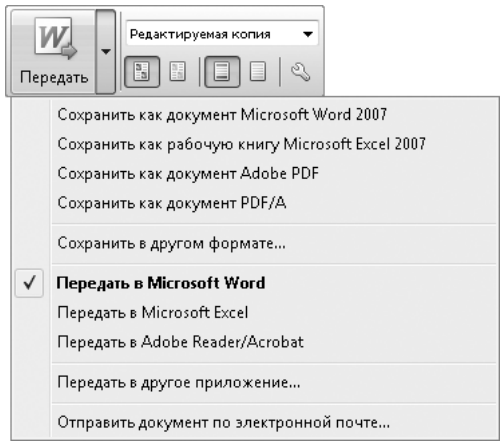


Рис. 8.2 ▾ Передача документа в Microsoft Word

кумент может быть передан в Microsoft Word или сохранен в формат Microsoft Word как *точная копия*, *редактируемая копия*, *форматированный текст* или *простой текст*. В Microsoft Excel тот же документ может быть передан или сохранен только как *форматированный текст*, а передать в Adobe Reader или сохранить в формат PDF его можно только как *точную копию*.

Передача документа в приложение

При передаче распознанного документа в приложение он не будет сохранен на жестком диске. Но вы сможете сделать это после передачи – уже из соответствующей программы. В FineReader предусмотрена возможность передачи распознанного документа в текстовый редактор Word, табличный редактор Excel, программу Adobe Reader/Acrobat и браузер. Непосредственно из FineReader 10 распознанный документ можно сохранить в одном из поддерживаемых форматов и прикрепить к сообщению электронной почты. Кроме того, содержимое документа может быть передано в буфер обмена.

СОВЕТ

Копирование распознанного текста в буфер обмена удобно, например, в тех случаях, когда вам нужно передать его в приложение, в которое FineReader не может этого сделать. В таком случае вы просто запускаете нужную программу и вставляете в нее текст из буфера обмена.

Далее мы на конкретных примерах рассмотрим, каким образом осуществляется его передача из окна **Текст** программы FineReader в другие приложения.

Пример передачи распознанного документа в Microsoft Word

Чтобы передать весь распознанный документ в текстовый редактор Microsoft Word без сохранения его на жесткий диск, нажмите стрелку справа от кнопки **Передать/Сохранить** и в открывшемся меню (рис. 8.2) выберите вариант **Передать в Microsoft Word**. На кнопке будут отображены значок Microsoft Word и слово **Передать**. В раскрывающемся списке (рис. 7.34) выберите требуемый режим сохранения оформления. Нажмите на кнопку **Передать/Сохранить**, и документ будет передан в Microsoft Word указанным способом, например как редактируемая копия.

Также вы можете воспользоваться командами меню. Выберите в меню FineReader команду **Файл ➢ Передать в ➢ Microsoft Word**. После этого на экране автоматически откроется окно Word, в котором будет представлен текст документа.

Обратите внимание, что документ передан тем способом, который был выбран в раскрывающемся списке рядом с кнопкой **Передать/Сохранить**, например как редактируемая копия или как простой текст. Непосредственно из меню вы не можете указать способ передачи: он задается либо с помощью раскрывающегося списка, либо на соответствующей вкладке диалога **Опции**.

Если распознанный документ состоит из нескольких страниц, вы можете передавать его в Word не полностью, а частично. Для этого в окне **Страницы** нужно выбрать одну или несколько страниц, после чего выполнить команду главного меню **Файл ➢ Передать выбранные страницы в ➢ Microsoft Word**.

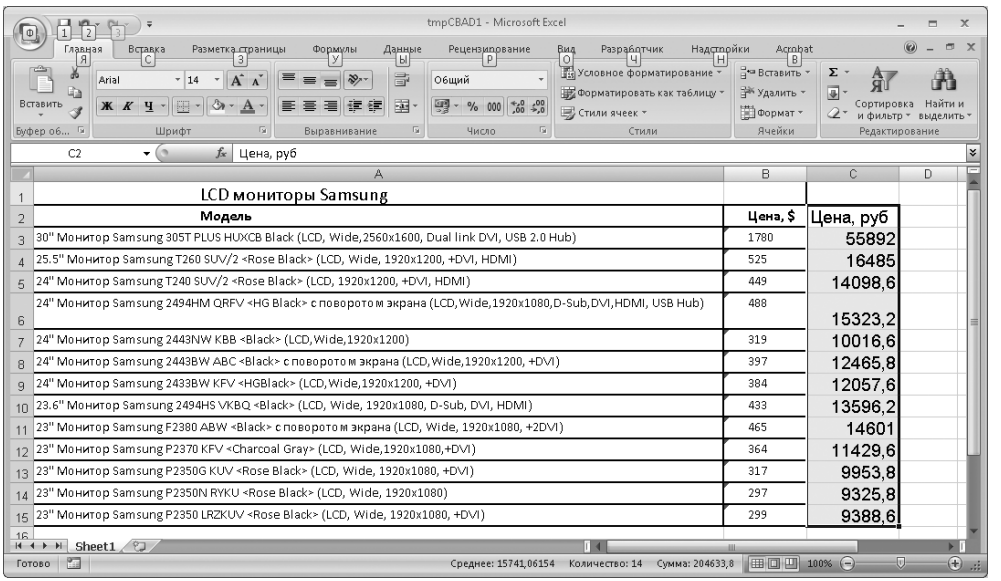
Выбранные страницы откроются в окне редактора Word, где их можно посмотреть, отредактировать и сохранить на жесткий диск.

Для сохранения переданного в Word текста на жесткий диск используйте команду главного меню **Файл ➢ Сохранить**, вызываемую также нажатием комбинации клавиш **Shift+F12**.

Пример передачи распознанного документа в Microsoft Excel

Теперь передадим распознанный текст (см. рис. 8.1) в табличный редактор Microsoft Excel. Для этого в главном меню выберем команду **Файл ➢ Передать в ➢ Microsoft Excel**. При активизации данной команды на экране откроется окно Excel, в котором будет представлен переданный документ.

В Microsoft Excel вы можете выполнять любые действия с документом: форматировать, выводить на печать, сохранять на жесткий диск и т. д. Самое важное, что электронная таблица позволяет добавлять колонки, строки, вставлять формулы и выполнять расчеты. Например, в распознанный прайс-лист легко вставить формулу, по которой цена пересчитывается из долларов в рубли (рис. 8.3).



	A	B	C	D
1	LCD мониторы Samsung			
2	**Модель**	**Цена, $**	**Цена, руб**	
3	30" Монитор Samsung 305T PLUS HUXCB Black (LCD, Wide,2560x1600, Dual link DVI, USB 2.0 Hub)	1780	55892	
4	25.5" Монитор Samsung T260 SUV/2 <Rose Black> (LCD, Wide, 1920x1200, +DVI, HDMI)	525	16485	
5	24" Монитор Samsung T240 SUV/2 <Rose Black> (LCD, 1920x1200, +DVI, HDMI)	449	14098,6	
6	24" Монитор Samsung 2494HM QRFV <HG Black> с поворотом экрана (LCD,Wide,1920x1080,D-Sub,DVI,HDMI, USB Hub)	488	15323,2	
7	24" Монитор Samsung 2443NW KBB <Black> (LCD,Wide,1920x1200)	319	10016,6	
8	24" Монитор Samsung 2443BW ABC <Black> с поворотом экрана (LCD,Wide,1920x1200, +DVI)	397	12465,8	
9	24" Монитор Samsung 2433BW KFV <HGBlack> (LCD,Wide,1920x1200, +DVI)	384	12057,6	
10	23.6" Монитор Samsung 2494HS VKBQ <Black> (LCD, Wide, 1920x1080, D-Sub, DVI, HDMI)	433	13596,2	
11	23" Монитор Samsung F2380 ABW <Black> с поворотом экрана (LCD, Wide, 1920x1080, +2DVI)	465	14601	
12	23" Монитор Samsung P2370 KFV <Charcoal Gray> (LCD, Wide,1920x1080,+DVI)	364	11429,6	
13	23" Монитор Samsung P2350G KUV <Rose Black> (LCD, Wide, 1920x1080, +DVI)	317	9953,8	
14	23" Монитор Samsung P2350N RVKU <Rose Black> (LCD, Wide, 1920x1080)	297	9325,8	
15	23" Монитор Samsung P2350 LRZKUV <Rose Black> (LCD, Wide, 1920x1080, +DVI)	299	9388,6	
16				

Рис. 8.3 ▾ Передача документа в Microsoft Excel

Вы можете передать в Excel не весь документ, а только некоторые его страницы. Для этого отметьте эти страницы в окне **Страницы** и выполните команду главного меню **Файл ➢ Передать выбранные страницы в ➢ Microsoft Excel**.

Пример передачи распознанного документа в Adobe Reader

Далее рассмотрим, каким образом в FineReader осуществляется передача распознанного текста в приложение Adobe Reader.

Данная возможность позволяет выполнить следующую процедуру: открыть и распознать документ (например, из того же PDF-файла), внести в окне **Текст** в него требуемые изменения и вновь сохранить в исходном формате. Таким образом, посторонние могут и не догадаться о том, что данный документ был модифицирован (особенно если впоследствии сохранить его под тем же именем).

Чтобы передать распознанный документ в программу Adobe Reader, выполним в главном меню команду **Файл ➢ Передать в ➢ Adobe Reader/Acrobat**. В результате на экране откроется окно Adobe Reader, содержащее распознанный текст.

Как известно, в данной программе никакое редактирование документов невозможно. Что касается сохранения, то вы можете сохранить документ как в формате pdf (рис. 8.4), так и в обычном текстовом формате. В первом случае используйте команду главного меню **Файл ➢ Сохранить копию** (она вызывается также нажатием комбинации клавиш **Ctrl**+**Shift**+**S**), во втором – команду

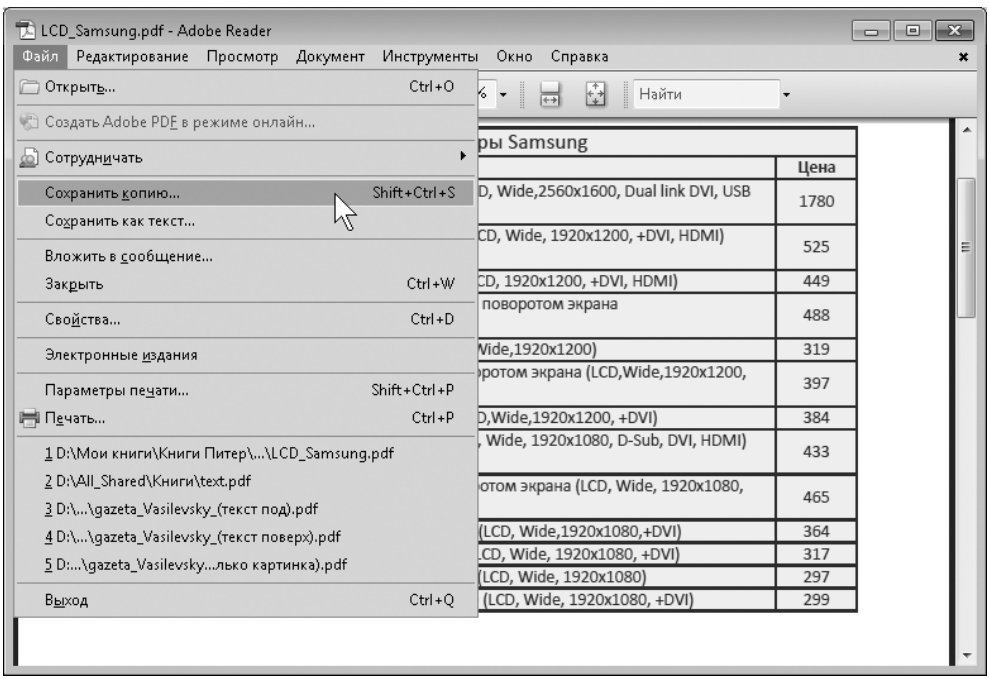


Рис. 8.4 ▾ Передача документа в программу Adobe Reader

Файл ➢ Сохранить как текст. В обоих случаях на экране откроется окно, в котором нужно будет указать путь для сохранения и имя файла.

Чтобы передать в Adobe Reader не весь документ, а лишь какую-то его часть, выберите в окне **Страницы** требуемые страницы и выполните команду главного меню **Файл ➢ Передать выбранные страницы в ➢ Adobe Reader/Acrobat**.

Пример передачи распознанного документа в веб-браузер

Возможности FineReader предусматривают передачу распознанного документа в используемый по умолчанию веб-браузер. В нашем примере таковым является Internet Explorer версии 7.0.

Чтобы передать наш документ в веб-браузер, выполним команду главного меню **Файл ➢ Передать в ➢ Web-браузер**. Результат показан на рис. 8.5.

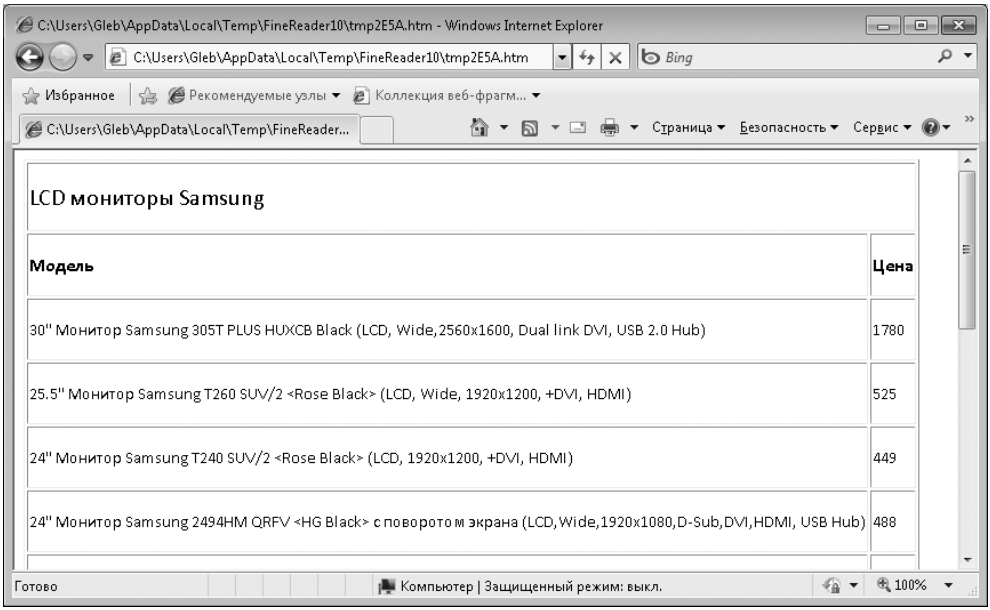


Рис. 8.5 ▾ Передача документа в веб-браузер

Если необходимо сохранить переданный документ в соответствующем формате, выберите в главном меню веб-браузера команду **Файл ➢ Сохранить как** и в открывшемся окне укажите путь для сохранения и имя файла.

Для выборочной передачи документа укажите требуемые страницы в окне **Страницы** и выполните команду главного меню **Файл ➢ Передать выбранные страницы в ➢ Web-браузер**.

Сохранение документа в файл

FineReader позволяет сохранять распознанные документы в файлы разных форматов: текстовый, doc-файл, html-файл и др. Однако, перед тем как сохранять документ в выбранном формате, рекомендуется просмотреть и, при необходимости, отредактировать соответствующие параметры настройки.

Поэтому вначале мы познакомимся с данными параметрами, а после этого на конкретных примерах посмотрим, как осуществляется сохранение распознанных документов в разных форматах.

Настройка параметров сохранения

Чтобы перейти к настройкам параметров сохранения, выберите в главном меню команду **Сервис ➢ Опции** (данная команда вызывается также нажатием комбинации клавиш **Ctrl+Shift+O**) и в открывшемся окне перейдите на вкладку **Сохранить** (рис. 8.6).

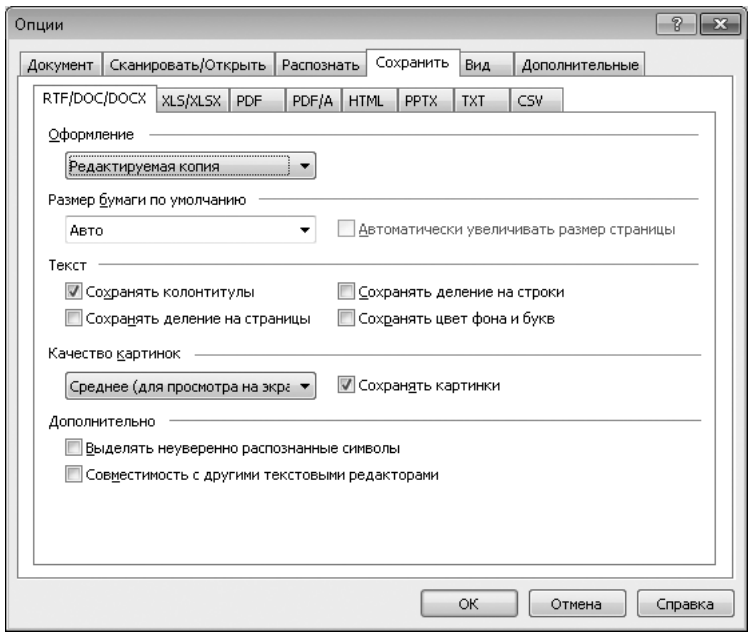


Рис. 8.6 ▾ Настройка параметров сохранения, вкладка **RTF/DOC/DOCX**

Данная вкладка включает в себя еще целый ряд подчиненных вкладок, каждая из которых предназначена для настройки параметров сохранения распознанных документов в соответствующем формате.

Настройка сохранения в формате Word осуществляется на вкладке **RTF/DOC/DOCX** (см. рис. 8.6). В поле **Оформление** из раскрывающегося списка выбирается режим оформления, который будет предлагаться по умолчанию при сохранении документа. Этот режим вы можете в любое время изменить с помощью раскрывающегося списка на главной панели инструментов.

ПРИМЕЧАНИЕ

Более подробно о режимах сохранения документов мы расскажем ниже, когда на конкретном примере будем рассматривать процесс сохранения документа в формате Word.

В поле **Размер бумаги по умолчанию** вы можете указать размер бумаги, который будет использоваться при сохранении документа. По умолчанию в данном поле предлагается значение **Авто**. При выборе любого другого значения становится доступным флажок **Автоматически увеличивать размер страницы**; если он установлен, то размер бумаги будет при необходимости увеличиваться автоматически. Вы можете выполнить тонкую настройку размера бумаги – это бывает полезно при работе с нестандартными документами. Для этого в данном поле выберите значение **Пользовательский размер бумаги**, после чего в открывшемся окне выполните требуемые настройки.

В области **Текст** объединено несколько флажков, с помощью которых можно указать параметры сохранения текста. Если установлен флажок **Сохранять колонтитулы** (это значение предлагается по умолчанию), то документ будет сохранен вместе с колонтитулами. Если установлены флажки **Сохранять деление на страницы** и **Сохранять деление на строки** (по умолчанию они сняты), то в сохраненном документе разделение текста соответственно на страницы и строки будет таким же, как и в изображении-источнике. С помощью флажка **Сохранять цвет текста** вы можете при сохранении документа оставить цвет символов, который использовался в исходном изображении.

В поле **Качество картинок** можно указать, с каким качеством должны быть сохранены имеющиеся в документе иллюстрации. Вы можете вообще отказаться от них – тогда будет сохранен лишь текст; для этого выберите в раскрывающемся списке значение **Без картинок**. По умолчанию в данном поле предлагается значение **Среднее (для просмотра на экране)**.

При выборе значения **Пользовательское** на экран вызывается дополнительный диалог (рис. 8.7). В этом диалоге задаются пользовательские параметры сохранения картинок.

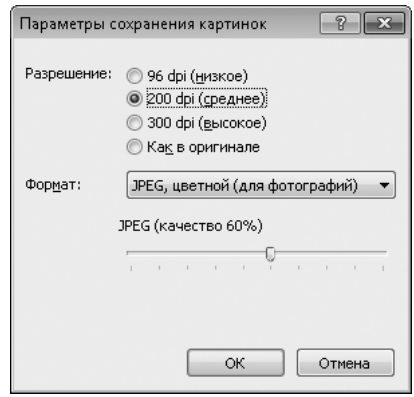


Рис. 8.7 ▾ Настройка дополнительных параметров сохранения

Чтобы неуверенно распознанные символы были в сохраненном документе выделены так же, как и в окне **Текст**, в данном окне нужно установить флажок **Выделять неуверенно распознанные символы**. Что касается второго параметра, то он позволяет получить документ, который можно будет открывать и редактировать в ранних версиях программы Microsoft Word и в других текстовых редакторах, поддерживающих формат RTF.

Настройка сохранения документа в формате Excel осуществляется на вкладке **XLS/XLSX** (рис. 8.8).

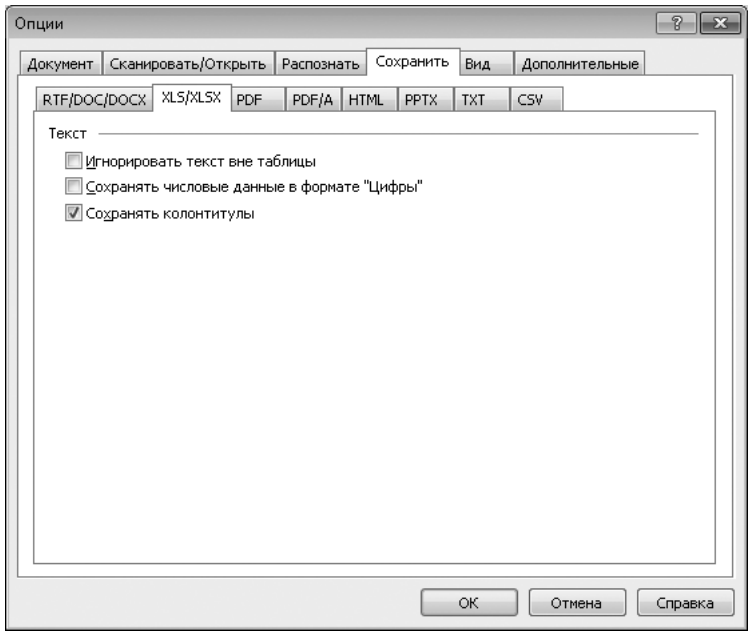


Рис. 8.8 ▾ Настройка сохранения документа в Excel

Данная вкладка содержит три параметра:

- **Игнорировать текст вне таблицы** – при установленном данном флажке в документ Excel будут включены только табличные данные. Вся остальная информация (например, заголовок таблицы и т. п.) будет проигнорирована;
- **Сохранять числовые данные в формате «Цифры»** – если включен этот параметр, то при сохранении документа в Excel все числовые значения будут сохранены в цифровом формате. Напомним, что Excel использует именно этот формат для ячеек с числовыми данными, участвующих в расчетах;
- **Сохранять колонтитулы** – при установленном данном флажке документ будет сохранен вместе с колонтитулами. В противном случае колонтитулы при сохранении будут проигнорированы.

Настройка сохранения документов в PDF-формате осуществляется на вкладке PDF (рис. 8.9).

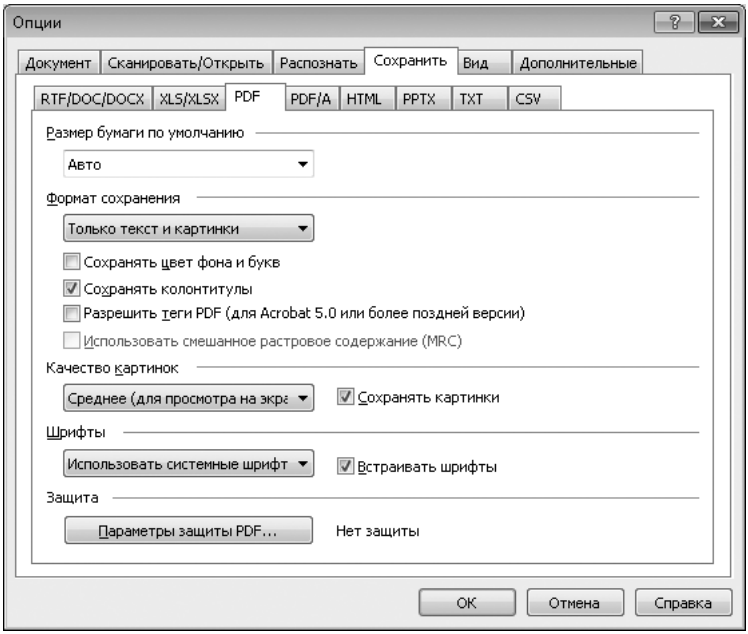


Рис. 8.9 ▾ Настройка сохранения документа в PDF

Как видно на рисунке, некоторые параметры данной вкладки нам уже знакомы. В частности, ими являются поля **Размер бумаги по умолчанию** и **Качество картинок**: их описание приведено выше (см. рис. 8.6), и здесь мы не будем на них останавливаться.

В поле **Формат сохранения** из раскрывающегося списка выбирается подходящий режим сохранения pdf-документа. Кратко охарактеризуем каждый из возможных вариантов.

- **Только текст и картинки** – при выборе этого режима сохранения будут сохранены и текст документа, и имеющиеся в нем иллюстрации. Получившийся в конечном итоге pdf-документ в некоторых случаях может иметь несущественные отличия от документа-источника. В нем можно осуществлять полнотекстовый поиск. Отметим, что данный режим предлагается использовать по умолчанию.
- **Текст поверх изображения страницы** – при использовании данного режима распознанный текст в процессе сохранения накладывается на фон страницы, а также на имеющиеся иллюстрации. Получившийся в конечном итоге pdf-документ в некоторых случаях может иметь несу-

щественные отличия от документа-источника. В нем можно осуществлять полнотекстовый поиск.

- **Текст под изображением страницы** – в данном случае текст документа помещается на внешне незаметном слое под изображением. Такой документ не имеет никаких отличий от документа-источника.
- **Только изображение** – при выборе данного режима будет сохранено изображение страницы. Получившийся в результате сохранения pdf-документ полностью соответствует оригиналу и не имеет никаких внешних отличий. Поскольку текстовый слой в таком документе отсутствует, поиск в нем не функционирует.

Далее на данной вкладке расположено несколько флажков, позволяющих выполнить тонкую настройку сохранения pdf-документа. Если необходимо, чтобы в pdf-документе цвет фона и символов совпадал с соответствующими цветами документа-источника, установите флажок **Сохранять цвет фона и букв**. При установленном флажке **Сохранять колонтитулы** документ будет сохранен вместе с колонтитулами (в противном случае при сохранении они будут проигнорированы).

Параметр **Разрешить теги PDF (для Acrobat 5.0 или более поздней версии)** требует более подробного пояснения. Дело в том, что pdf-файлы могут содержать не только иллюстрации и текст, но и сведения о структуре текущего документа: имеющихся в нем таблицах, логических блоках и т. п. Для хранения подобных сведений используются pdf-теги: они позволяют работать с документом на экранах разных размеров (в том числе и на экранах КПК). Так вот, если данный флажок установлен, то при сохранении pdf-документа будут сохранены и имеющиеся в нем pdf-теги, в противном же случае они будут проигнорированы. Отметим, что использование данного флажка имеет смысл только для версий программы Acrobat, начиная с 5.0.

Чтобы сохранить качество документа даже при сильном его сжатии, установите флажок **Использовать смешанное растровое содержание (MRC)**. При этом размер получившегося pdf-файла будет относительно небольшим.

В поле **Шрифты** указывается, какой тип шрифтов следует использовать при сохранении распознанных документов в pdf-формат. Если выбрать значение **Использовать стандартные шрифты**, то при сохранении будут использоваться стандартные шрифты Acrobat. При выборе значения **Использовать системные шрифты** при сохранении будут использоваться системные шрифты, установленные на данном компьютере.

С помощью кнопки **Параметры защиты PDF** осуществляется переход в режим настройки параметров защиты pdf-документа.

ПРИМЕЧАНИЕ

В данном случае под защитой подразумевается наложение ограничений на просмотр, редактирование и печать сохраняемого документа.

При нажатии данной кнопки на экране отображается окно, которое показано на рис. 8.10.

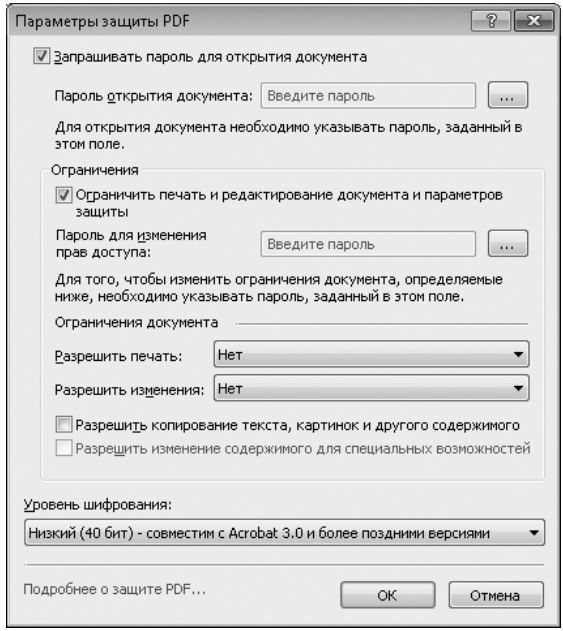


Рис. 8.10 ▾ Настройка параметров защиты pdf-документа

Вы можете сделать так, что открытие документа будет возможным только после ввода заданного вами пароля. Для этого нужно установить флажок **Пароль открытия документа**, нажать расположенную справа кнопку и в открывшемся окне (рис. 8.11) ввести требуемый пароль, после чего нажать кнопку **ОК**.

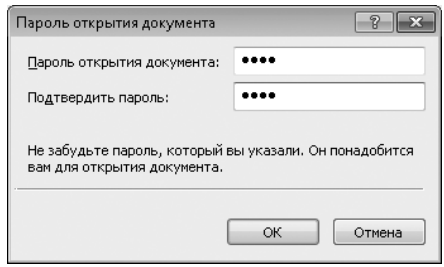


Рис. 8.11 ▾ Защита документа паролем

Пароль требуется ввести дважды – во избежание ошибок при вводе.

ВНИМАНИЕ

Обязательно запомните пароль, а еще лучше – сохраните его в надежном месте. Без ввода пароля последующее открытие документа будет невозможным.

Аналогичным образом можно ввести ограничения на внесение несанкционированных изменений в документ, а также на его печать. Для этого нужно установить флажок **Ограничить печать и редактирование документа и параметров защиты**, нажать расположенную справа кнопку, затем в открывшемся окне дважды ввести пароль и нажать кнопку **ОК**.

ПРИМЕЧАНИЕ

Если документ защищен паролями одновременно и на открытие, и на внесение изменений, то пароль на редактирование/печать будет иметь более высокий приоритет. Иначе говоря, ввод пароля на внесение изменений позволяет и открыть документ.

Отметим, что после ввода пароля на редактирование/печать документа вы можете отдельно указать, какие именно действия ограничиваются этим паролем. Например, если в поле **Разрешить печать** указано значение **Нет**, то после ввода пароля печать документа будет запрещена. Если же в этом поле указать значение **С высоким разрешением**, то даже после защиты документа паролем на редактирование/печать его можно будет распечатать с высоким разрешением.

То же самое касается и внесения изменений в защищенный паролем документ. Если в поле **Разрешить изменения** выбрано значение **Нет**, то любое его редактирование будет запрещено. Если же в этом поле выбрать значение **Изменение документа (за исключением извлечения страниц)**, то даже с защищенным паролем документом можно выполнять любые действия, кроме извлечения страниц.

С помощью флажка **Разрешить копирование текста, картинок и другого содержимого** вы можете разрешить другим пользователям копировать содержимое защищенного паролем документа. По умолчанию данный параметр отключен.

Выбор опции **Разрешить изменение содержимого для специальных возможностей** позволяет делать снимки открытого на экране pdf-документа.

В поле **Уровень шифрования** из раскрывающегося списка можно выбрать уровень шифрования защищенного паролем pdf-документа (при снятом флажке **Ограничить печать и редактирование документа и параметров защиты** данный параметр недоступен для редактирования). Учтите, что выбранный уровень шифрования влияет на возможность открытия документа разными версиями программы Acrobat.

На вкладке **Документ PDF/A** задаются опции сохранения в формат PDF/A. Данный формат есть смысл использовать в тех случаях, когда необходимо получить pdf-документ без визуальных потерь качества и поддерживающий возможность полнотекстового поиска, который в дальнейшем будет храниться в архиве. Параметры сохранения в формат PDF/A совпадают с настройками сохранения в формат PDF, за исключением выбора используемых шрифтов и защиты документа.

Настройка сохранения документа в HTML-формате осуществляется на вкладке **HTML** (рис. 8.12).

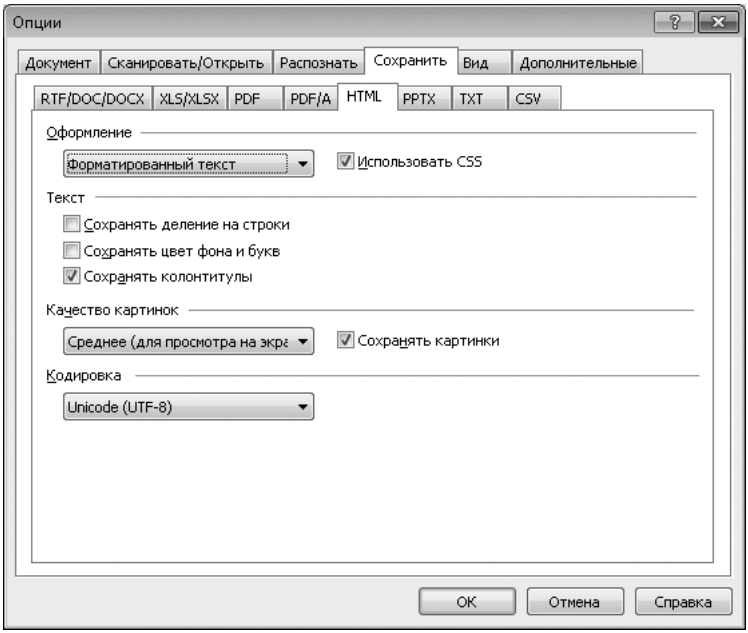


Рис. 8.12 ▾ Настройка сохранения документа в HTML

Некоторые параметры данного окна нам уже знакомы. В частности, таковыми являются флажки **Сохранять деление на строки**, **Сохранять цвет фона и букв** и **Сохранять колонтитулы**, а также поле **Качество картинок**. Их описание приведено выше (см. рис. 8.6), поэтому здесь на них останавливаться мы не будем.

В поле **Оформление** из раскрывающегося списка выбирается требуемый режим сохранения оформления документа. При выборе значений **Гибкая копия** или **Форматированный текст** документ будет сохранен в HTML-формат с сохранением шрифтового оформления. Форматирование сохраненного документа будет достаточно точно соответствовать документу-источнику. Если при этом флажок **Использовать CSS** снят, данные об оформлении сохраняются непосредственно в файле HTML в виде тегов. Если же он установлен, в файле HTML сохраняются только ссылки на таблицу стилей CSS, сохраняемую вместе с файлом.

Если же выбрать режим **Простой текст**, то при сохранении документа будет использовано стандартное оформление. При просмотре такого документа форматирование может заметно отличаться от форматирования документа-источника и полностью определяется настройками браузера, используемого для просмотра. Отметим, что формат документа при выборе режима сохране-

ния с таблицами стилей CSS может не поддерживаться некоторыми версиями веб-браузеров, текстовый же режим от подобных ограничений свободен.

В раскрывающемся списке **Кодировка символов** задается кодировка символов национальных алфавитов. Отметим, что в большинстве случаев оптимальными являются значения, предложенные по умолчанию.

На вкладке **PPTX** (рис. 8.13) выполняется настройка параметров сохранения документа в формат презентаций Microsoft Power Point 2007.

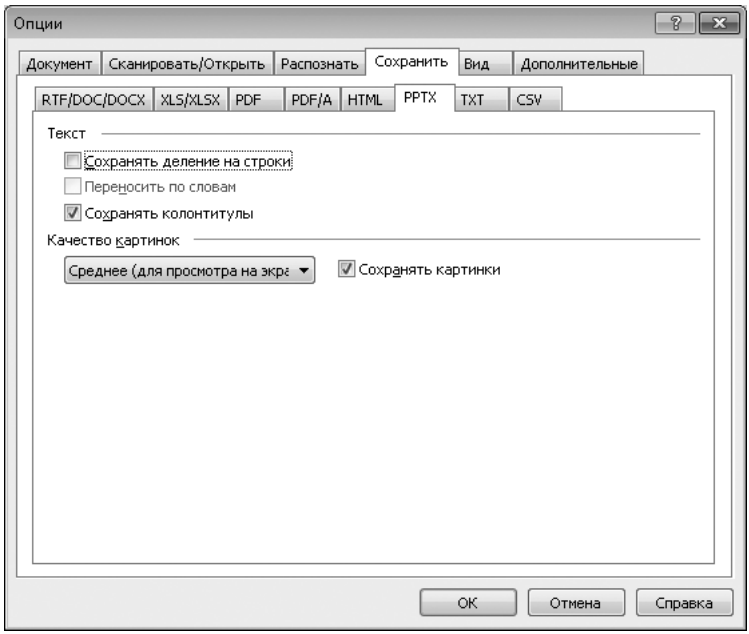


Рис. 8.13 ▾ Настройка сохранения документа в формат PPT/PPTX

На данной вкладке нам знакомы все параметры, кроме флажка **Переносить по словам**. Если он установлен, то распознанный и сохраненный текст будет умещен по ширине текстового блока слайда. Отметим, что данный флажок доступен только при установленном флажке **Сохранять деление на строки**.

На вкладке **TXT** выполняется настройка сохранения документа в текстовом формате (рис. 8.14).

Как видно на рисунке, некоторые параметры данной вкладки нам также уже знакомы, поэтому их описание мы здесь приводить не будем. Исключением являются флажки **Разделять страницы символом конца страницы (#12)** и **Разделять абзацы пустыми строками**.

С помощью параметра **Разделять страницы символом конца страницы (#12)** можно включить режим разделения текста в сохраненном txt-файле на страницы – так же, как он был разделен в окне **Страницы** программы Fine-

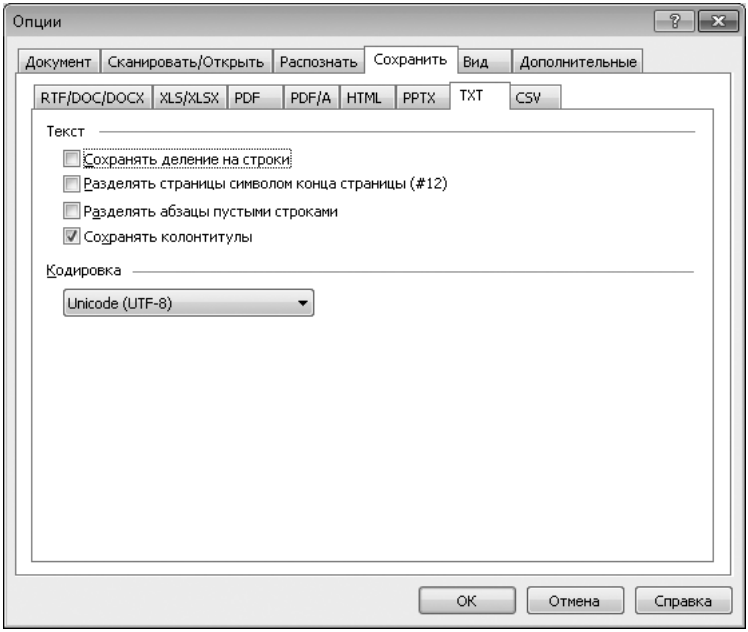


Рис. 8.14 ▾ Настройка сохранения документа в текстовом формате

Reader. При этом страницы будут отделены одна от другой специальным символом.

Если установить флажок **Разделять абзацы пустыми строками**, то в сохраненном в txt-формате тексте после каждого абзаца будет следовать пустая строка. Данная функциональность позволяет намного повысить наглядность и эргономичность текста (именно этих качеств не хватает обычно txt-документам).

Формат CSV часто используется для сохранения файлов с целью последующего импорта в базы данных или электронные таблицы (рис. 8.15).

На вкладке **CSV** имеется лишь один параметр, с которым мы ранее не встречались: это поле **Разделитель**. В нем из раскрывающегося списка выбирается символ, который будет использоваться в качестве разделителя между полями данных в csv-файле. По умолчанию в данном поле предлагается точка с запятой.

Все изменения, выполненные на вкладках, находящихся в подчинении вкладки **Сохранить**, вступают в силу только после нажатия в данном окне кнопки **ОК**. С помощью кнопки **Отмена** осуществляется выход из данного режима без сохранения выполненных изменений.

Отметим, что к диалогу **Опции** можно перейти из любого диалога сохранения или открытия файла. Для этого нужно нажать кнопку **Опции**, которая присутствует в нижней части диалога сохранения или открытия файла.

Итак, мы уже знаем, каким образом настроить требуемый режим сохранения распознанных документов. Далее мы на конкретных примерах рассмотрим порядок сохранения документов в каждом формате.

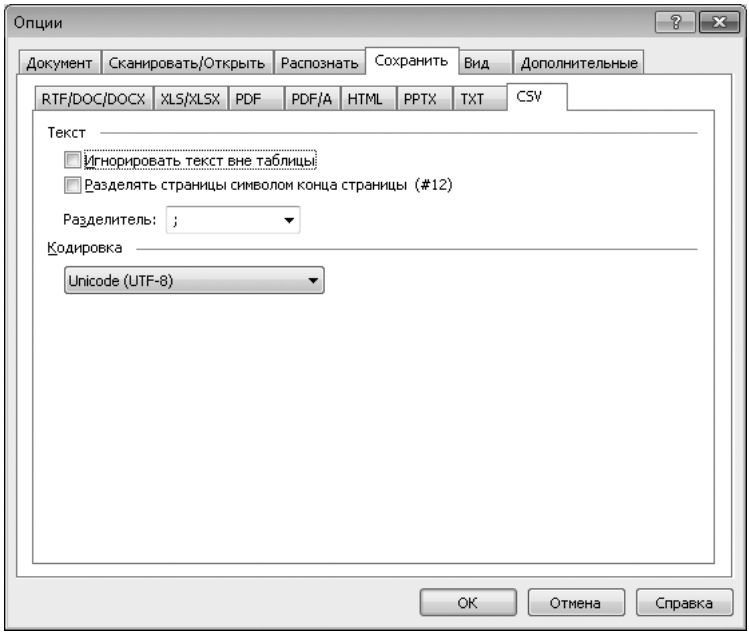


Рис. 8.15 ▾ Настройка сохранения документов в csv-формате

Примеры сохранения распознанного документа в формате Word

Предположим, что нам нужно сохранить в формат Word распознанный документ, который был показан на рис. 5.17. Это образец с достаточно сложной структурой: в нем присутствуют колонтитул, текст и таблица с заголовком.

Чтобы сохранить документ с помощью кнопки **Передать/Сохранить** на главной панели инструментов, нажмите треугольную стрелку справа от кнопки. В открывшемся меню выберите пункт **Сохранить как документ Microsoft Word 2007**. На кнопке отобразятся значок Word и слово **Сохранить**.

Затем в расположенном рядом раскрывающемся списке (рис. 8.2) выберем режим сохранения. Этот режим будет использоваться и при сохранении документа после нажатия кнопки **Передать/Сохранить**, и при вызове команды меню **Файл ➢ Сохранить как ➢ Документ Microsoft Word 2007**.

Чтобы сохранить документ в соответствии с установленными параметрами, нажмите кнопку **Сохранить**, или вызовите команду меню **Файл ➢ Сохранить как ➢ Документ Microsoft Word 2007.** В открывшемся окне укажите путь для сохранения и имя файла. В результате сохранения распознанного текста в режиме **Точная копия** будет получен документ, который будет оформлен в полном соответствии с документом-источником (рис. 8.16).

Обратите внимание: в данном случае отдельные фрагменты распознанного текста помещены в документ как объекты. Чтобы убедиться в этом, достаточно

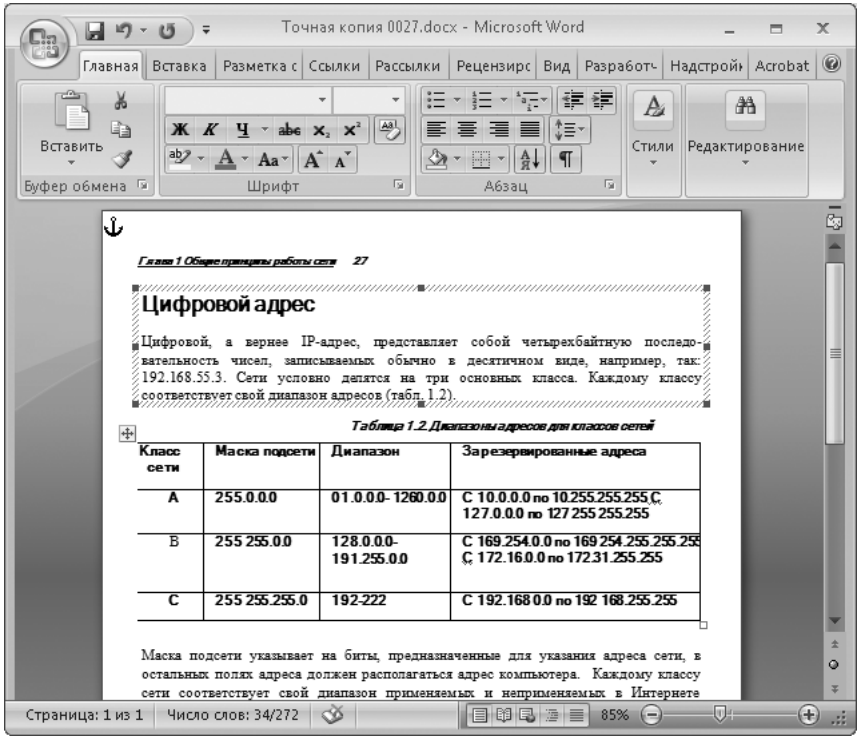


Рис. 8.16 ▾ Документ, сохраненный в режиме точной копии

щелкнуть кнопкой мыши на границе текста. Часть текста будет обведена рамкой с маркерами – это и есть границы объекта, а в углу страницы отобразится значок якоря – положение объекта зафиксировано относительно этого угла страницы. Использование объектов позволяет очень точно позиционировать текст, рисунки и таблицы на странице, но вместе с тем усложняет редактирование и форматирование содержимого документов.

Таким образом, сохранение или передача в режиме точной копии оправданы для документов со сложной структурой: обилием таблиц и рисунков с обтеканием текстом и т. п. Подробнее о том, что такое объекты и как с ними работать, вы можете узнать из справочной системы Microsoft Word.

Если сохранить распознанный текст в режиме **Редактируемая копия**, то внешний вид документа может незначительно отличаться от документа-источника (рис. 8.17). В большинстве случаев такие отличия практически незаметны. В нашем примере документ занял две страницы вместо одной в оригинале. Заметьте, что колонтитул появился и на второй странице, причем с соблюдением последовательной нумерации страниц. При сохранении в режиме **Редактируемая копия** текст помещается в документ Microsoft Word как обычные абзацы, а таблицы и рисунки также вставляются в виде абзацев.

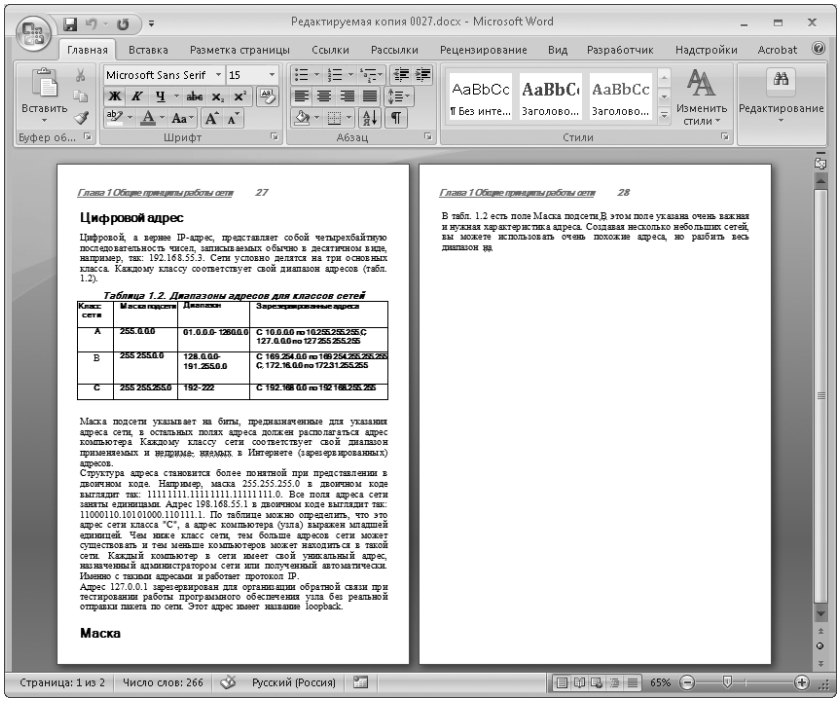


Рис. 8.17 ▾ Документ, сохраненный в режиме редактируемой копии

При сохранении документа в данном режиме его можно впоследствии редактировать безо всяких ограничений. Режим редактируемой копии наиболее удобен для последующей обработки документа в Microsoft Word и в большинстве случаев является оптимальным выбором.

Если сохранить распознанный документ в режиме **Форматированный текст**, то будет сохранено использованное в документе-источнике шрифтовое оформление. Однако при этом будут проигнорированы межстрочные интервалы, а также выравнивание абзаца. Иначе говоря, в результате будет получен отформатированный текст, который выровнен по левому краю страницы (рис. 8.18).

Что касается варианта сохранения **Простой текст**, то его название говорит само за себя: в данном случае будет получен простой текст, примерно как в документах формата txt. Как и при сохранении форматированных текстов, он будет выровнен по левому краю, но с использованием шрифта, принятого в Microsoft Word по умолчанию (рис. 8.19). Тем не менее таблицы в этом режиме сохраняются с соблюдением числа строк и столбцов.

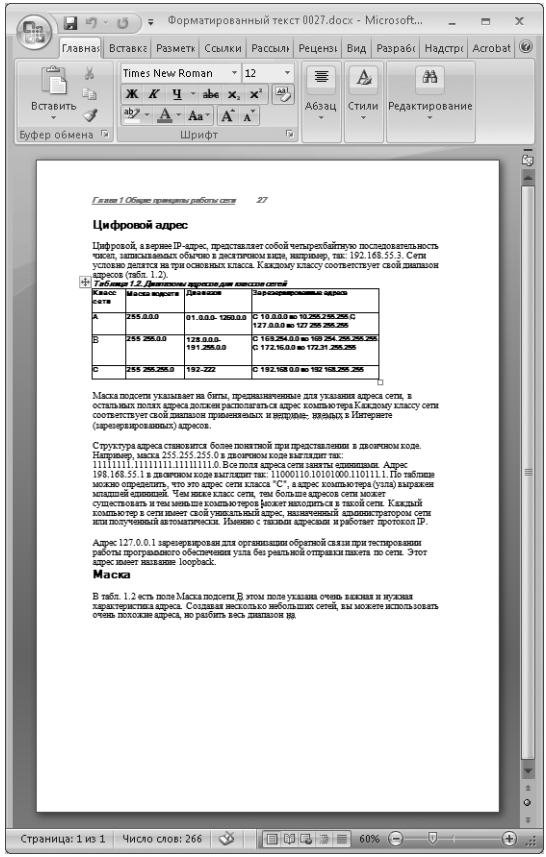


Рис. 8.18 ▾ Документ, сохраненный в режиме форматированного текста

Таким образом, один и тот же документ вы можете сохранить в формате Word в четырех разных режимах. Какому из этих режимов отдать предпочтение, зависит от того, как вы собираетесь работать с этим документом в дальнейшем. Например, если вы предполагаете использовать фрагменты сохраняемого документа для вставки в другие документы, удобным окажется режим **Простой текст**: в этом случае исключаются проблемы с появлением чужеродных стилей и упрощается форматирование итогового документа.

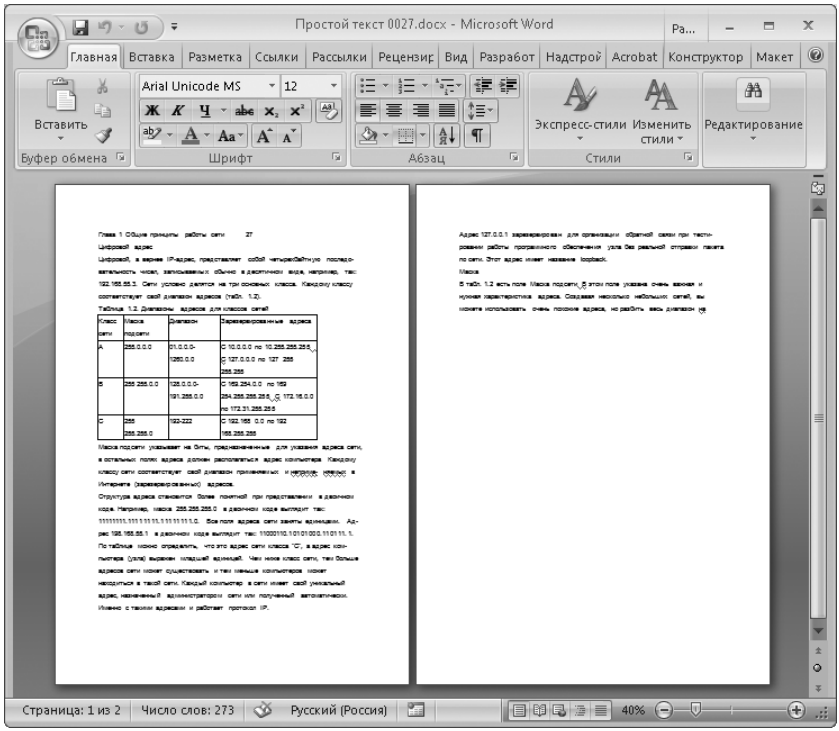


Рис. 8.19 ▾ Документ, сохраненный в режиме простого текста

Пример сохранения распознанного документа в формате PDF

Чтобы сохранить распознанный документ в pdf-формате, нужно в меню, открывающемся при нажатии на стрелку справа от кнопки **Передать/Сохранить**, выбрать значение **Сохранить как документ Adobe PDF** (рис. 8.21) или **Сохранить как PDF/A**, а затем нажать эту кнопку. При сохранении распознанного документа в pdf-формате в раскрывающемся списке доступен лишь один способ сохранения – **Точная копия**.

Если вы предпочитаете пользоваться не кнопками панелей инструментов, а меню, выберите в меню команду **Файл ➤ Сохранить документ как ➤ Документ PDF** (рис. 8.21). Чтобы сохранить только страницы документа, предварительно выбранные в окне **Страницы**, выберите в меню команду **Файл ➤ Сохранить выбранные страницы как ➤ Документ PDF**.

В каждом случае откроется диалоговое окно сохранения файла (рис. 8.22). Выберите папку и введите в поле **Имя файла:** имя, под которым будет сохранен файл. В диалоге **Сохранить как** вы можете еще раз уточнить параметры сохранения: в поле **Тип файла:** выбрать тип файла (например, PDF или PDF/A),

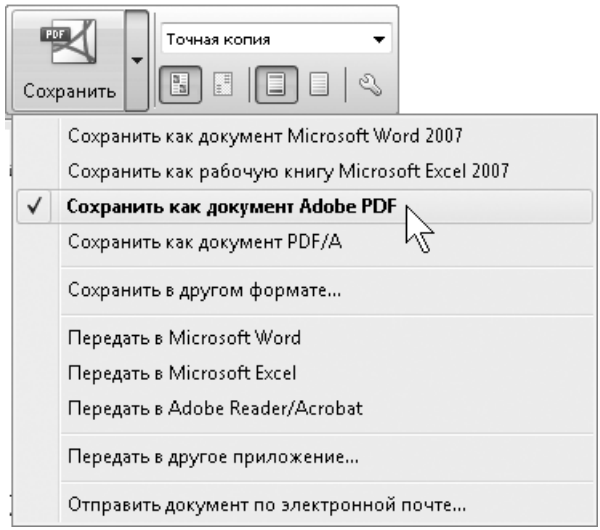


Рис. 8.20 ▾ Выбор функции кнопки **Передать/Сохранить**

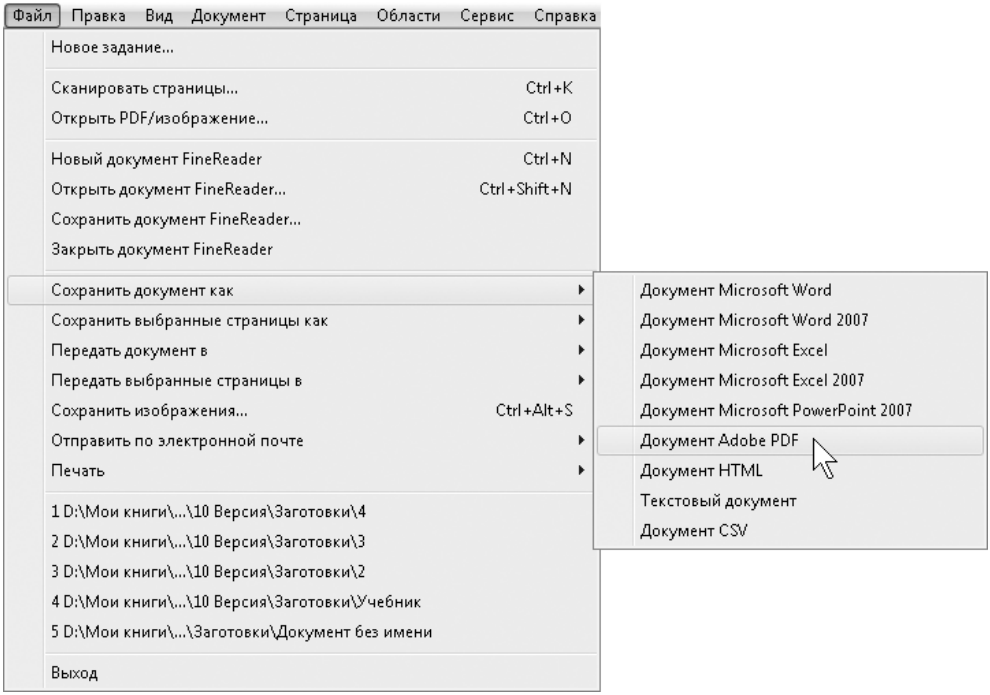


Рис. 8.21 ▾ Меню **Файл** ➢ **Сохранить документ как**

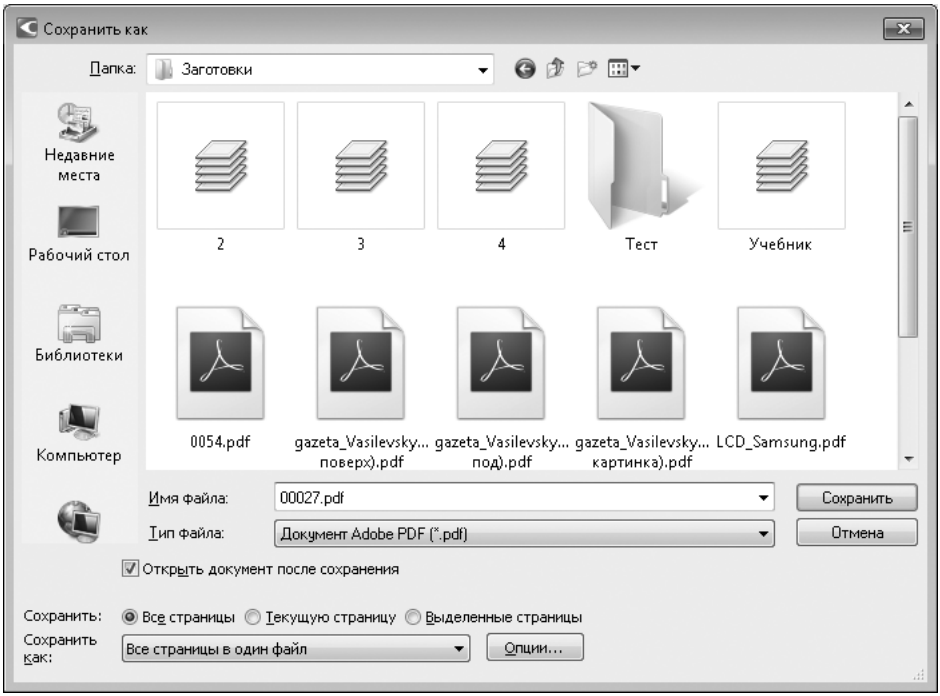


Рис. 8.22 ▾ Диалог **Сохранить как:**

с помощью переключателя **Сохранить:** указать, какие страницы документа следует сохранить, а в раскрывающемся списке **Сохранить как:** выбрать способ сохранения (все страницы в один файл либо каждую страницу в отдельный файл).

Кроме того, нажав кнопку **Опции**, вы можете вызвать из этого диалогового окна диалог **Опции**. Проверив параметры сохранения, нажмите кнопку **Сохранить**. Документ будет сохранен в формате PDF.

При сохранении используются настройки, предварительно заданные в диалоге **Опции**. Еще раз обратимся к такому свойству файла PDF, как наличие и взаимное расположение слоев. В качестве примера возьмем изображение газетной статьи (см. рис. 4.10). Выясним, как влияет на вид сохраненного документа то, какой из четырех вариантов был выбран в раскрывающемся списке **Формат сохранения** на вкладке **Сохранить ➢ PDF** диалога **Опции**.

На рис. 8.23–8.25 показаны PDF-документы, сохраненные в различных режимах. Режим **Только текст и картинки** предлагается в FineReader 10 в качестве режима по умолчанию. Документы, сохраненные в режимах **Текст под изображением страницы** и **Только изображение**, выглядят идентично, поэтому последний пример относится к обоим режимам.

Рис. 8.23 ▾ Документ PDF, сохраненный в режиме **Только текст и картинки**

Рис. 8.24 ▾ Документ PDF, сохраненный в режиме **Текст поверх изображения страницы**

Таким образом, при сохранении распознанного документа в формат PDF перед вами открывается широкий выбор возможностей. В большинстве случаев опции сохранения, предлагаемые FineReader 10 по умолчанию, являются оптимальными.

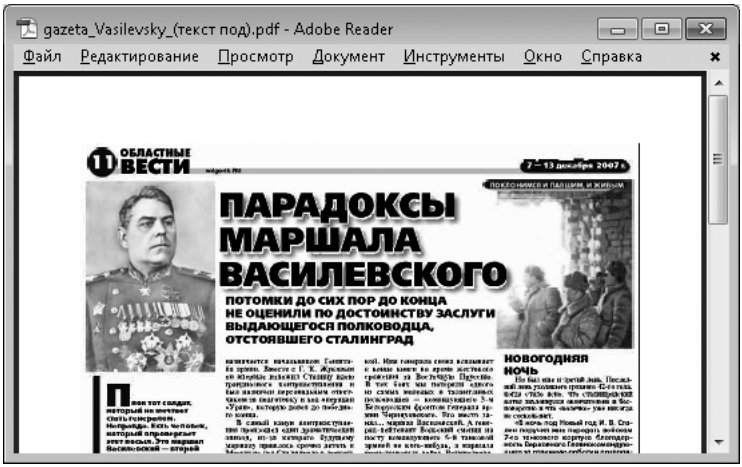


Рис. 8.25 ▾ Документ PDF, сохраненный в режиме **Текст под изображением страницы**

Пример сохранения распознанного документа в формате HTML

Чтобы сохранить распознанный документ в формате HTML, выберите в меню команду **Файл ➢ Сохранить документ как ➢ Документ HTML**. Чтобы сохранить только страницы документа, предварительно выбранные в окне **Страницы**, выберите в меню команду **Файл ➢ Сохранить выбранные страницы как ➢ Документ HTML**.

Откроется диалоговое окно сохранения файла (см. рис. 8.22). Выберите папку и введите в поле **Имя файла:** имя, под которым будет сохранен файл. Как и при сохранении в формат PDF, вы можете указать в этом диалоге параметры сохранения страниц документа.

Режим сохранения оформления (**Гибкая копия, Форматированный текст** или **Простой текст**) следует предварительно задать в диалоге **Опции** на вкладке **Сохранить ➢ HTML**. Указанный режим программа FineReader будет использовать при сохранении каждого документа HTML до тех пор, пока вы вновь не измените режим на вкладке **Сохранить ➢ HTML** диалога **Опции**.

Также вы можете в меню, открывающемся при нажатии на стрелку справа от кнопки **Передать/Сохранить**, выбрать значение **Сохранить в другом формате** и в открывшемся диалоге сохранения выбрать нужный формат. При сохранении распознанного документа в формате html в раскрывающемся списке, находящемся правее кнопки **Передать/Сохранить**, доступны три режима сохранения: **Гибкая копия**, **Форматированный текст** и **Простой текст**.

В качестве примера взята страница из учебника английского языка. Если мы сохраним распознанный документ в HTML-формат, когда в диалоге **Опции** был выбран вариант **Гибкая копия**, то документ HTML (веб-страница) будет выглядеть так, как показано на рис. 8.26.

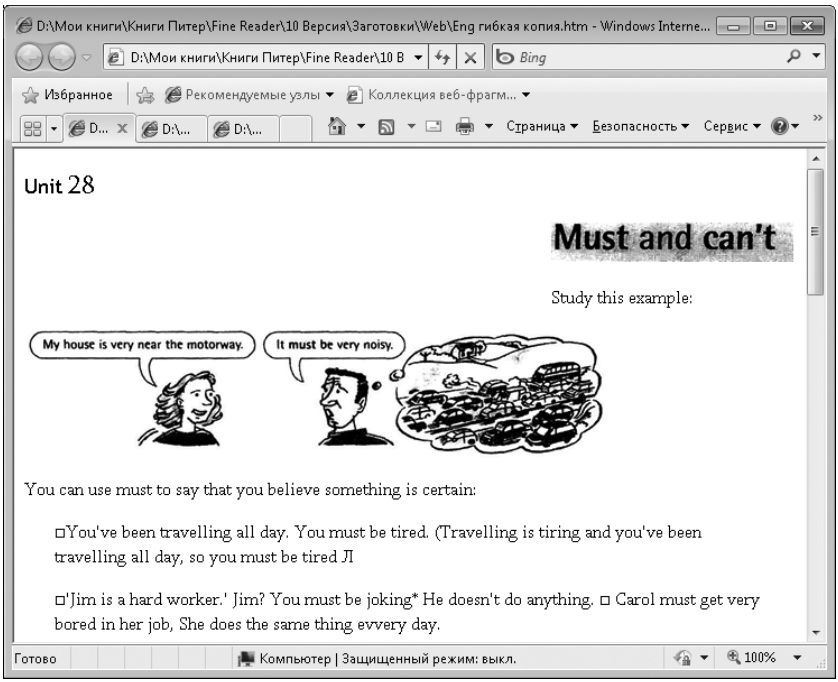


Рис. 8.26 ▾ Сохранение точной копии документа в HTML-формате

Сохранение распознанного документа в режиме **Форматированный текст** даст следующий результат (рис. 8.27).

Что касается варианта **Простой текст**, то в данном случае сохраненный HTML-документ будет выглядеть следующим образом (рис. 8.28).

Таким образом, возможности FineReader позволяют сохранить один и тот же распознанный документ в HTML-формат в трех разных вариантах.

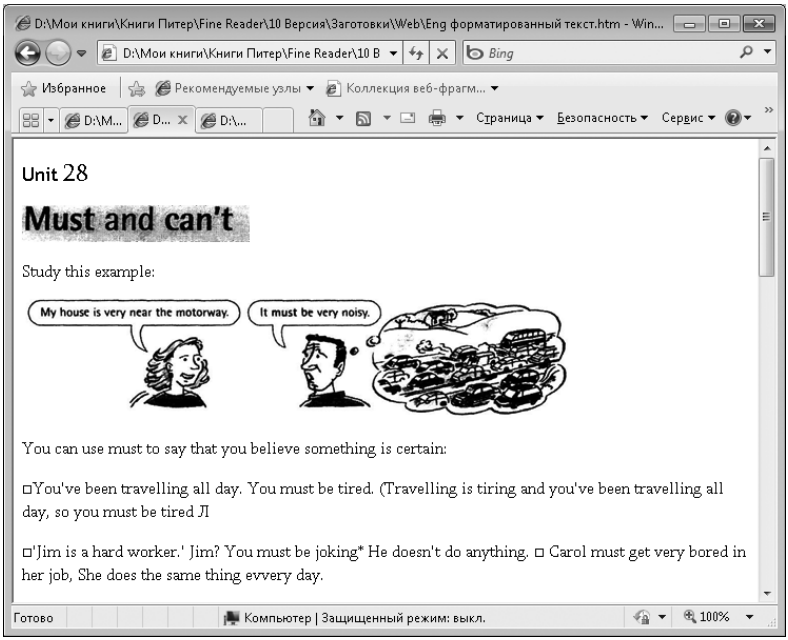


Рис. 8.27 ▾ Сохранение форматированного текста в HTML-формате

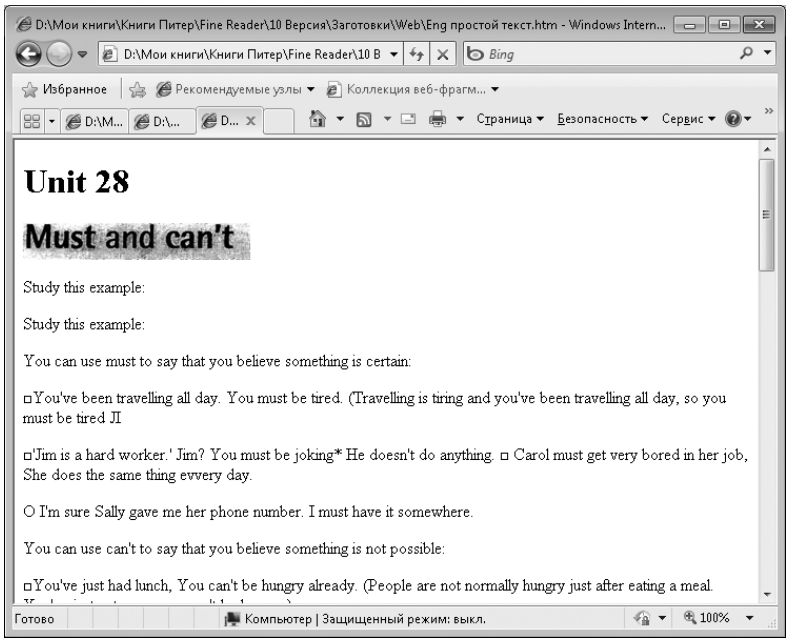


Рис. 8.28 ▾ Сохранение простого текста в HTML-формате

Пример сохранения распознанного документа в формате TXT

При сохранении распознанных документов в текстовый формат надо учитывать, что таким образом сохраняется только текст: все элементы оформления и рисунки будут утеряны. Если в распознанном документе содержались таблицы, то при сохранении в текстовый файл они преобразуются в «чистый» текст: содержимое ячеек разделяется символами табуляции, а каждая строка таблицы становится отдельным абзацем текста.

Чтобы сохранить распознанный документ в формате TXT, выберите в меню команду **Файл ➢ Сохранить документ как ➢ Текстовый документ**. Чтобы сохранить только страницы документа, предварительно выбранные в окне **Страницы**, выберите в меню команду **Файл ➢ Сохранить выбранные страницы как ➢ Текстовый документ**.

Откроется диалоговое окно сохранения файла (см. рис. 8.22). Выберите папку и введите в поле **Имя файла:** имя, под которым будет сохранен файл. В диалоге вы можете указать параметры сохранения страниц документа.

Нажмите кнопку **Сохранить**. Документ будет сохранен в текстовый файл. При этом используются настройки, которые были заданы в диалоге **Опции** на вкладке **Сохранить ➢ TXT**.

Таким же образом осуществляется сохранение распознанного документа в csv-формат. Основное отличие файла CSV от текстового файла в том, что значения, содержащиеся в ячейках таблиц, отделяются друг от друга символом-разделителем, а в конце каждой строки вставляется символ конца строки.

Пример сохранения распознанного документа в формате Excel

Теперь рассмотрим пример сохранения распознанного документа в Excel-файл. Чтобы сделать пример более наглядным, будем работать с распознанным документом, представляющим собой таблицу.

Для сохранения документа в формат XLSX вы можете воспользоваться командами меню **Файл ➢ Сохранить документ как ➢ Документ Microsoft Excel 2007** или **Файл ➢ Сохранить выбранные страницы как ➢ Документ Microsoft Excel 2007** либо кнопкой **Передать/Сохранить**, выбрав для нее функцию **Сохранить как рабочую книгу Microsoft Excel 2007** (рис. 8.29).

После нажатия кнопки **Сохранить** в открывшемся диалоговом окне укажем путь для сохранения и имя файла. Отметим, что при сохранении распознанных документов в Excel-формат возможно использование только одного варианта сохранения – **Форматированный текст**.

Как уже сказано, в диалоге **Опции** на вкладке **XLSX** задаются дополнительные параметры сохранения распознанного документа в формат XLSX. По умолчанию флажок **Игнорировать текст вне таблицы** снят, и в файл формата Microsoft Excel сохраняется все содержимое документа (рис. 8.30).

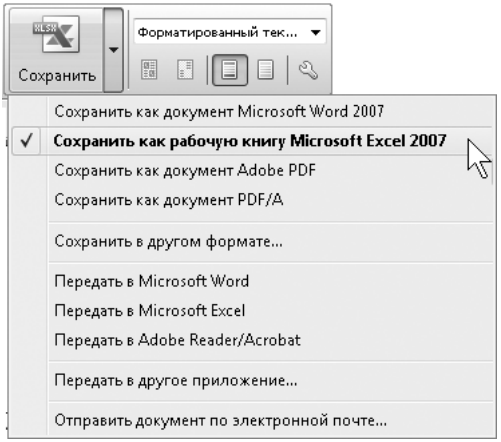


Рис. 8.29 ▾ Выбор формата для сохранения Excel-документа

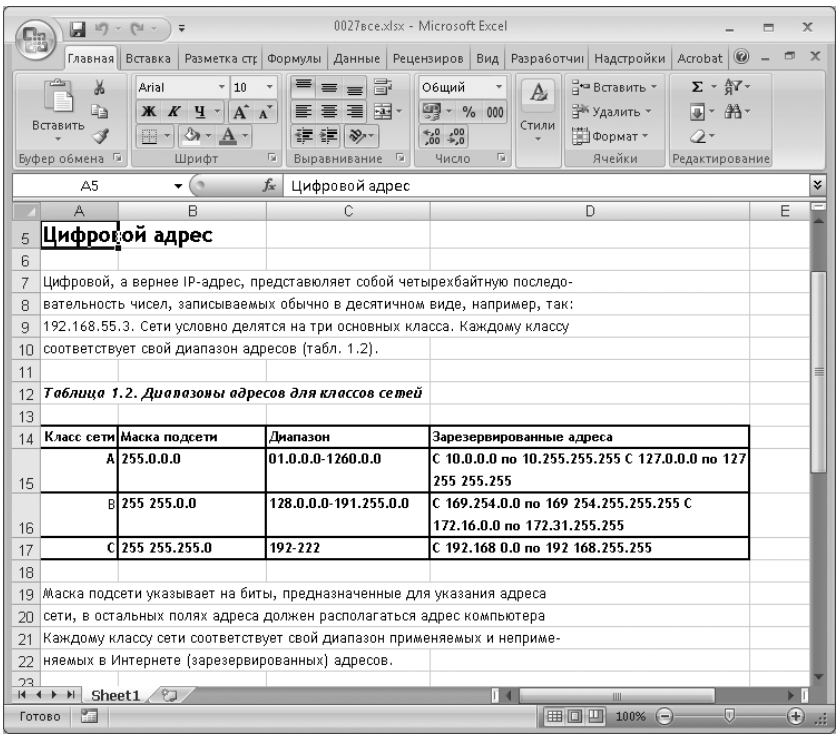
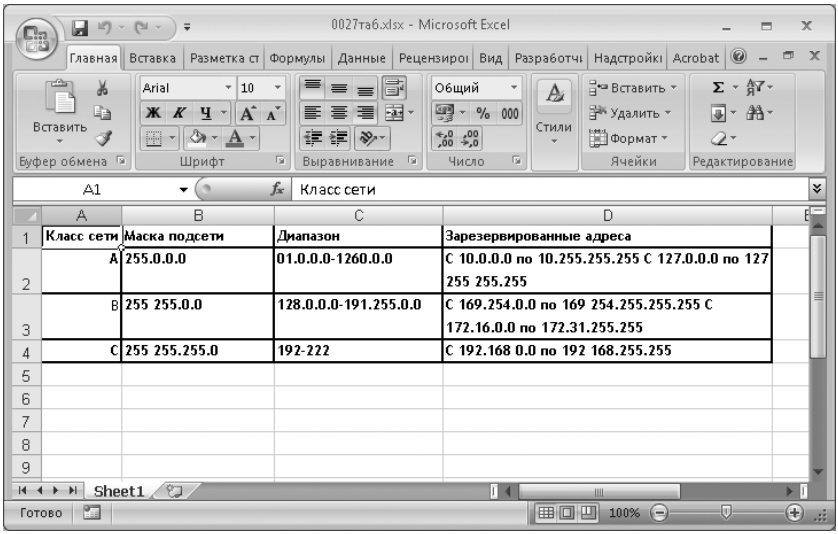


Рис. 8.30 ▾ Документ, сохраненный в Excel-формате

Чтобы в таблицу Microsoft Excel сохранялись только таблицы, находящиеся в распознанном документе, перед сохранением файла вызовите диалог **Опции** и на вкладке **XLSX** установите флажок **Игнорировать текст вне таблицы**. После этого сохраните документ в формате XLSX. В результате в документе Microsoft Excel будет сохранена лишь табличная часть распознанного документа (рис. 8.31).



	A	B	C	D
1	Класс сети	Маска подсети	Диапазон	Зарезервированные адреса
2	A	255.0.0.0	01.0.0.0-1260.0.0	С 10.0.0.0 по 10.255.255.255 С 127.0.0.0 по 127 255 255.255
3	B	255 255.0.0	128.0.0.0-191.255.0.0	С 169.254.0.0 по 169 254.255.255.255 С 172.16.0.0 по 172.31.255.255
4	C	255 255.255.0	192-222	С 192.168 0.0 по 192 168.255.255
5				
6				
7				
8				
9				

Рис. 8.31 ▾ В документ XLSX сохранена только табличная часть распознанного документа

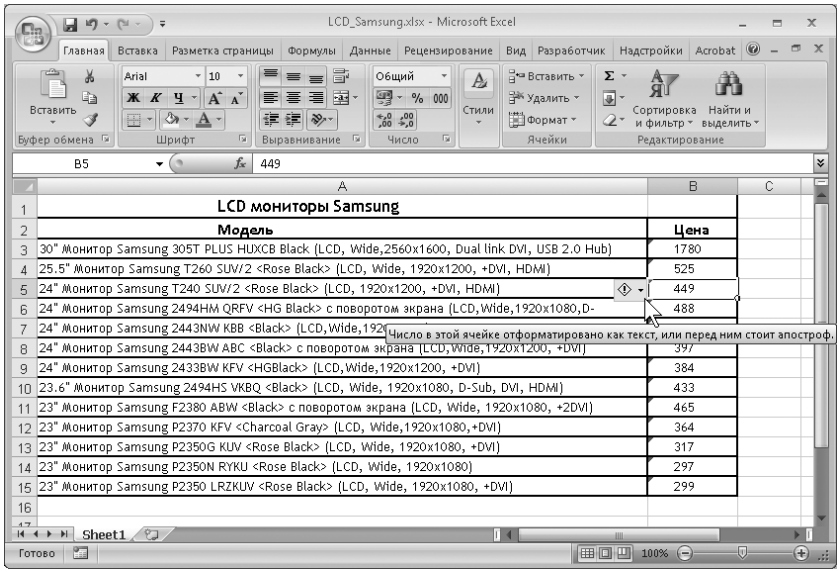
При сохранении документов, в которых после открытия их в Microsoft Excel вы предполагаете вставлять формулы, возможны проблемы с вычислением результатов математических операций с содержимым ячеек. Это связано с тем, что некоторым ячейкам, содержащим числа, при сохранении документа может быть присвоен формат *текста*.

Выполнение математических операций с ячейками, которым присвоен текстовый формат, ведет к ошибке. Ячейки, которые содержат числа, но отформатированы как текст, в окне Microsoft Excel помечаются треугольными маркерами. При наведении указателя мыши на такие ячейки программа Microsoft Excel выдает предупреждения о несоответствии формата ячеек (рис. 8.32).

Чтобы устранить эту проблему, сохраним распознанный документ заново, предварительно изменив соответствующую настройку FineReader.

В диалоге **Опции** на вкладке **XLS/XLSX** (рис. 8.8) установите флажок **Сохранять числовые данные в формате «Цифры»**. Закройте диалог **Опции**, нажав кнопку **OK**.

Вновь сохраните документ в формате XLSX. При открытии этого файла в программе Microsoft Excel видно, что пометки, свидетельствующие об ошиб-



	A	B	C
1	LCD мониторы Samsung		
2	Модель	Цена	
3	30" Монитор Samsung 305T PLUS HUXCB Black (LCD, Wide,2560x1600, Dual link DVI, USB 2.0 Hub)	1780	
4	25.5" Монитор Samsung T260 SUV/2 <Rose Black> (LCD, Wide, 1920x1200, +DVI, HDMI)	525	
5	24" Монитор Samsung T240 SUV/2 <Rose Black> (LCD, 1920x1200, +DVI, HDMI)	449	
6	24" Монитор Samsung 2494HM QRFV <HG Black> с поворотом экрана (LCD,Wide,1920x1080,D-	488	
7	24" Монитор Samsung 2443NW KBB <Black> (LCD,Wide,192		
8	24" Монитор Samsung 2443BW ABC <Black> с поворотом экрана (LCD,Wide,1920x1200, +DVI)	397	
9	24" Монитор Samsung 2433BW KFV <HGBlack> (LCD,Wide,1920x1200, +DVI)	384	
10	23.6" Монитор Samsung 2494HS VKBQ <Black> (LCD, Wide, 1920x1080, D-Sub, DVI, HDMI)	433	
11	23" Монитор Samsung F2380 ABW <Black> с поворотом экрана (LCD, Wide, 1920x1080, +2DVI)	465	
12	23" Монитор Samsung P2370 KFV <Charcoal Gray> (LCD, Wide,1920x1080,+DVI)	364	
13	23" Монитор Samsung P2350G KUV <Rose Black> (LCD, Wide, 1920x1080, +DVI)	317	
14	23" Монитор Samsung P2350N RYKU <Rose Black> (LCD, Wide, 1920x1080)	297	
15	23" Монитор Samsung P2350 LRZKUV <Rose Black> (LCD, Wide, 1920x1080, +DVI)	299	
16			

Рис. 8.32 ▾ Предупреждение о несоответствии формата ячеек

ках, исчезли. Теперь ячейки с числовыми данными имеют цифровой формат и, следовательно, с их содержимым могут выполняться математические действия.

Как сохранить документ FineReader

Вы можете сохранить текущий документ FineReader. Это бывает целесообразно, например, когда вы не успеваете завершить работу с ним сразу после распознавания и потому планируете вернуться к нему позже, а также в иных случаях.

При сохранении документа FineReader сохраняется как исходный, так и распознанный вариант, а также различного рода вспомогательные файлы. Все эти объекты хранятся в папке, имя которой указывает пользователь в процессе сохранения.

Предположим, что нам нужно сохранить текущий документ FineReader, чтобы впоследствии вернуться к работе с ним. Для этого выполним команду главного меню **Файл ➢ Сохранить документ FineReader** – в результате на экране откроется диалоговое окно, изображенное на рис. 8.33.

В данном окне в поле **Имя документа** нужно ввести имя документа FineReader (как мы уже отмечали, фактически этот документ представляет собой папку с файлами документа) и нажать кнопку **Сохранить**.

Чтобы впоследствии вернуться к работе с сохраненным ранее документом FineReader, выберите в главном меню программы команду **Файл ➢ Открыть документ FineReader** или нажмите комбинацию клавиш **Ctrl**+**Shift**+**N**. В результате на экране откроется окно **Открыть документ FineReader** (рис. 8.34).

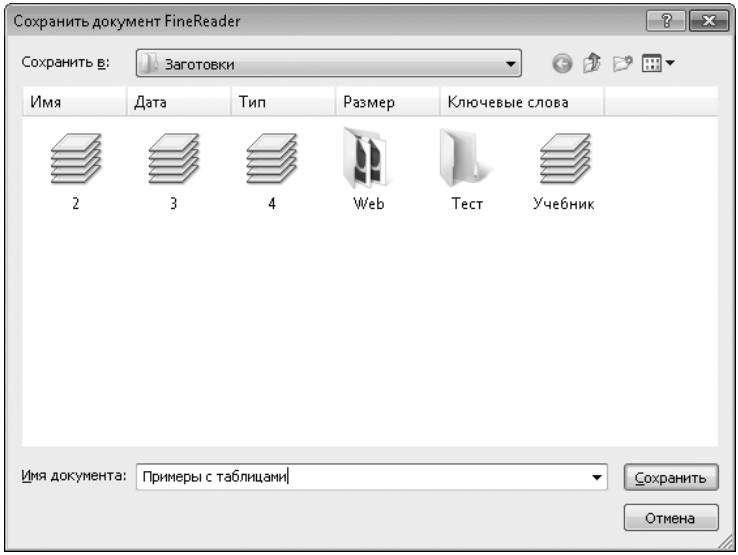


Рис. 8.33 ▾ Сохранение документа FineReader

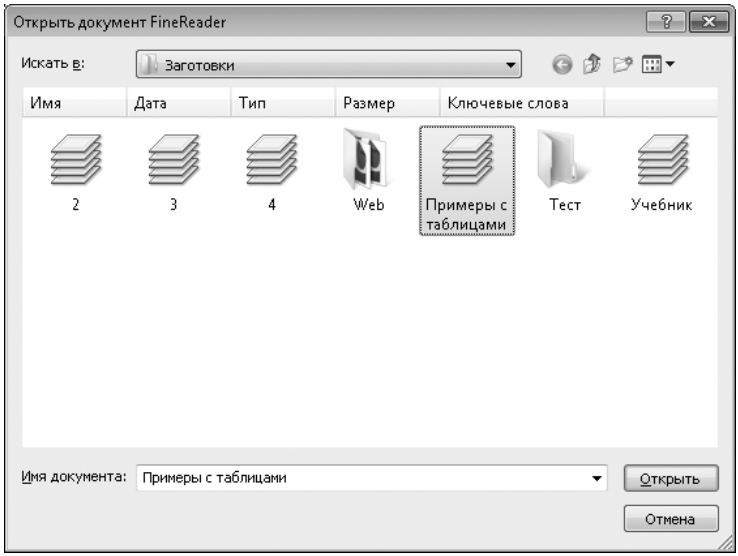


Рис. 8.34 ▾ Открытие документа FineReader

В данном окне в поле **Искать в** нужно указать каталог, в котором находится требуемый документ, затем выделить его щелчком мыши и нажать кнопку **Открыть**. В данном примере наш документ называется **Test**. Обратите внимание: папка документа FineReader отличается характерным значком (см. рис. 8.34).

Резюме

Прочитав эту главу, вы убедились в том, что важно уметь не только распознать и, в случае надобности, обработать документ, но и сохранить его в требуемом формате. Причем сохранить корректно, при необходимости выполнив соответствующие настройки: ведь нередко бывает так, что установка или снятие какого-нибудь одного флажка кардинальным образом влияет на результат.

Теперь вы знаете, что если работа с документом требует большого количества времени, вы можете отложить ее «на потом», сохранив документ FineReader и вернувшись к нему тогда, когда появится такая возможность. Это очень полезная функциональность, которую особенно удобно применять при работе с объемными документами.

Стоит отметить, что работа с документами в программе FineReader может быть довольно однообразной, складываясь из одних и тех же этапов (например, сканирование, распознавание, сохранение в файл). Чтобы не дублировать одни и те же процессы при работе с разными документами, вы можете их автоматизировать, объединив в пользовательский сценарий. О том, как это делается, мы расскажем в следующей главе нашей книги.

Глава 9 Сценарии

Со встроенными сценариями программы FineReader мы уже познакомились в главе «Быстрый старт». Они позволяют выполнить типовые задачи «одним нажатием кнопки».

Встроенные сценарии выполняются с текущими настройками программы. Если в процессе работы вы изменили какие-то настройки в диалоговом окне **Опции**, а потом закрыли программу, при следующем запуске она будет использовать эти измененные настройки. Поэтому при выполнении сценария будут использованы те настройки обработки изображения, его анализа, распознавания и передачи/сохранения файла, которые были заданы в диалоге **Опции** в последний раз. При распознавании используются языки, которые были выбраны в раскрывающемся списке окна **Новое задание** перед запуском сценария.

Пользовательские сценарии вы создаете самостоятельно. Они нужны для выполнения повторяющихся задач, которые не вполне подходят под один из готовых сценариев программы.

Создание пользовательского сценария

Перед тем как создавать сценарий, нужно четко обрисовать задачу, представить себе все этапы ее решения. Возможно, она вполне подойдет и под один из существующих встроенных сценариев!

Основная сфера применения пользовательских сценариев – распознавание с использованием шаблонов областей или с сохранением результатов в файлы не самых распространенных форматов. Кроме того, сценарий стоит создавать, если одну и ту же задачу предстоит выполнять достаточно регулярно.

Сначала подробно разберем один пример. В нем задействованы многие из функций и настроек, с которыми вы встретитесь при создании любого сценария. Затем отдельно рассмотрим те моменты, которые в этот пример не вошли.

Пример: распознавание платежного поручения

Предположим, вы систематически сканируете и распознаете в программе FineReader платежные поручения для внесения в базу данных. Чтобы выделить из всего документа и распознать только текст, расположенный в определенных местах, применяются шаблоны областей. Создание и сохранение подобного шаблона мы рассмотрели при обсуждении анализа изображений.

Передавать в базу данных наборы значений, каждое из которых должно попасть в определенное поле базы данных, удобно через файлы формата CSV. Саму процедуру импорта данных мы рассматривать не будем – она зависит от той программы, в которую эти данные должны попасть. Скажем лишь, что возможность импорта записей из файлов CSV предусмотрена практически в любой программе для ведения учета и бухгалтерии.

Таким образом, последовательность действий по работе с платежным поручением состоит из четырех основных шагов:

1) отсканировать оригинал или несколько оригиналов подряд;
2) загрузить шаблон областей и применить его ко всем изображениям;
3) распознать;
4) сохранить результат распознавания в файл CSV – каждую «платежку» в отдельный файл.

Для автоматизации такой задачи целесообразно создать особый сценарий. В его основу ляжет названная последовательность шагов.

Для управления сценариями откройте окно **Менеджера сценариев**: выберите команду меню **Сервис ➢ Менеджер сценариев** или нажмите сочетание клавиш **Ctrl+T**. Откроется окно **Менеджера сценариев** (рис. 9.1).

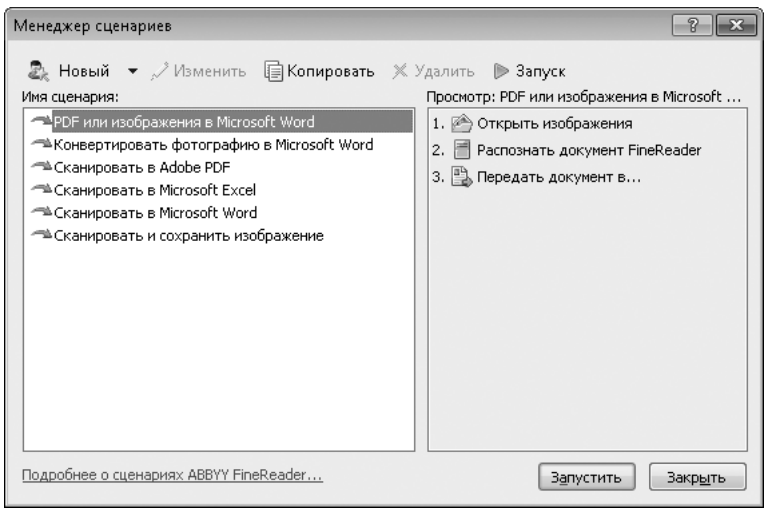


Рис. 9.1 ▾ Менеджер сценариев

1. Чтобы создать пользовательский сценарий, нажмите на панели инструментов в верхней части окна кнопку **Новый**. Откроется диалог **Новый сценарий** (рис. 9.2).

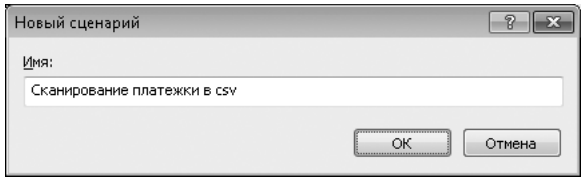


Рис. 9.2 ▾ Ввод имени нового сценария

2. Введите в текстовое поле имя создаваемого сценария, например `Сканирование платежки в csv`. Нажмите кнопку **OK**. Диалог закроется, и появится окно **Мастера сценариев** (рис. 9.3).

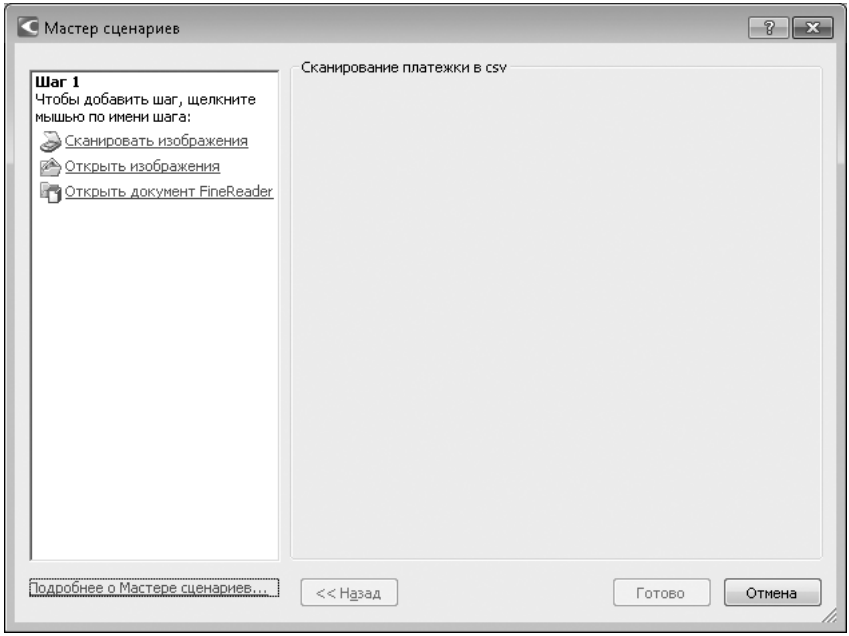


Рис. 9.3 ▾ Мастер сценариев

В левой части окна указывается номер шага и перечисляются доступные на данном этапе действия. Чтобы внести действие в сценарий, щелкните кнопкой мыши на соответствующей ему ссылке. В правой части показываются действия, уже внесенные в настоящий сценарий. Первоначально этот список пуст.

На первом шаге Мастер всегда предлагает выбрать одно из трех действий:
- сканировать изображения;
- открыть изображения;
- открыть документ FineReader.

3. Щелкните кнопкой мыши на ссылке **Сканировать изображения**. Это действие будет внесено в сценарий. Мастер предложит выбрать действие для второго шага сценария (рис. 9.4).

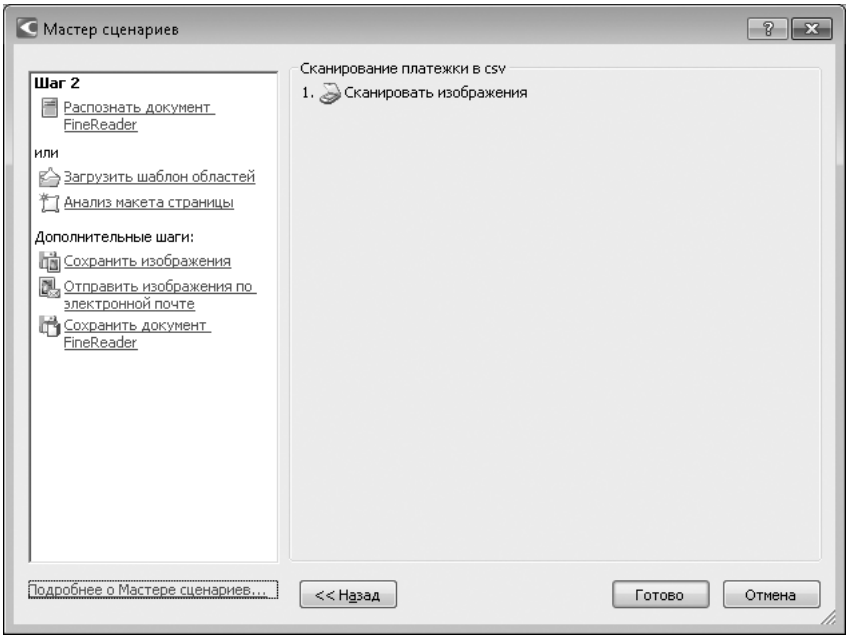


Рис. 9.4 ▾ Выбор второго шага сценария

На втором шаге список доступных действий тоже всегда одинаков и не зависит от того, какое действие было выбрано в качестве первого шага. В качестве основных вариантов предлагаются три действия:
- распознать документ FineReader. При выборе этого шага по сценарию будут запущены автоматический анализ и распознавание полученных изображений;
- загрузить шаблон областей. Этот тот вариант, который нам и нужен по замыслу;
- анализ макета страницы. Когда в сценарии задан этот вариант действий, программа только проведет анализ изображений. Как правило, такой вариант выбирают, чтобы при выполнении сценария после автоматического анализа получить возможность оценить его результаты и при необходимости внести коррективы вручную.

Кроме того, Мастер предлагает еще три варианта дополнительных действий:

- сохранить изображения;
- отправить изображения по электронной почте;
- сохранить документ FineReader.

Те же самые действия Мастер будет предлагать и на каждом последующем шаге. Иначе говоря, сохранение изображений, всего документа FineReader или отправку изображений по электронной почте можно задать в любом месте сценария.

Остальные действия, предлагаемые Мастером на следующих шагах, зависят от того, каковы предшествующие шаги сценария. Поэтому доведем до конца создание настоящего сценария в соответствии с планом нашего примера, а затем рассмотрим некоторые действия из тех, которые в этот «учебный» сценарий не вошли.

4. Выберите вариант **Загрузить шаблон областей**. В сценарий добавлен второй шаг. Запись об этом шаге появится в правой части окна **Мастера сценариев** (рис. 9.5). В левой же части окна Мастер предлагает задать третий шаг сценария.

   Для действия **Загрузить шаблон областей**, как и для многих других, доступны дополнительные настройки. Поэтому рядом с записью о втором шаге показывается ссылка **Изменить...**. Прежде чем задать следующий шаг сценария, уточним, как именно будет загружаться шаблон областей.

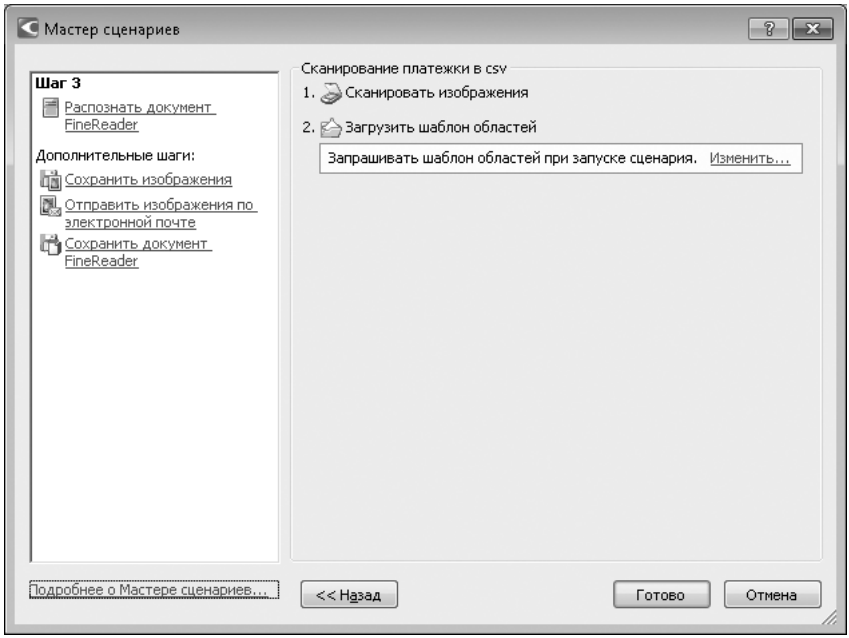


Рис. 9.5 ▾ Второй шаг добавлен в сценарий

5. Щелкните кнопкой мыши на ссылке **Изменить...**. Откроется диалог настройки указанного действия (рис. 9.6).

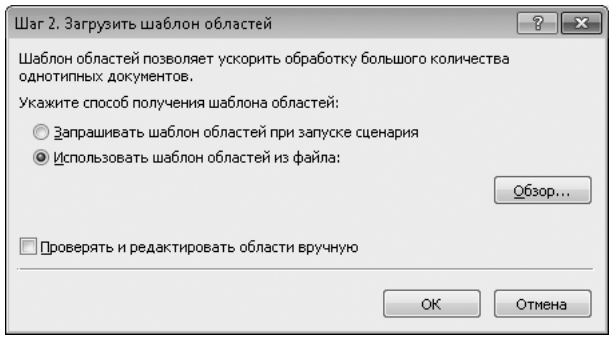


Рис. 9.6 ▾ Настройка загрузки шаблона областей

Укажите способ получения шаблона областей, установив переключатель в одно из положений:

- **Запрашивать шаблон областей при запуске сценария**. В этом случае в ходе выполнения сценария на экране появится диалог открытия файла шаблонов (рис. 5.15). После того как вы выберете и откроете файл с шаблоном областей, сценарий продолжит работу дальше;
- **Использовать шаблон областей из файла:**. В этом случае указанный файл шаблона загружается автоматически без каких-либо запросов.

Чтобы указать файл, нажмите кнопку **Обзор** и в открывшемся диалоге выберите файл шаблона областей.

Флажок **Проверять и редактировать области вручную** определяет поведение программы после применения шаблона к изображениям.

- Когда флажок установлен, выполнение сценария приостановится, чтобы вы могли проверить и при необходимости откорректировать разметку документа на области. При этом в главном окне программы под строкой меню появляется панель желтого цвета с кнопкой **Перейти к следующему шагу** (рис. 9.7). Проверив разметку, нажмите эту кнопку. Панель исчезнет, а сценарий будет выполняться дальше.

Рис. 9.7 ▾ Панель продолжения выполнения сценария

- Когда флажок снят, программа выполняет сценарий без остановки на данном этапе.

Если вы уверены, что шаблон точно попадет на нужные участки изображения (интервалы между строками на бланке достаточно широкие, обла-

сти на шаблоне заданы с определенным запасом по ширине и высоте, а оригинал в сканер вы всегда укладываете одинаково), снимите флажок. Если же вы предполагаете, что при обработке отдельных документов области иногда придется корректировать, установите флажок **Проверять и редактировать области вручную**. В общем случае, так надежнее, хотя при выполнении сценария каждый раз нужно будет нажимать кнопку **Перейти к следующему шагу**.

Настроив загрузку шаблона областей, нажмите в диалоге **Шаг 2. Загрузить шаблон областей** кнопку **OK**. Диалог закроется, а вы вернетесь к **Мастеру сценариев**.

6. Теперь выберите в **Мастере сценариев** следующий шаг. Нажмите в левой части окна ссылку **Распознать документ FineReader**. Очередной, третий шаг добавится в сценарий (рис. 9.8).

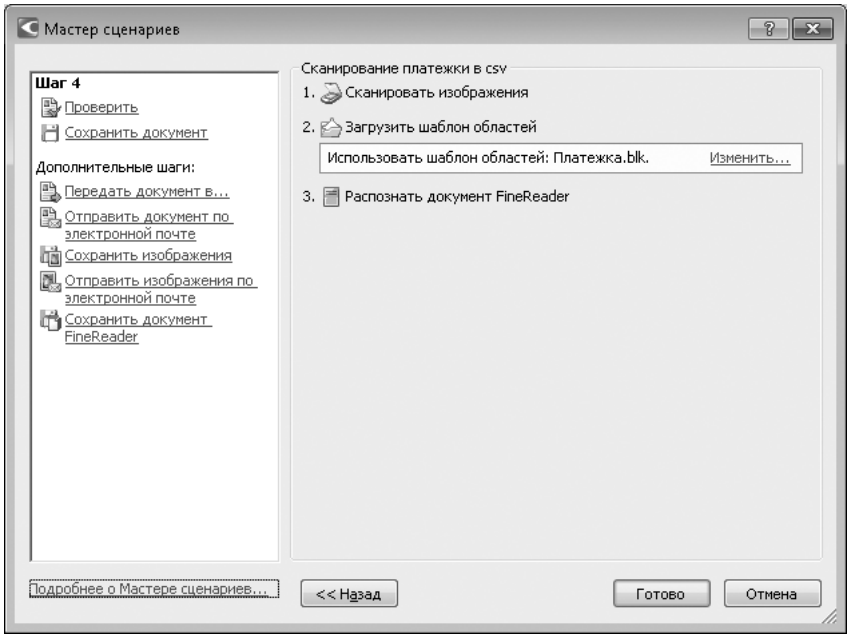


Рис. 9.8 ▾ В сценарий добавлен третий шаг

В качестве четвертого шага Мастер предлагает на выбор два основных действия:

- проверить;
- сохранить документ.

7. Выберите в левой части окна ссылку **Проверить**. В сценарий добавится очередной шаг.
8. Настройте проверку распознанного документа. Щелкните кнопкой мыши на ссылке **Изменить** напротив четвертого шага в правой части окна

Мастера. Откроется диалог, в котором уточняются опции проверки распознанного документа (рис. 9.9).

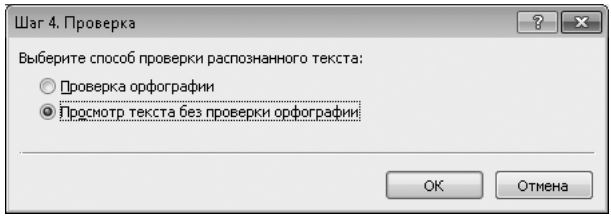


Рис. 9.9 ▾ Настройка проверки

В нашем конкретном примере проверка орфографии вряд ли целесообразна: номера счетов и другие реквизиты, состоящие из цифр, проверить по словарю нельзя. Фамилии плательщиков проверять тоже бесполезно – написание имен собственных далеко не всегда подчиняется правилам орфографии. Все-таки стоит просмотреть распознанный текст перед сохранением: явные ошибки человек уловит чисто интуитивно.

9. Установите переключатель **Выберите способ проверки распознанного текста** в положение **Просмотр текста без проверки орфографии**. Нажмите кнопку **OK**. Диалог закроется, а вам остается задать последний шаг сценария.

   В таком случае, распознав страницу, программа приостановит сценарий и покажет ее в окне **Текст**. В главном окне программы под строкой меню появится панель с сообщением **Проверка. По завершении щелкните** и кнопкой **Перейти к следующему шагу**. Визуально оценив результат распознавания и внеся, при необходимости, изменения в распознанный текст, нажмите кнопку **Перейти к следующему шагу**. Выполнение сценария продолжится.
10. В качестве последнего шага выберите в левой части окна Мастера ссылку **Сохранить документ**. Это действие будет добавлено в сценарий.
11. В раскрывающемся списке выберите формат файла, в который должен сохраняться распознанный текст. В нашем случае это **Документ CSV (*.csv)**, как показано на рис. 9.10.

Сценарий почти готов. Осталось уточнить, в какую папку и под каким именем будут сохраняться файлы с распознанными данными.

1. В правой части окна Мастера сценариев щелкните кнопкой мыши на ссылке **Изменить...** около записи **Сохранить страницы**. Откроется диалог настройки сохранения файлов (рис. 9.11).

   По умолчанию переключатель стоит в положении **Запрашивать имена файлов при сохранении**, а все остальные элементы управления неактивны. В этом случае в конце выполнения сценария появится диалог, в котором вам предлагается выбрать формат файла, указать папку для сохранения и имя файла.

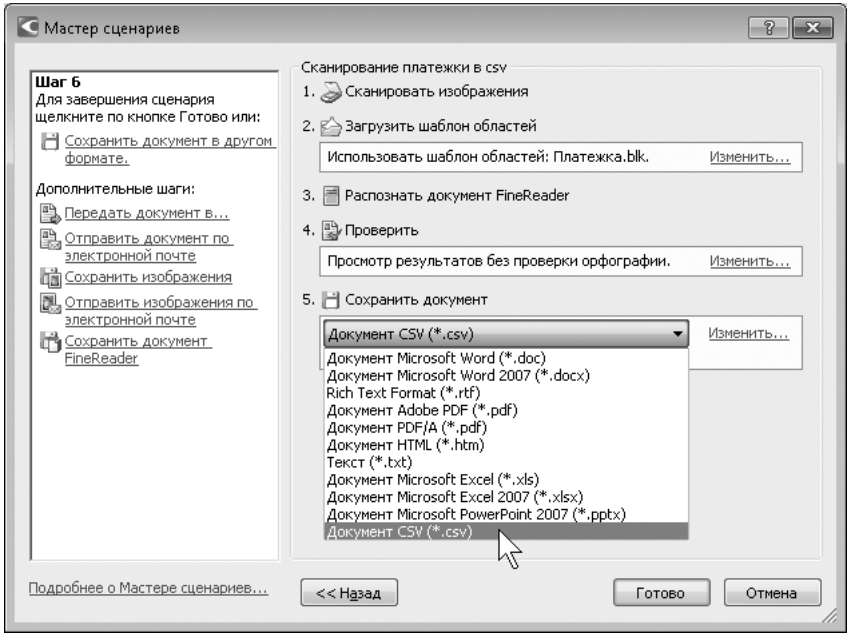


Рис. 9.10 ▾ Выбор формата сохраняемого файла

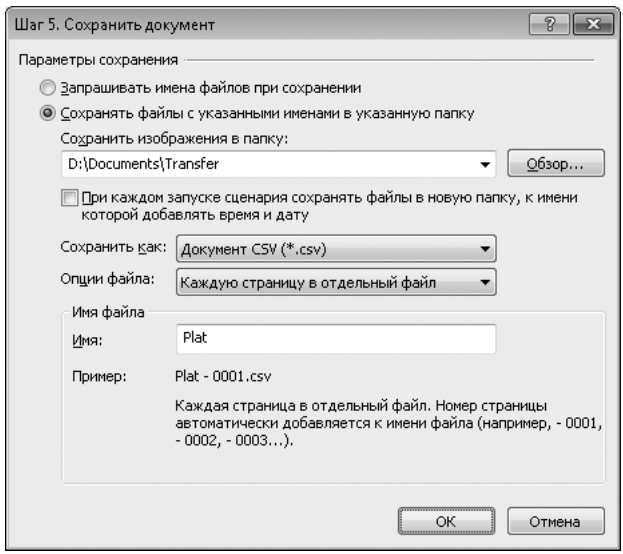


Рис. 9.11 ▾ Настройка параметров сохранения страниц

2. Чтобы по сценарию файлы сохранялись автоматически, без запросов, установите переключатель в положение **Сохранять файлы с указанными именами в указанную папку:**.
3. Введите в поле **Сохранить изображения в папку:** путь к папке, куда по сценарию должны сохраняться файлы, либо нажмите кнопку **Обзор** и укажите нужную папку в дереве.
4. В раскрывающемся списке **Сохранить как:** выберите формат файла для сохранения результатов распознавания.
5. В раскрывающемся списке **Опции файла:** выберите один из трех вариантов сохранения:
   - **Все страницы в один файл**;
   - **Каждая страница в отдельный файл**;
   - **Пофайловое деление по пустым страницам**. В этом случае все страницы документа FineReader, идущие до первой пустой страницы, будут сохранены в один файл. Страницы, находящиеся между этой пустой страницей и до следующей, – в другой файл и т. д. Опция удобна при использовании сканера с автоподатчиком: чтобы каждый из сканируемых подряд многостраничных документов сохранялся в отдельный файл, достаточно вложить между этими оригиналами по чистому листу бумаги.
6. В поле **Имя:** введите имя файла. Если все страницы документа FineReader сохраняются в один файл, то введенное имя и будет присвоено этому файлу. Если в раскрывающемся списке **Опции файла:** выбрано сохранение в несколько файлов, то каждому файлу будет автоматически присвоено имя, составленное из указанного имени и порядкового номера. Пример того, как будут выглядеть имена файлов в каждом конкретном случае, показывается сразу под полем **Имя:**.

   В нашем примере был выбран вариант сохранения **Каждая страница в отдельный файл** и введено имя `Plat`. Файлам в таком случае будут присваиваться имена **Plat – 0001.csv**, **Plat – 0002.csv** и т. д.

При каждом очередном запуске сценария файлы вновь будут помещаться в ту же папку, и им снова будут присваиваться те же имена. Все файлы с такими же именами, сохраненные в указанную папку при предыдущем выполнении сценария, будут переписаны новыми.

В нашем конкретном примере это и не страшно. Предполагается, что после сканирования пачки платежных поручений и завершения сценария вы импортируете все полученные файлы в базу данных, а затем удалите их.

Чаще, наоборот, важно не допустить утраты сохраненных ранее файлов. В этом случае установите флажок **При каждом запуске сценария сохранять файлы в новую папку, к имени которой добавлять время и дату**. Выполняя сценарий, программа FineReader создаст внутри папки, указанной в поле

Сохранить изображения в папку:, папку с именем наподобие **Результаты экспорта 15.12.2009 13.33.43** и сохранит все файлы в нее. При следующем запуске сценария будет создана еще одна папка, имя которой образовано из текущей даты и времени, допустим **Результаты экспорта 16.12.2009 10.17.05**, и так каждый раз.

Настроив параметры сохранения, нажмите в диалоге **Шаг 5. Сохранить документ** кнопку **ОК**. Диалог закроется, а вы вернетесь в **Мастер сценариев**. Теперь в нем собран и настроен весь сценарий из пяти шагов (рис. 9.12).

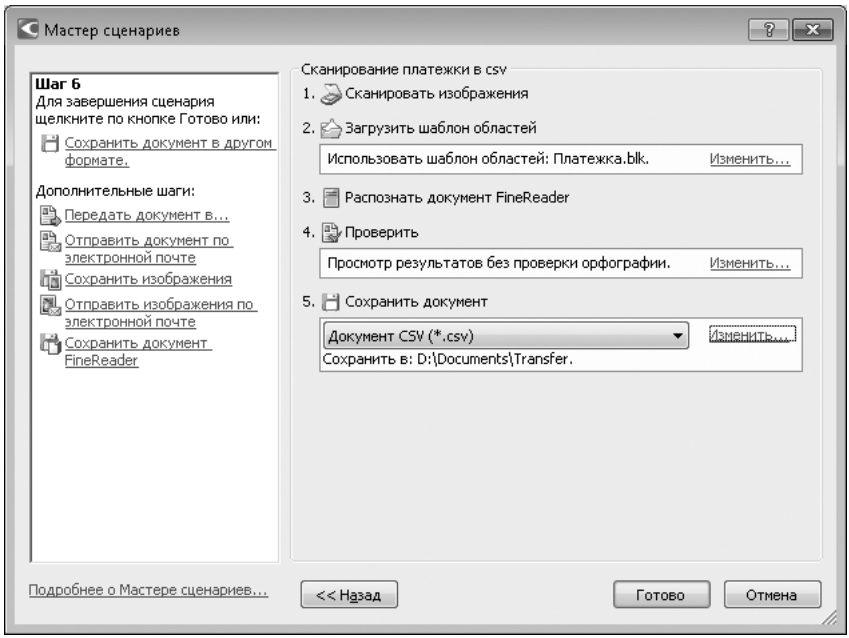


Рис. 9.12 ▾ Сценарий готов

Чтобы сохранить созданный сценарий и завершить работу мастера, нажмите в окне **Мастера сценариев** кнопку **Готово**. Новый пользовательский сценарий появится в списке в окне **Менеджера сценариев** (рис. 9.13).

Выберите этот сценарий в списке в левой части окна **Менеджера сценариев**. В правой части окна будут перечислены шаги, входящие в сценарий. Для проверки запустите созданный сценарий: нажмите кнопку **Запуск** на панели инструментов **Менеджера сценариев** или кнопку **Запустить** в нижней части окна.

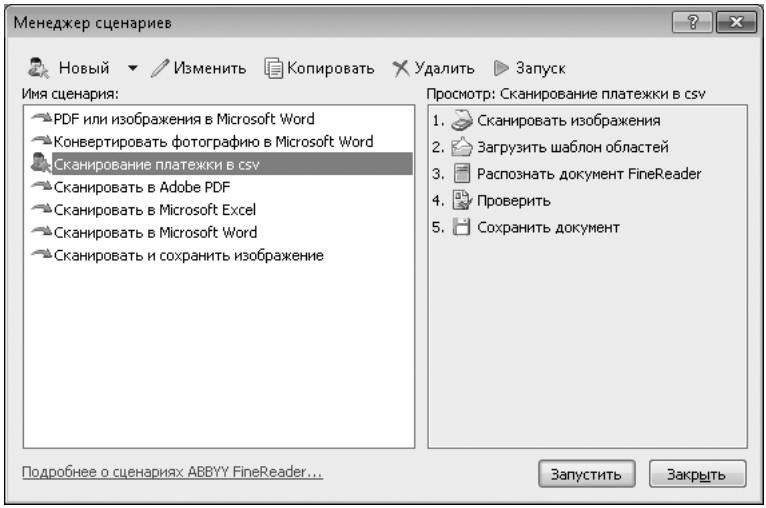


Рис. 9.13 ▾ Новый сценарий появился в **Менеджере сценариев**

Другие действия сценариев

Рассмотрим остальные действия, которые могут быть заложены в сценарии. После добавления шага в сценарий в правой части окна Мастера сценариев рядом с названием действия отображается ссылка **Изменить**. По ссылке **Изменить** открывается диалоговое окно, в котором настраиваются параметры шага. Ссылка **Изменить** доступна не для всех действий – например, для шага **Сканировать** дополнительные параметры не предусмотрены.

- Действие **Открыть изображение**. По ссылке **Изменить** открывается диалог настройки **Шаг 1. Открыть изображения** (рис. 9.14).
  - По умолчанию переключатель установлен в положение **Запрашивать имена файлов изображении при запуске сценария**. В этом случае при запуске сценария появится диалог открытия файлов (рис. 2.3). Вы всякий раз должны будете указать файлы, которые следует открыть.

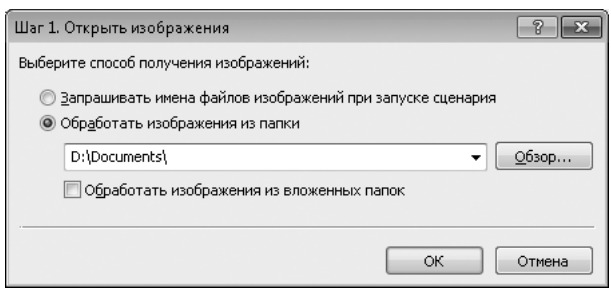


Рис. 9.14 ▾ Диалог **Шаг 1. Открыть изображения**

- Когда переключатель установлен в положение **Обработать изображения из папки:**, при выполнении сценария автоматически будут открываться все файлы изображений из папки, указанной в текстовом поле ввода. Вы можете либо непосредственно ввести путь к этой папке в поле, либо нажать кнопку **Обзор** и выбрать папку в дереве файлов и папок.
- При установленном флажке **Обработать изображения из вложенных папок** будут открыты также все файлы изображений из всех папок, лежащих внутри указанной.

❑ Действие **Открыть документ FineReader**. Настройки этого действия предусматривают три варианта (рис. 9.15).

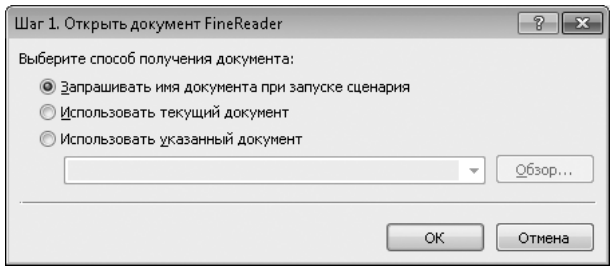


Рис. 9.15 ▾ Диалог **Шаг 1. Открыть документ FineReader**

Чтобы выбрать способ получения документа, установите переключатель в нужное положение:

- **Запрашивать имя документа при запуске сценария**. Это настройка по умолчанию;
- **Использовать текущий документ**. Будет использоваться текущий документ;
- **Использовать указанный документ:**. Имя документа и путь к нему указываются в текстовом поле ввода под переключателем.

Возможное применение сценариев, начинающихся с шага **Открыть документ FineReader**, – отложенное распознавание заранее подготовленных документов на «маломощных» компьютерах. Например, вы отключаете автоматическое распознавание в диалоге **Опции** и быстро сканируете многостраничный документ. На скорость получения изображений быстродействие компьютера практически не влияет. Затем вы запускаете сценарий, по которому происходят распознавание текущего документа FineReader и сохранение результатов в документ Word или PDF, а сами идете заниматься другими делами. Без использования сценария пришлось бы дожидаться окончания распознавания, чтобы потом запустить сохранение в формат **DOC** или **PDF**, – на «слабом» компьютере это тоже длительный процесс.

Когда в качестве второго шага выбрано действие **Распознать документ FineReader**, происходят полностью автоматический анализ изображений и их распознавание без возможности вмешательства. Именно так работают встроенные сценарии.

Альтернатива – в качестве второго шага запрограммировать действия **Загрузить шаблон областей** (этот вариант мы рассмотрели на примере) или **Анализ макета страницы**, а распознавание сделать третьим шагом. В этом случае появляется возможность просмотреть и откорректировать результат анализа до распознавания страницы.

- Действие **Анализ макета страницы** предусматривает два варианта (рис. 9.16). Для выбора установите переключатель в одно из двух положений:
  - **Анализировать страницы автоматически, а затем вносить исправления вручную (рекомендуется)**;
  - **Выделять области вручную**.

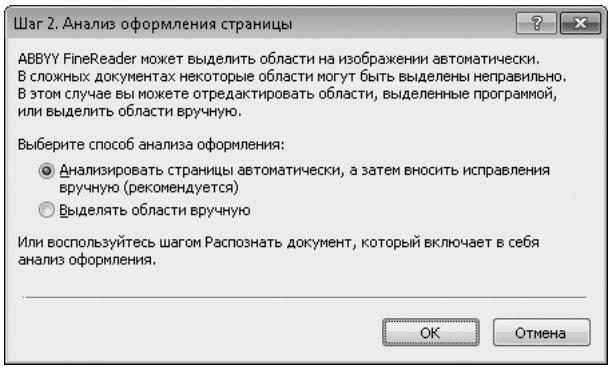


Рис. 9.16 ▾ Диалог **Шаг 2. Анализ оформления страницы**

В первом случае на изображении области будут размечены автоматически. Вам останется убедиться, что программа правильно разметила изображение, и, при необходимости, внести свои коррективы. Во втором случае анализ изображения проводиться не будет, и вы должны будете разметить области на изображении вручную.

В обоих случаях на этом шаге выполнение сценария приостановится. В главном окне программы появится панель с кнопкой **Перейти к следующему шагу.** Чтобы после проверки и корректировки областей продолжить выполнение сценария, нужно нажать эту кнопку.

Если по характеру оригиналов (много иллюстраций, формулы, таблицы и т. д.) вы предполагаете, что в ходе анализа и разметки областей программа может допустить ошибки, выберите в качестве второго шага сценария **Анализ макета страницы**.

Следующим шагом после загрузки шаблона или анализа страницы должно стать распознавание. Для этого действия дополнительные настройки не предусмотрены.

Последним шагом сценария обычно являются передача распознанного документа в другое приложение, сохранение его в файл, отправка его как вложения в сообщение электронной почты. Обратите внимание, что в качестве последнего шага можно указать в любой последовательности и в любом сочетании несколько действий, например **Передать документ в...**, **Сохранить изображения** и **Сохранить документ FineReader**. Выполняя сценарий, программа совершит все указанные действия.

Кроме того, при создании сценария вы можете завершить его на любом из промежуточных шагов. Например, если вы настроите шаг **Распознать документ FineReader** и нажмете в Мастере сценариев кнопку **Готово**, этот шаг станет в сценарии последним. Выполняя такой сценарий, программа распознает документ, отобразит его в окне **Текст** и предоставит вам совершать дальнейшие действия вручную.

Менеджер сценариев

Менеджер сценариев (меню **Сервис ➢ Менеджер сценариев**) служит для управления сценариями. В левой части окна **Менеджера сценариев** перечислены все существующие сценарии, как встроенные, так и пользовательские.

Встроенные сценарии можно только копировать или запустить. Любой из пользовательских сценариев можно изменить, копировать, переименовать, удалить или запустить.

В верхней части окна **Менеджера сценариев** находится панель инструментов. На ней расположены кнопки для создания, изменения, копирования, удаления и запуска сценариев.

Изменить, копировать, переименовать, удалить или запустить сценарий можно также с помощью команд контекстного меню. Для вызова контекстного меню щелкните правой кнопкой мыши на названии сценария.

Копирование сценария

Чтобы скопировать существующий сценарий, выберите его в списке в левой части окна **Менеджера сценариев** и нажмите кнопку **Копировать** на панели инструментов. Иначе щелкните правой кнопкой мыши на названии сценария и в контекстном меню выберите команду **Копировать**. Третий способ – выберите сценарий в списке и нажмите сочетание клавиш **Ctrl+Shift+N**. В списке появится копия выбранного сценария. По умолчанию ей присваивается имя исходного сценария с порядковым номером в скобках.

Копирование позволяет быстро создать новый сценарий на основе существующего, состоящий из тех же шагов, но отличающийся некоторыми параметрами. Вместо создания нового сценария «с нуля» достаточно скопировать уже проверенный и испытанный образец, а затем изменить в нем отдельные

настройки. Как правило, эти настройки касаются имен файлов и опций сохранения.

Копию сценария желательно переименовать – пусть из названия будет ясно, для работы с какими документами этот сценарий предназначен. Чтобы переименовать сценарий, два раза щелкните на его названии кнопкой мыши, или щелкните правой кнопкой мыши и в контекстном меню выберите команду **Переименовать**. На месте названия сценария появится текстовое поле ввода. Введите новое имя сценария и щелкните кнопкой мыши в любом месте окна. Сценарий переименован.

Например, вы уже создали сценарий для распознавания платежных поручений, показанный в предыдущей главе. Теперь вы хотите создавать сценарий для сканирования и распознавания других бланков с использованием другого шаблона областей и сохранения результатов в таблицу Microsoft Excel.

В таком случае создайте и сохраните шаблон областей для нового бланка. Скопируйте пользовательский сценарий **Сканирование платежки в csv**. Переименуйте копию, например в **Сканирование накладной в XLS**. Затем измените эту копию так, чтобы по ней загружался другой шаблон областей, а документ сохранялся в файл формата **XLS**.

Изменение сценария

Для изменения сценария выберите этот сценарий и нажмите на панели инструментов кнопку **Изменить**. Другой способ – выберите команду **Изменить** в контекстном меню или нажмите сочетание клавиш **Ctrl+M**. Откроется уже знакомое окно **Мастера сценариев** (рис. 9.13).

Каждое нажатие кнопки **Назад** удаляет из сценария последний шаг. После этого вы можете выбрать в качестве этого шага другое действие, доступное в левой части окна Мастера. Действие всегда вносится в сценарий с параметрами по умолчанию, и его настройку придется выполнить полностью.

Чтобы изменить параметры какого-либо шага, щелкните кнопкой мыши на ссылке **Изменить** рядом с записью об этом шаге в правой части окна Мастера. Откроется диалог настройки соответствующего действия. Примеры таких диалогов показаны на рис. 9.9, 9.11, 9.14–9.16.

Закончив редактирование сценария, нажмите в окне **Мастера сценариев** кнопку **Готово**. Все изменения в сценарии будут сохранены.

Удаление сценария

Чтобы удалить ненужный пользовательский сценарий, выберите этот сценарий и нажмите на панели инструментов кнопку **Удалить**. Другие способы – выберите команду **Удалить** в контекстном меню, или нажмите клавишу **Del**. Сценарий будет удален.

Экспорт и импорт сценариев

Экспорт сценария – запись сценария в отдельный файл. Этот файл специального формата FTA содержит информацию о последовательности шагов и их

параметрах. Импорт сценария – загрузка сведений о сценарии из такого файла. Один сценарий – один файл FTA. Экспорт и импорт сценариев преследуют две основные цели.

- Сценарий, созданный на одном компьютере, можно экспортировать в файл, скопировать этот файл по сети или на флэш-диск, а потом импортировать в программу FineReader, установленную на другом компьютере. Гарантированно сценарий можно импортировать в программу FineReader такой же или более новой версии, но не в программу предыдущей версии.
- Сценарии можно экспортировать в файлы и сохранить эти файлы на случай переустановки всей системы или программы FineReader. После переустановки вы импортируете сохраненные сценарии, вместо того чтобы создавать их заново.

Сама процедура экспорта и импорта чрезвычайно проста.

1. Чтобы экспортировать сценарий, выберите этот сценарий в списке в левой части окна **Менеджера сценариев** и нажмите стрелку рядом с кнопкой **Новый** на панели инструментов. Откроется короткое меню, состоящее всего из двух пунктов (рис. 9.17).
2. Выберите в этом меню команду **Экспорт...**. Откроется стандартное диалоговое окно сохранения файла (рис. 9.18).
3. Выберите в диалоговом окне **Экспорт сценария** нужный диск и папку. По умолчанию в качестве имени файла предлагается название экспорти-

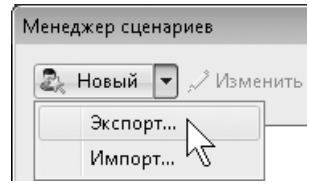


Рис. 9.17 ▾ Меню экспорта/импорта сценариев

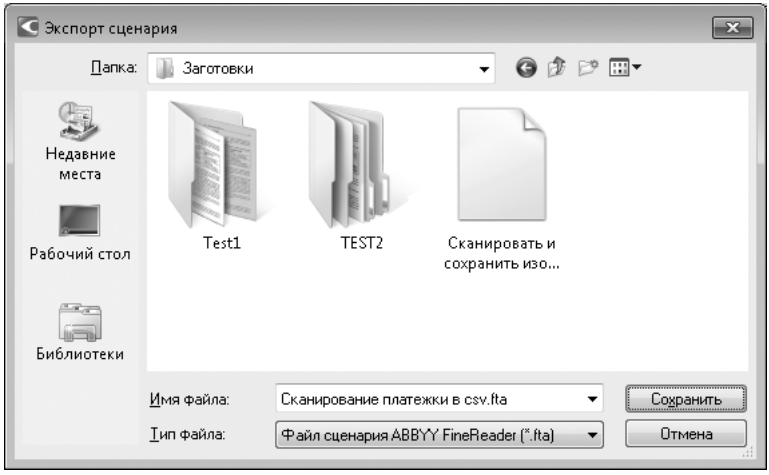


Рис. 9.18 ▾ Экспорт сценария

руемого сценария, и файлу присваивается расширение **FTA**. Нажмите кнопку **Сохранить**. Сценарий будет экспортирован в файл с заданным именем.

Для импорта сценария нажмите на стрелку рядом с кнопкой **Новый** и в выпадающем меню (рис. 9.17) выберите команду **Импорт**. Откроется диалоговое окно открытия файла. Выберите в нем нужный файл сценария и нажмите кнопку **Открыть**. Импортированный сценарий появится в списке в левой части окна **Менеджера сценариев**.

ПРИМЕЧАНИЕ

Если сценарий использует шаблон областей, при переносе его с компьютера на компьютер нужно скопировать также и файл шаблона областей. На другом компьютере следует изменить импортированный сценарий и указать в нем правильный путь к файлу шаблона, а также пути для открытия и сохранения файлов.

Использование сценариев

Для запуска любого сценария откройте окно **Менеджера сценариев**: меню **Сервис ➢ Менеджер сценариев** (рис. 9.19), или сочетание клавиш **Ctrl+T**. В окне **Менеджера сценариев** выберите нужный сценарий и нажмите кнопку **Запуск** на панели инструментов, или кнопку **Запустить** в нижней части окна. Кроме того, команда **Запуск** присутствует в контекстном меню сценария, открывающемся при щелчке правой кнопкой мыши на названии сценария.

Для решения рутинных задач вы можете создать несколько пользовательских сценариев. При этом надо помнить, что сам сценарий содержит лишь *часть* настроек, с которыми он будет выполняться. Другая часть настроек задается в диалоговом окне **Опции**. Они действуют до тех пор, пока вы в очередной

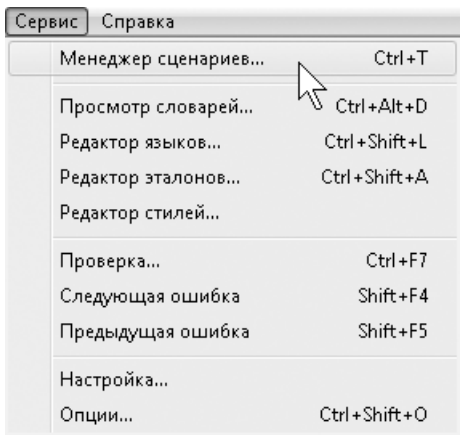


Рис. 9.19 ▾ Вызов **Менеджера сценариев**

раз прямо не измените их через диалог **Опции**, и не меняются при вызове тех или иных сценариев.

На конечный результат совместно влияют настройки сценария, общие настройки программы, а при использовании шаблона областей – еще и настройки этого шаблона. Обобщим, какие из настроек содержатся непосредственно в сценариях, а какие – в настройках программы в целом.

- В сценариях хранятся:
  - пути и имена открываемых и сохраняемых файлов;
  - включение/выключение проверки орфографии в распознанном документе;
  - настройка сохранения всех станиц в один файл или каждой страницы в отдельный файл;
  - выбор формата файла для сохранения изображений.
- В общих настройках программы FineReader (диалог **Опции**) задаются:
  - язык (языки) распознавания для документа в целом;
  - тип печати (авто, пишущая машинка, факс) для документа в целом;
  - функции обработки изображения;
  - эталоны;
  - опции сохранения распознанного документа в различных форматах и передачи его в другие приложения.
- При загрузке шаблона областей этот шаблон переопределяет некоторые общие настройки программы, но только применительно к определенным областям:
  - язык (языки) распознавания для отдельных областей;
  - тип печати (авто, пишущая машинка, факс) для отдельных областей;
  - инверсия и зеркальное отражение изображения для отдельных областей.

Если помнить об этом «разделении полномочий», некоторых проблем при использовании сценариев удастся избежать. Проблемы состоят в том, что после выполнения совершенно «правильного» сценария результаты иногда отличаются от ожидаемых.

Например, вы создали сценарий, по которому фотографии страниц из указанной папки должны распознаваться и сохраняться в документ Word как точные копии, со всеми полутоновыми иллюстрациями, и все страницы сохраняются в один файл. Сценарий создавался, когда в диалоге **Опции** на вкладке **Сохранить ➢ RTF/DOC/DOCX** в раскрывающемся списке **Оформление** было выбрано значение **Точная копия**. При опробовании сценария результат вас полностью устроил: выходной документ выглядел именно так, как надо.

После этого, работая с каким-то очередным документом, вы изменили одну из настроек – на вкладке **Сохранить ➢ RTF/DOC/DOCX** диалогового окна **Опции** выбрали значение **Простой текст**. Во всех документах, сохраненных после этого, текст будет отформатирован стандартным шрифтом и выровнен по левому краю страницы, а картинки не появятся вовсе, поскольку изменились опции сохранения в документ Microsoft Word.

Избежать подобных «накладок» очень просто. Достаточно перед запуском сценария вернуть настройки программы к тем, которые требуются для получения нужного результата. Чтобы каждый раз не вспоминать, какие настройки вы меняли и какими они должны быть для данного сценария, обратитесь к функции сохранения настроек программы.

Если какой-либо сценарий должен выполняться с особенными настройками программы FineReader, заранее сохраните такие настройки в файл. Пока все настройки программы подходят для выполнения данного сценария, откройте диалоговое окно **Опции** и на вкладке **Дополнительные** (рис. 3.14) нажмите кнопку **Сохранить опции...**. В открывшемся окне сохранения файла укажите такое имя файла, чтобы было понятно, к каким сценариям или случаям распознавания этот набор настроек подходит. Изменив настройки применительно к другим работам, вновь сохраните файл с набором опций и т. д. В результате вы сохраните несколько наборов настроек (файлов с расширением FBT), подходящих для разных повторяющихся ситуаций.

Перед тем как вызвать сценарий, откройте диалоговое окно **Опции** и на вкладке **Дополнительные** нажмите кнопку **Загрузить опции...**. На экране появится диалоговое окно открытия файла. Выберите в нем файл настроек (с расширением **FBT**) с настройками, нужными для работы по данному сценарию, и нажмите кнопку **Открыть**. В результате программа начнет использовать набор настроек, записанных в указанном файле.

Резюме

Пользовательские сценарии упрощают выполнение регулярно повторяющихся заданий. Выгода от применения сценариев особенно заметна при работе с автоподатчиком документов, использовании шаблонов областей, распознавании стандартных документов. Время, затраченное на создание и настройку сценария, обычно оправдывается уже при обработке первого десятка однотипных документов.

В этой главе мы коснулись проблемы автоматического распознавания стандартных документов наподобие платежных извещений или накладных. При небольшом объеме работы программа ABBYY FineReader благодаря сценариям вполне справляется и с такой задачей. Для массовой автоматической обработки форм компания ABBYY предлагает другое, специализированное решение – программу ABBYY FormReader.

Если вы не полностью доверяете автоматическому анализу изображения или проверке орфографии в документах со сложной структурой, в сценарий целесообразно заложить шаги с «остановками» на этих этапах: действия **Анализ макета страницы** и **Проверить**. Когда сценарий требует особых настроек программы, удобно заранее сохранить такие настройки в файл, а затем загружать их перед запуском сценария.

На этом мы фактически закончили знакомство с программой FineReader. В последней главе речь пойдет об отдельном приложении, входящем в состав этой программы.

Глава 10 ABBYY Screenshot Reader

ABBYY Screenshot Reader – программа для распознавания «снимков экрана», то есть всего, что отображается на экране компьютера. Слово Screenshot в переводе с английского означает «снимок экрана». Это приложение обладает собственным, очень простым интерфейсом с минимумом настроек. При работе оно использует компоненты программы FineReader.

Данная программа способна захватывать изображение с экрана, распознавать содержащиеся в этом изображении символы, а затем передавать результаты распознавания в буфер обмена, приложения Microsoft Office или сохранять в файл. Кроме того, программа Screenshot Reader может без распознавания передать захваченное изображение в буфер обмена, программу FineReader, сохранить его в файл или прикрепить к сообщению электронной почты.

Интерфейс и настройки программы

Запустите программу ABBYY Screenshot Reader: **Пуск ➢ Все программы ➢ ABBYY FineReader 10 ➢ ABBYY Screenshot Reader.** Окно программы мало по размеру (рис. 10.1).

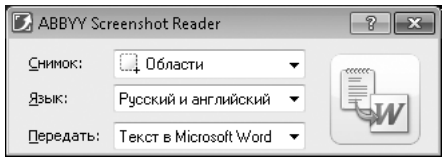


Рис. 10.1 ▾ Окно ABBYY Screenshot Reader

В этом окне присутствуют четыре элемента управления:

- раскрывающийся список **Снимок:**. Из него вы можете выбрать способ захвата изображения с экрана: выделенной области, окна, всего экрана или всего экрана с отсрочкой;
- раскрывающийся список **Язык:** служит для выбора языков распознавания;
- раскрывающийся список **Передать:** позволяет указать действие, которое программа выполнит после захвата изображения. Фактически каждое из этих действий является коротким встроенным сценарием;
- при нажатии кнопки **Сделать снимок**, находящейся в правой части окна, программа делает снимок экрана.

Все время, пока программа ABBYY Screenshot Reader запущена, в области уведомлений панели задач отображается ее значок. Даже если закрыть окно программы, она продолжает работать в фоновом режиме, и этот значок по-прежнему отображается в области уведомлений. Чтобы вновь отобразить окно программы на экране, щелкните на значке кнопкой мыши. Чтобы завершить работу программы, щелкните правой кнопкой мыши на значке в области уведомлений и в открывшемся контекстном меню выберите команду **Выход**.

Для изменения настроек программы вызовите контекстное меню и перейдите к пункту **Настройки**. Откроется вложенное меню (рис. 10.2).

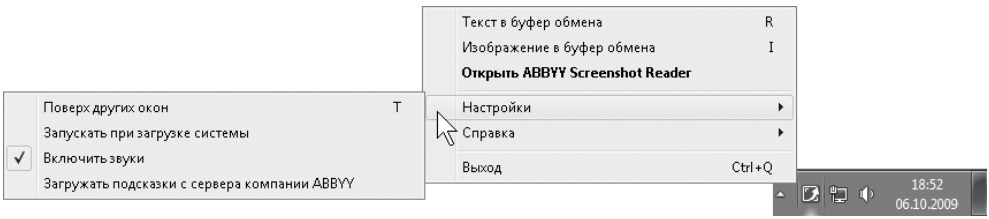


Рис. 10.2 ▾ Меню настройки

Меню настройки содержит четыре пункта. Включенные функции отмечены флажками. Назначение настроек ясно из названий пунктов меню:

- **Поверх других окон**. Если флажок установлен, окно ABBYY Screenshot Reader постоянно отображается поверх других окон;
- **Запускать при загрузке системы**. Когда флажок установлен, программа ABBYY Screenshot Reader запускается автоматически при загрузке ОС Windows;
- **Включить звуки**. Когда флажок установлен, при завершении копирования в буфер обмена звучит сигнал;
- **Загружать подсказки с сервера компании ABBYY**.

Чтобы задействовать функцию, щелкните мышью на соответствующем пункте меню. Напротив него будет установлен флажок, а само меню закроется. Что-

бы отключить функцию, вновь вызовите меню настройки и еще раз щелкните кнопкой мыши на соответствующем пункте. Флажок будет снят, меню закроется.

Работа с программой

Начиная работать с программой ABBYY Screenshot Reader, выполните три настройки. Для этого воспользуйтесь раскрывающимися списками в окне приложения.

В раскрывающемся списке **Язык:** выберите языки распознавания. По умолчанию задано значение **Авто**: программа автоматически выбирает нужные языки. Также вы можете в раскрывающемся списке **Язык:** указать определенные языки вручную.

Последний элемент списка, **Выбор языков...**, открывает почти такое же диалоговое окно **Редактор языков** (рис. 6.20), как и в программе FineReader. Отличие в том, что из программы Screenshot Reader редактор языков вызывается в режиме «только для чтения» – в нем нельзя создавать или изменять пользовательские языки. Вместе с тем если ранее в программе FineReader был создан какой-либо пользовательский язык, то он доступен и из программы Screenshot Reader.

Решите, каким образом вы хотите получить изображение с экрана. Выберите из раскрывающегося списка **Снимок:** один из способов захвата изображения:

- **Области**. Если выбрать этот вариант, при нажатии кнопки **Сделать снимок** будет получен снимок области, определяемой пользователем;
- **Окна**. При выборе этого варианта захватывается изображение одного из элементов оконного интерфейса Windows: рабочей области какого-либо окна, диалога, панели;
- **Экрана**. При нажатии кнопки **Сделать снимок** будет захвачено изображение всего экрана;
- **Экрана с отсрочкой**. В таком случае при нажатии кнопки **Сделать снимок** также будет сделан снимок всего экрана, но с задержкой в 5 секунд.

В раскрывающемся списке **Передать:** (рис. 10.3) выберите действие, которое программа должна выполнить с полученным снимком. Рисунок на кнопке **Сделать снимок** изменяется в зависимости от того, какое действие выбрано.

Таким образом, вы задали язык распознавания, указали программе способ захвата изображения и то, что программа должна будет сделать с полученным снимком экрана. Теперь рассмотрим собственно работу с программой ABBYY Screenshot Reader.

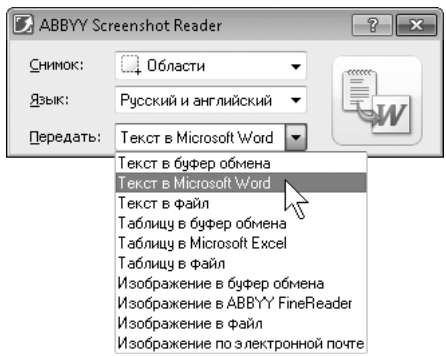


Рис. 10.3 ▾ Выбор действия

Захват изображения

Первый шаг – захват изображения всего экрана или определенной его области. Чтобы выполнить эту операцию, нажмите кнопку **Сделать снимок**. Дальнейшие ваши действия зависят от того, какой метод захвата был предварительно выбран в раскрывающемся списке **Снимок:**.

Если был выбран вариант **Области**, после нажатия кнопки **Сделать снимок** окно программы Screenshot Reader скроется, а указатель мыши превратится в символ графического выделения  .

Нажмите левую кнопку мыши и, удерживая ее, обведите нужный участок экрана. Захватываемый участок показывается пунктирной рамкой (см. рис. 10.12, 10.14). Отпустите кнопку мыши. Программа «сфотографировала» все, что оказалось внутри рамки.

Если в раскрывающемся списке **Снимок:** был выбран вариант **Окна**, после нажатия кнопки **Сделать снимок** окно программы Screenshot Reader скроется, а указатель мыши приобретет вид стрелки с «прицельной рамкой» . Наводя этот указатель на различные объекты на рабочем столе Windows, вы можете выбрать один из них. Выбранный объект заключается в рамку розового цвета, а над ним отображается всплывающее сообщение (рис. 10.4).

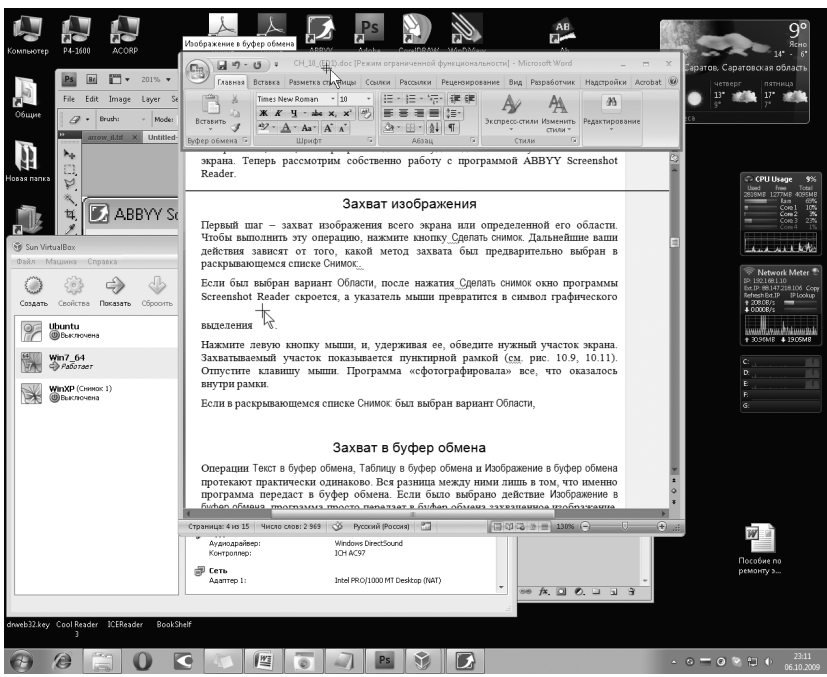


Рис. 10.4 ▾ Выбор окна или другого объекта

Чтобы выбрать какое-либо окно целиком, наведите указатель на его угол или рамку. Для выбора только рабочей области окна поместите указатель мыши над его рабочей областью. Заметим, что данный режим позволяет выбирать даже такие элементы некоторых окон, как панели инструментов, строки меню или заголовков окон, полосы прокрутки и т. п. Например, рис. 10.4 демонстрирует, как получить снимок ленты, строки меню и панели быстрого доступа Microsoft Word 2007. При этом изображения рабочей области окна и остальных объектов интерфейса захвачены не будут.

Щелкните кнопкой мыши, и программа Screenshot Reader сделает снимок выбранного объекта. Описанный режим удобен для получения снимков интересующих вас окон или отдельных их частей.

Если в раскрывающемся списке **Снимок:** был выбран вариант **Экрана**, после нажатия кнопки **Сделать снимок** окно программы Screenshot Reader скроется, и программа выполнит захват изображения всего экрана.

Захват изображения в режиме **Экрана с отсрочкой** требует особого пояснения. Без использования этой функции получить снимки некоторых объектов достаточно сложно. К таким объектам относятся всплывающие сообщения, меню, а также некоторые активные элементы, отображаемые на веб-страницах в окне браузера. Они отображаются на экране, лишь пока вы удерживаете указатель мыши в определенном месте. Достаточно переместить указатель мыши или нажать клавишу, как эти объекты скрываются, и сделать их снимок не удается.

Когда в раскрывающемся списке **Снимок:** выбран вариант **Экрана с отсрочкой**, после нажатия кнопки **Сделать снимок** на экране появляется таймер, ведущий обратный отсчет времени (рис. 10.5).

У вас остается пять секунд на то, чтобы вызвать всплывающее сообщение или другой элемент, снимок которого иным способом получить не удается. По истечении этого времени программа Screenshot Reader автоматически производит захват изображения всего экрана.

Таким образом, программа ABBYY Screenshot Reader получила изображение с экрана компьютера. То, как программа поступит с захваченным изображением, зависит от варианта действий, который вы заранее выбрали из раскрывающегося списка **Передать:** (см. рис. 10.3).

Передача в буфер обмена

Операции **Текст в буфер обмена**, **Таблицу в буфер обмена** и **Изображение в буфер обмена** протекают практически одинаково. Вся разница между ними лишь в том, что именно программа передаст в буфер обмена.

- Если было выбрано действие **Изображение в буфер обмена**, программа просто передает в буфер обмена захваченное изображение. Затем вы сможете вставить его в любое приложение, поддерживающее вставку графики.
- Если было выбрано действие **Текст в буфер обмена**, программа проведет распознавание и затем передаст в буфер обмена распознанный текст.

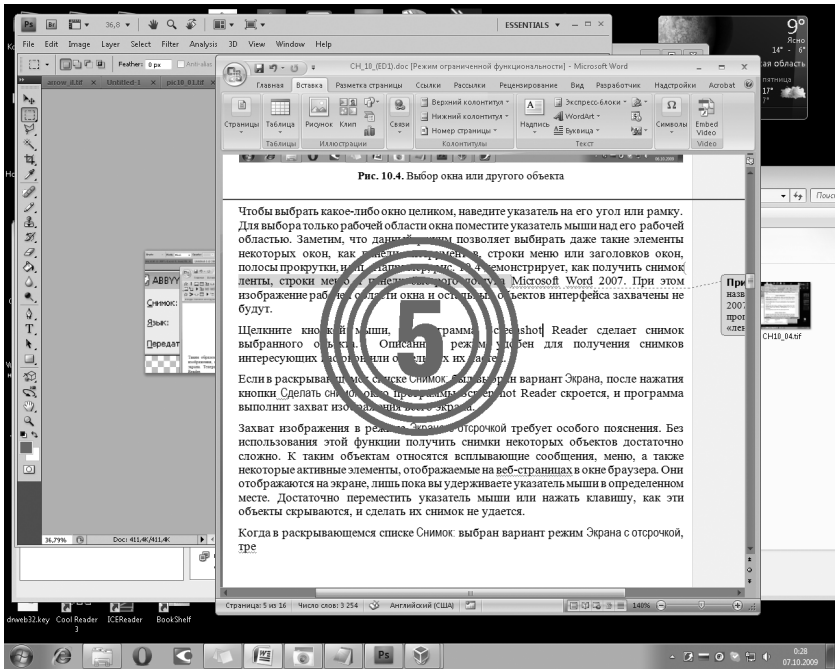


Рис. 10.5 ▾ Захват экрана с отсрочкой

- При выборе действия **Таблицу в буфер обмена** программа анализирует захваченное изображение и пытается выявить в нем структуру таблицы. Если таковая обнаружена, производится распознавание, и распознанная таблица передается в буфер обмена.

В случае если программа не обнаружила в изображении характерных признаков таблицы (наличия столбцов и строк), над значком в области уведомлений появится предупреждение (рис. 10.6). Тем не менее символы, найденные на изображении, будут распознаны и переданы в буфер обмена как простой текст.

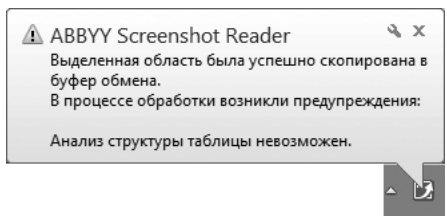


Рис. 10.6 ▾ Предупреждение о невозможности анализа таблицы

Вставку содержимого из буфера обмена поддерживают практически все программы для работы с текстом и изображениями. При этом самые простые текстовые редакторы, например Блокнот, способны вставлять из буфера обмена только текст. Другие программы могут импортировать из буфера обмена и текст, и изображения. Некоторые программы, например Microsoft Word, Microsoft Excel, Corel Draw, поддерживают, кроме того, и вставку таблиц. Как правило, для вставки содержимого буфера обмена в прикладных программах служат команда меню **Правка ➢ Вставить** (**Edit ➢ Paste**) и сочетания клавиш **Ctrl+C** или **Shift+Insert**.

В некоторых текстовых и графических редакторах предусмотрена также команда **Специальная вставка** (**Paste Special**). По этой команде программа предлагает уточнить, как именно следует вставить содержимое буфера обмена: как форматированный текст, текст без форматирования, таблицу, изображение и т. д.

Перейдите в окно приложения с документом, в который вы хотите вставить содержимое буфера обмена. Щелкните кнопкой мыши в том месте документа, куда следует выполнить вставку: в этом месте замигает курсор. Выполните команду вставки. В большинстве приложений Windows стандартной командой вставки является сочетание клавиш **CTRL+C**, хотя в некоторых программах для этой цели могут использоваться другие сочетания клавиш. Текст или изображение из буфера обмена будут помещены в указанное место документа.

Передача в приложения Microsoft Office

Два действия: **Текст в Microsoft Word** и **Таблицу в Microsoft Excel** – передают результат распознавания в приложения Microsoft Office. После захвата с экрана и распознавания текста (таблицы) автоматически запускается названное приложение с новым документом, и в этот документ вставляется распознанный текст (таблица).

Удобство в том, что не нужно отдельно запускать приложение Microsoft Office или создавать новый документ. Выберите действие **Текст в Microsoft Word** и произведите снимок экрана или его области. Через короткое время на экране откроется окно Microsoft Word с документом, в котором содержится распознанный текст.

Операция Изображение в ABBYY FineReader

По этому сценарию после захвата запускается программа ABBYY FineReader, и захваченное с экрана изображение передается в нее. Далее вы сможете выполнить полноценный анализ изображения, распознать области и сохранить результат распознавания в любом из поддерживаемых программой форматов.

Целесообразно выбирать действие **Изображение в ABBYY FineReader**, когда в захватываемом изображении одновременно присутствуют и текст, и таблицы, и иллюстрации, а в результате вы хотите получить достаточно похожую копию. Один из примеров – копирование содержимого сайта или файла PDF – мы рассмотрим чуть позже.

Сохранение в файл

Три операции: **Текст в Файл**, **Таблицу в Файл** и **Изображение в Файл** – завершаются тем, что программа Screenshot Reader предлагает вам сохранить в файл захваченное изображение или результат его распознавания. Различие между этими тремя операциями состоит в типе сохраняемого содержимого и доступных форматах файлов.

Если из раскрывающего списка **Передать:** было выбрано действие **Текст в Файл**, после выделения и захвата участка экрана появляется диалоговое окно **Сохранить текст как...** (рис. 10.7).

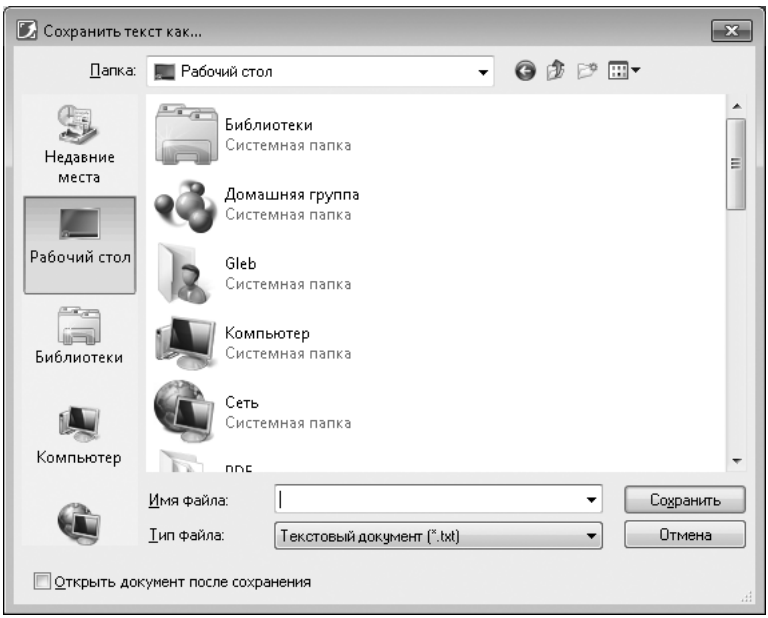


Рис. 10.7 ▾ Сохранение текста

Выберите в окне папку для сохранения файла. В поле **Имя файла:** ведите имя сохраняемого файла. Из раскрывающегося списка **Тип файла:** выберите формат, в котором этот файл должен быть сохранен. Доступными форматами в данном случае являются обычный текстовый файл, текст в формате RTF или документ Microsoft Word.

Когда из раскрывающего списка было выбрано действие **Таблицу в Файл**, после выделения и захвата участка экрана появляется диалоговое окно **Сохранить таблицу как...** (рис. 10.8). Доступными типами файлов для сохранения таблицы являются документ Microsoft Excel (XLS) и документ CSV.

Если программа ABBYY Screenshot Reader не смогла выделить в захваченном изображении структуру таблицы, вместо диалога сохранения появится со-

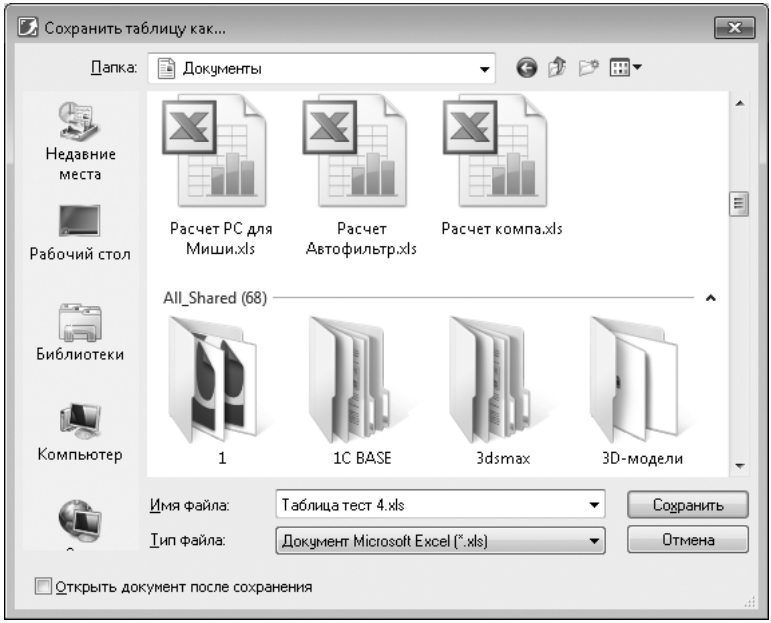


Рис. 10.8 ▾ Сохранение таблицы

общение о том, что анализ структуры таблицы невозможен (рис. 10.9). В таком случае нажмите в сообщении кнопку **Закрыть**.

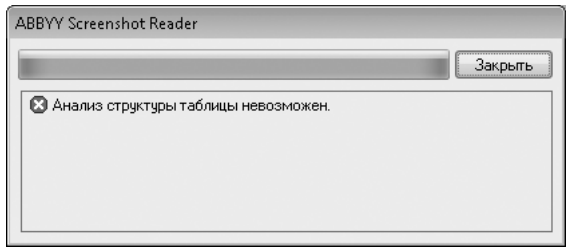


Рис. 10.9 ▾ Сообщение о невозможности анализа таблицы

Если вы считаете, что изображение с экрана может быть распознано именно как таблица, обратитесь к более гибким средствам программы ABBYY FineReader. В окне ABBYY Screenshot Reader выберите из раскрывающегося списка **Передать:** действие **Изображение в ABBYY FineReader** и выполните захват повторно. Изображение будет передано в программу FineReader. Разметьте на захваченном изображении области автоматически или вручную и запустите распознавание таблицы. Затем из программы FineReader сохраните результат распознавания.

При выборе действия **Изображение в Файл** программа ABBYY Screenshot Reader работает как обычный «граббер экрана». Она предлагает сохранить захваченное изображение в файл одного из графических форматов (рис. 10.10). Доступными форматами в данном случае являются JPEG (цветной или в оттенках серого), Bitmap (цветное изображение в файле с расширением BMP, RLE или DIB), PNG (цветной, в оттенках серого, черно-белый).

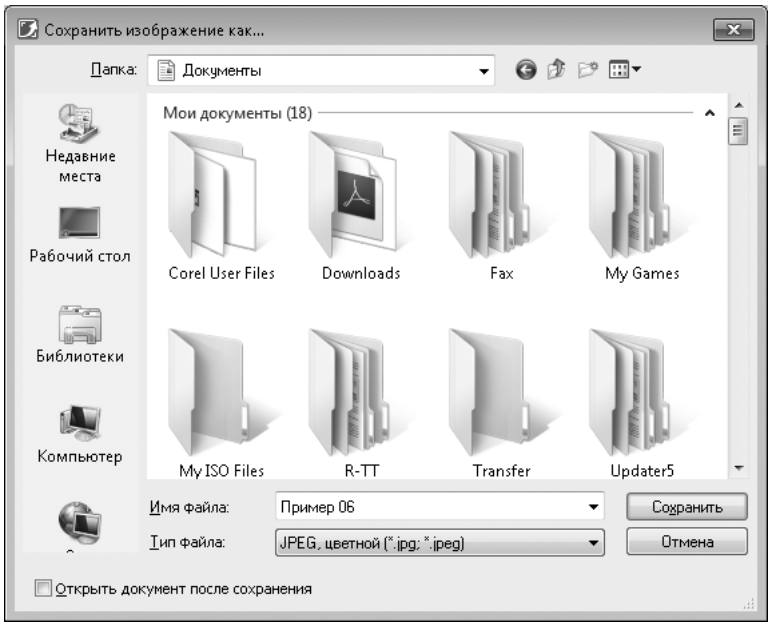


Рис. 10.10 ▾ Сохранение изображения

Если вы хотите сохранить снимок экрана в файл другого формата, вместо действия **Изображение в Файл** выберите действие **Изображение в буфер обмена**. Выполните захват и вставьте изображение из буфера обмена в какой-нибудь графический редактор, например Paint, GIMP или Adobe Photoshop. Сохраните изображение в нужном формате средствами этого графического редактора.

Во всех трех диалоговых окнах сохранения файла (рис. 10.7–10.10) присутствует флажок **Открыть документ после сохранения**. По умолчанию он снят. Если установить этот флажок, после сохранения файл будет сразу же открыт в приложении, связанном с данным типом файлов.

Примеры использования программы

Программе ABBYY Screenshot Reader найдется множество применений, как вполне серьезных, так и не очень. Прежде всего это любые ситуации, когда нужно скопировать какой-нибудь текст из меню, диалогов настройки, всплывающих сообщений и других элементов интерфейса программ. Другое очевидное применение – копирование текста документов, которые иным способом скопировать не удается.

Копирование текста документов с экрана

На протяжении всей книги мы рассматривали работу с документами, форматы которых поддерживаются программой FineReader. Мы открывали такие файлы программой FineReader и далее обрабатывали и распознавали их содержимое. Однако порой нужно скопировать текст из таких источников, которые средствами FineReader открыть невозможно или затруднительно. Программа ABBYY Screenshot Reader решает эту проблему: она распознает любой текст, который отображается на экране.

Файлы PDF могут быть защищены от копирования, редактирования или печати. При попытке открыть такой файл в программе FineReader появляется диалог с запросом пароля, который, скорее всего, вам неизвестен. Если открыть документ в программе Adobe Reader, выделить в нем текст и щелкнуть на выделении правой кнопкой мыши, команда **Копировать** в контекстном меню недоступна. Точно так же невозможно и распечатать подобный документ.

> *ПРИМЕЧАНИЕ*
>
> *Извлекая текст из защищенного от копирования документа, вы, возможно, затрагиваете чьи-то авторские права. Программа предоставляет лишь* техническую возможность, *но ответственность целиком и полностью лежит на том, кто копирует содержимое защищенного документа. На снятие копий исключительно для личного пользования пока еще не обращают серьезного внимания, но за распространение подобных документов правообладатели вполне могут предъявить претензии. Поэтому решение о том, насколько правомерны подобные действия, остается на вашей совести.*

Простому копированию текста с сайтов может помешать применение на веб-страницах технологий Java Script или Flash. Хотя содержимое такой страницы удается читать и прокручивать в окне браузера, при попытке выделить и скопировать текст либо вообще ничего не происходит, либо текст внезапно «убегает» из-под указателя мыши, а страница обновляется. Неудачей заканчиваются и попытки сохранить страницу на диск – например, в сохраненной копии присутствуют только пустые рамки.

«Электронные книги» в виде исполняемых файлов – своеобразные программы. При запуске такой программы на экране появляется окно с текстом, но этот текст нельзя ни выделить, ни скопировать, ни распечатать.

Во всех случаях простейший выход – сделать снимок экрана, а потом распознать то, что на нем изображено. Программа ABBYY Screenshot Reader легко справляется с такой задачей. Единственное условие – для качественного распознавания текст на экране должен быть достаточно крупным.

1. Откройте документ PDF в программе Adobe Reader или веб-страницу из Интернета в браузере. Разверните окно программы Adobe Reader или браузера на весь экран. Установите масштаб изображения **100%** или **по ширине страницы**.
2. Откройте программу Блокнот (**Пуск ➢ Все программы ➢ Стандартные ➢ Блокнот**) или другой текстовый редактор, например WordPad или Microsoft Word.
3. Запустите программу Screenshot Reader и в ней выберите действие **Текст в буфер обмена**. Проверьте, что включен режим отображения поверх остальных окон.
4. Сделайте снимок интересующего вас текста в окне браузера или программы Adobe Reader.
5. Вставьте распознанный текст в документ в окне редактора.
6. Если на экране поместился не весь текст, прокрутите исходный документ в окне браузера и сделайте снимок следующего фрагмента. Вставьте очередной фрагмент в текстовый редактор и т. д.
7. Скопировав все, что требовалось, сохраните текстовый документ или документ Word.

При необходимости вы можете также скопировать из исходного документа иллюстрации (выберите действие **Изображение в буфер обмена**) или таблицы (выберите действие **Таблицу в буфер обмена**) и вставить их в документ Microsoft Word. В результате получается достаточно качественная редактируемая копия защищенного документа.

Список файлов

Иногда нужно получить список имен файлов и сохранить его. Например, вы решили привести в порядок свою коллекцию компакт-дисков и снабдить коробки этикетками с перечислением названий песен, клипов и исполнителей.

Разумеется, можно набрать текст вручную в редакторе, а потом распечатать. Гораздо приятнее и быстрее сгенерировать список автоматически – ведь имена файлов обычно совпадают с названиями песен, клипов и т. д.

Опытные пользователи знают о существовании команды **Dir**. Она запускается из командной строки и выводит список имен всех файлов, находящихся в папке. Однако работать с командной строкой любят далеко не все, а некоторых необходимость набирать какие-то команды и помнить их параметры просто пугает. С помощью снимка экрана задача решается быстро и изящно.

Откройте папку с файлами, список которых вы хотите получить. Выберите в меню **Вид** подходящий режим отображения, например **Таблица**.

ПРИМЕЧАНИЕ

В ОС Windows XP строка меню в Проводнике отображается всегда, а в ОС Windows 7 для вывода строки меню необходимо нажать клавишу ***Alt****.*

При желании настройте детали отображения: меню **Вид ➢ Выбор столбцов в таблице**. Откроется диалоговое окно **Выбор столбцов в таблице** (рис. 10.11).

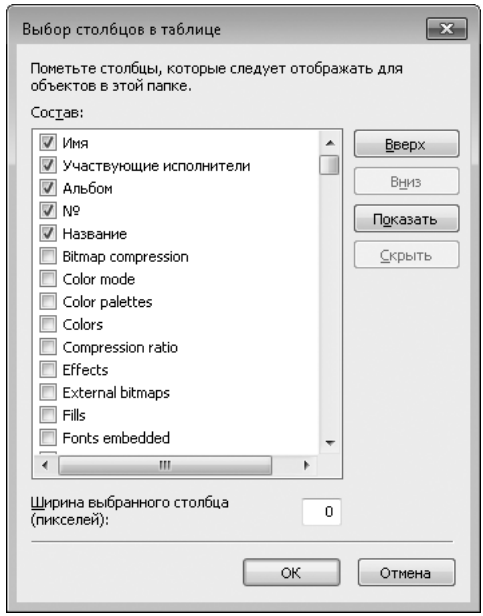


Рис. 10.11 ▾ Выбор столбцов, отображаемых в папке

Установите флажки напротив тех сведений о файлах, которые вы хотите видеть в окне **Проводника Windows**. Для музыкальных записей, например, целесообразно выбрать столбцы **Имя**, **Исполнитель** и **Альбом**. Нажмите кнопку **ОК**, и окно Проводника приобретет примерно такой вид, как на рис. 10.12.

Запустите программу Screenshot Reader и выберите в ней действие **Таблицу в Microsoft Excel**. Нажмите кнопку **Снимок** и обведите список файлов (рис. 10.12).

В открывшемся окне программы Microsoft Excel вы увидите таблицу со списком файлов (рис. 10.13). При желании ее можно откорректировать и оформить: вставить заголовок, номера записей, изменить шрифты, добавить ком-

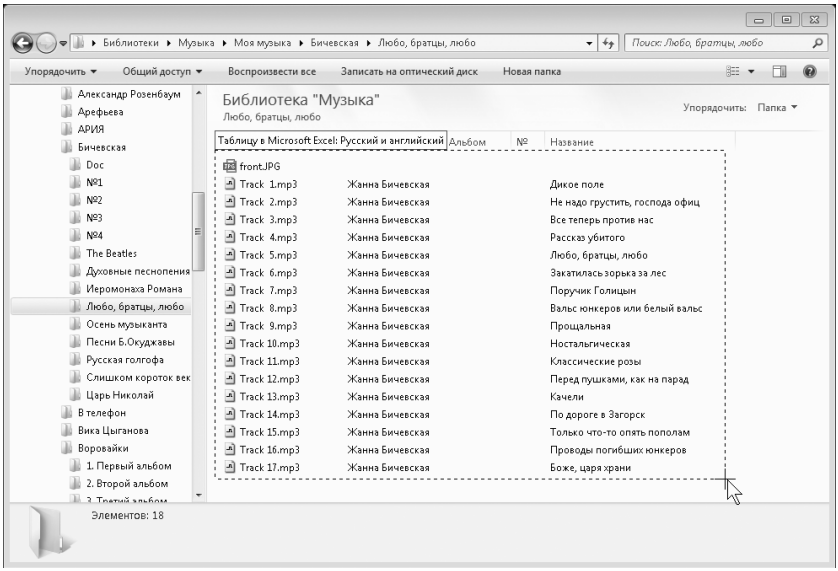


Рис. 10.12 ▾ Захват списка файлов в папке

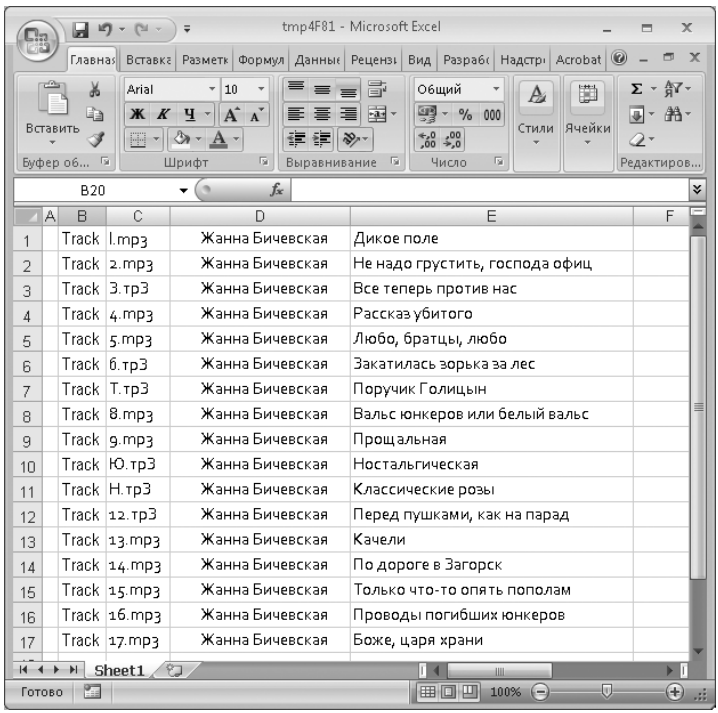


Рис. 10.13 ▾ Список файлов в таблице Excel

ментарии, нарисовать рамки и т. п. Получится обложка для компакт-диска, которую останется лишь распечатать.

Списки файлов и папок с описаниями и комментариями полезно заводить во многих случаях. Например, если в какой-то папке с дистрибутивами программ у вас хранится несколько десятков дополнений и плагинов, заготовок, текстур, моделей, сложно запомнить, «что есть что». По названиям догадаться можно, но не всегда. Гораздо удобнее составить небольшой текстовый файл, в котором, например, будет сказано: «phpsuk.exe – чертежи 6 простых кухонь, Польша 2005г.; KateTr-i.exe – кух. гарнитур Kate, модерн» и т. п.

Снимки интерфейса

При написании этой книги многие названия элементов управления в диалоговых окнах мы копировали как раз с помощью программы ABBYY Screenshot Reader. Точно так же программа отлично захватывает и распознает текст с кнопок, заголовков окон, сообщений и т. д. Этот прием помогает не только при составлении руководств, но и во многих других случаях.

Например, какая-то программа выдала сообщение, причем на английском языке. Еще забавнее, если это сообщение на немецком или итальянском! Естественное желание – понять, что сказано в сообщении. Можно отправить его текст по электронной почте или ICQ кому-нибудь из знакомых и попросить разъяснить смысл.

1. Не закрывая окно сообщения, запустите программу Screenshot Reader.
2. Выберите нужный язык. Как правило, основные европейские языки мы узнаем почти безошибочно, даже не зная их. Если не уверены, что это за язык, выберите вариант **Авто**. Закройте диалог настроек.
3. Выберите в раскрывающемся списке окна программы Screenshot Reader действие **Текст в буфер обмена**.
4. Нажмите кнопку **Снимок** и захватите текст, который вас интересует (рис. 10.14).

Через несколько секунд распознанный текст окажется в буфере обмена. Дальше с этим текстом можно поступать по-разному.

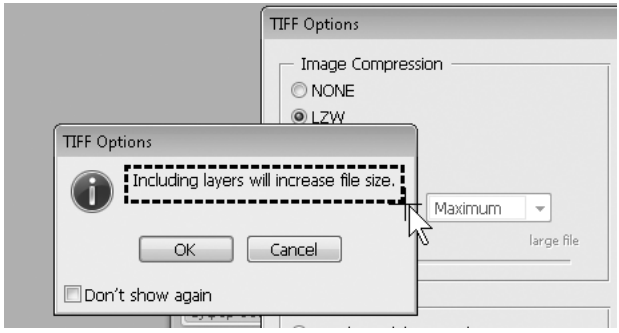


Рис. 10.14 ▾ Захват текста из окна сообщения

1. Подключитесь к Интернету и откройте в браузере веб-страницу какого-либо интерактивного переводчика, например **http://translate.google.com/translate_t#**. Вставьте текст в поле ввода, выберите направление перевода и нажмите кнопку **Перевести**. Через короткое время вы увидите перевод (рис. 10.15). Пусть он небезупречен, но смысл сообщения понять можно.

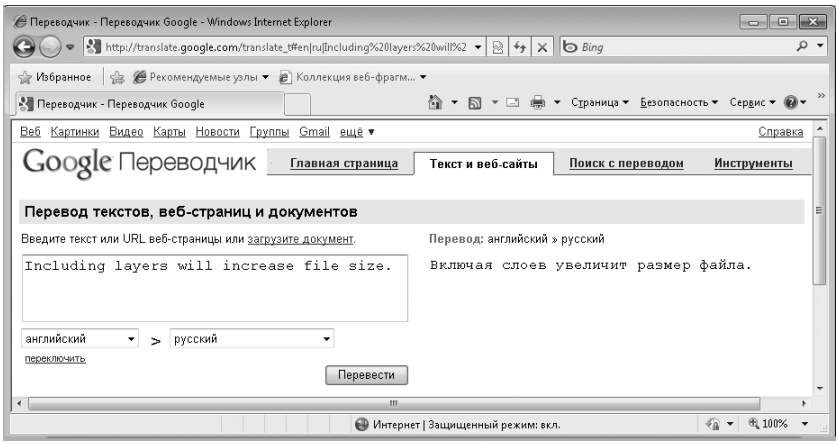


Рис. 10.15 ▾ Перевод на сайте Google

2. Если на компьютере установлены, например, словарь ABBYY Lingvo или программа-переводчик Prompt, вставьте текст в одну из таких программ. Вы также получите перевод, даже без подключения к Интернету.
3. Создайте сообщение электронной почты и вставьте в него текст из буфера обмена. Отправьте письмо тому, кто может подсказать вам решение проблемы.

Сама идея «распознать и перевести» простирается гораздо дальше. Так можно понять, что написано на упаковках, маркировке иностранных товаров или панелях импортной техники. Сфотографируйте надписи цифровой камерой, распознайте фотографию в программе FineReader, предварительно выбрав язык, а потом переведите текст любой программой-переводчиком или с помощью Интернета.

Резюме

ABBYY Screenshot Reader – приятное дополнение к программе FineReader. Фактически это еще один способ получения изображений – не с цифровой камеры или сканера, а непосредственно с экрана компьютера. Захваченное изображение или результат его распознавания (текст или таблицу) программа передает

в буфер обмена, приложение Microsoft Office либо сохраняет в файл. Программа Screenshot Reader способна распознать все изображение либо как текст, либо как таблицу. При выборе функции **Изображение в ABBYY FineReader** «снимок экрана» передается для дальнейшей обработки в программу FineReader. В этом случае внутри одного изображения можно выделить области разных типов и распознать их по отдельности.

Распознавание «снимков экрана» незаменимо в тех случаях, когда документ-источник сложно или невозможно открыть средствами FineReader. Если текст виден на экране, то он доступен и для захвата.

Книги издательства «ДМК Пресс» можно заказать в торгово-издательском холдинге «АЛЬЯНС-КНИГА» наложенным платежом, выслав открытку или письмо по почтовому адресу: **123242, Москва, а/я 20** или по электронному адресу: **orders@alians-kniga.ru**.

При оформлении заказа следует указать адрес (полностью), по которому должны быть высланы книги; фамилию, имя и отчество получателя. Желательно также указать свой телефон и электронный адрес.

Эти книги вы можете заказать и в Internet-магазине: **www.alians-kniga.ru**.

Оптовые закупки: тел. **(495) 258-91-94, 258-91-95**; электронный адрес **books@alians-kniga.ru**.

Жадаев Александр Геннадьевич

Сканирование и распознавание текстов

Самоучитель по работе с ABBYY® FineReader 10

Главный редактор *Мовчан Д. А.*
dm@dmk-press.ru
Корректор *Синяева Г. И.*
Верстка *Чаннова А. А.*
Дизайн обложки *Мовчан А. Г.*

Подписано в печать 04.03.2010. Формат 70×100 $^{1}/_{16}$.
Гарнитура «Петербург». Печать офсетная.
Усл. печ. л. 23,25. Тираж 1000 экз.
№

Web-сайт издательства: www.dmk-press.ru

САМОУЧИТЕЛЬ
по работе с программой
ABBYY® FineReader 10

Работать с «электронными» документами во многом удобнее и проще, чем с их «бумажными» аналогами. Электронный документ можно редактировать, использовать при создании собственных работ, его легко копировать и пересылать по электронной почте. Вместе с тем, многие материалы изначально доступны нам в нередактируемом виде (бумажные или отсканированные документы, цифровые фотографии). Программа ABBYY® FineReader – лучший инструмент для создания «электронных копий» любых печатных материалов: книг, справочников, журналов, договоров, бланков.

Книга включает описание приемов сканирования и распознавания разных оригиналов – от простых книжных страниц до сложнооформленных документов. А приведенные скриншоты программы позволят читателю быстро освоить интерфейс ABBYY® FineReader и получить практические навыки по работе с программой.

Изложение материала сопровождается практическими примерами. Читатели, которые еще не пробовали самостоятельно переводить печатные материалы в электронный вид, найдут в этой книге простое пошаговое руководство. Для тех же, кто хочет в совершенстве освоить работу с программой, книга откроет многочисленные тонкости настройки для эффективного использования ABBYY® FineReader.

*DVD содержит ознакомительные версии программ компании ABBYY®.

ДМК ПРЕСС
ИЗДАТЕЛЬСТВО

ABBYY® PRESS

Internet-магазин: www.alians-kniga.ru

Книга – почтой:
Россия, 123242, Москва, а/я 20
e-mail: book@alians-kniga.ru

Оптовая продажа: «Альянс-книга»
Тел./факс: (495) 258-9195
e-mail: books@alians-kniga.ru

978-5-94074-595-2

9 785940 745952