Тема 3.2 Системи перекладу та розпізнавання текстово-графічної інформації

Програма PROMT

Цю програму (її більш ранні версії відомо під назвою Stylus) розроблено російською фірмою PROMT. Програма є додатком до операційних систем, таких як Windows 95, 98, NT 4.0, 2000 і може бути інтегрована в комплект програм Microsoft Office, зокрема, у програми Microsoft Word та Excel.

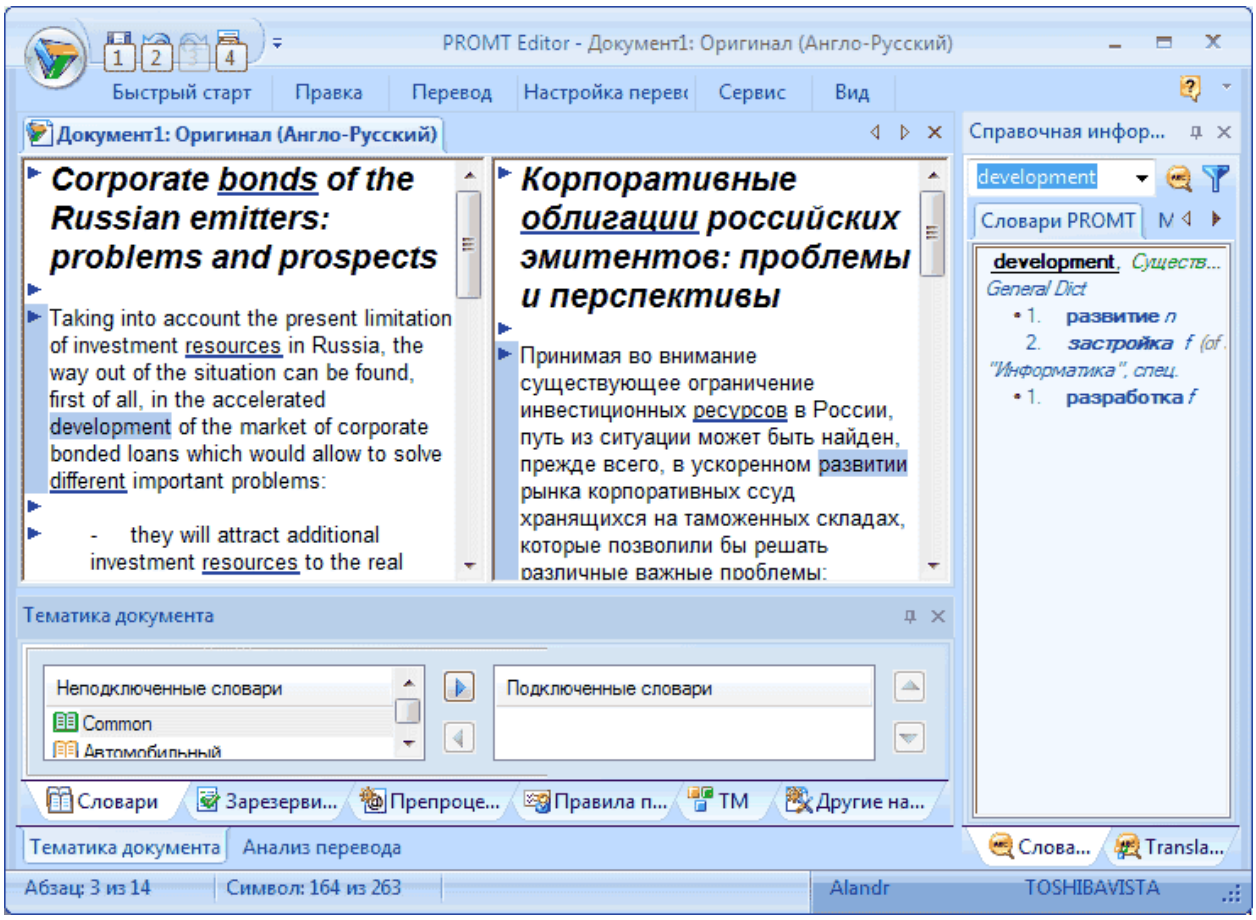
Можливості програми PROMT:

- забезпечення перекладу документів з англійської, німецької та французької мов на російську і навпаки;
- до неї можна підключати кілька десятків спеціалізованих словників, що забезпечує правильний переклад термінів, які стосуються певної області знань;
- динамічне відслідковування напрямку перекладу, тобто визначення мови оригіналу і перекладу;
- переклад вмісту буфера обміну, поточного параграфа, виділеного фрагмента тексту або всього тексту;
- забезпечення будь-якого з можливих напрямків перекладу, підключення й відключення словників, доповнення та виправлення їх, складання списку зарезервованих слів, які не перекладаються;
- робота безпосередньо з програмами розпізнавання текстів, наприклад, FineReader;
- не виходячи з програми можна використати відомі способи редагування й форматування оригіналу та перекладу;
- забезпечення перевірки орфографії оригіналу і перекладу після встановлення прикладних програм для перевірки правопису (LingvoCorrector, Пропис, Орфо, Hugo).

Головне вікно програми складається з трьох частин: дві призначені для відображення оригіналу тексту і його перекладу, третя - утворює інформаційну панель, де відображаються інформація про перекладений документ і спеціальні настройки. Вікно має стандартні елементи керування вікна Windows - заголовок, рядок меню, панелі інструментів і т.д.

Для швидкого запуску всіх програм, що входять до складу PROMT, призначений Інтегратор PROMT у вигляді окремої панелі робочого стола Windows. Кнопки панелі Інтегратора, а також пункти контекстного меню, яке викликається клацанням правою клавішею миші на значку Інтегратора, що є на панелі задач, дають змогу вибрати такі функції програми PROMT :

1. Переклад Clipboard (вміст буфера обміну).
2. Відкрити файл.
3. Відкрити WWW-вузол.
4. Пошук у WWW.
5. Запустити програму PROMT.
6. Запустити File Translator - програму перекладу файлів у пакетному режимі.
7. Запустити WebView - броузер-перекладач, що забезпечує синхронний переклад Web-сторінок при навігації у Internet.
8. Запустити Quick Translator - програму швидкого перекладу тексту, набраного з клавіатури.



Переклад окремих слів і виділених фрагментів можна здійснити прямо в тексті, навівши на них вказівник миші. Окремі фрагменти тексту можна перекласти без попереднього запуску програми PROMT. Для цього досить, знаходячись в будь-якому текстовому редакторі, наприклад, Notepad або Microsoft Word, скопіювати виділений фрагмент у буфер обміну і викликати функцію Переклад Clipboard Інтегратора PROMT.

Переклад документа за допомогою програми PROMT передбачає проведення кількох етапів:

1. **Введення документа,** який необхідно перекласти. Документ може бути завантажений з файлу. Для цього слід виконати стандартну операцію відкриття файлу. Текст для перекладу може також бути набраний на клавіатурі у власному редакторі програми. Для цього треба спочатку створити новий документ за допомогою відповідної команди. Для перекладу введеного з клавіатури тексту без виклику основного вікна програми PROMT можна також скористатися функцією Quick Translator Інтегратора PROMT. У вікні програми Quick Translator, крім перекладу, можна виконати також інші дії з оригіналом і перекладеним текстом: скопіювати переклад у буфер обміну, змінити напрямок перекладу, підключити додаткові словники тощо.

2. **Уточнення параметрів перекладу**. Після того, як підготовлено оригінал тексту, що підлягає перекладу, слід визначається напрямок перекладу, тобто з якої мови на яку мову буде здійснюватися переклад, а також уточнити формат тексту оригіналу (формат файлу тексту оригінала, наприклад MS Word файл, форматований текст RTF і т.д.).

3. **Підготовка тексту до перекладу**. Вибраний документ відображається в області тексту оригіналу. Перед початком перекладу доцільно перевірити орфографію, оскільки неправильно написані слова будуть сприйматися програмою як невідомі і залишаться без перекладу. У разі необхідності текст можна зберегти для подальшої роботи як документ

PROMT. У документі можуть бути слова і словосполучення, які не повинні перекладатися, наприклад, прізвища, назви програмних продуктів (Windows 98, Microsoft Word 2000 тощо). Іноді застосовують транслітерацію - запис із використанням іншого алфавіту, що відповідає написанню або вимові мовою оригіналу (наприклад, прізвище Brown бажано перекласти не як Коричневий, а Браун).

Інколи доводиться відмовлятися від перекладу цілих абзаців, наприклад, текстів програм на алгоритмічних мовах. Щоб відмовитися від перекладу окремих слів, їх треба зарезервувати, тобто встановити на цьому слові курсор, а потім клацнути мишею на відповідній кнопці панелі інструментів або вибрати пункт Зарезервировать... у контекстному меню чи меню Перевод.

Можна зарезервувати фрагмент тексту, заздалегідь виділивши його або цілий абзац. У тексті всі зарезервовані слова й абзаци, що мають залишитися без перекладу, виділяють зеленим кольором. Якість перекладу визначається повнотою словників, які використовуються, з урахуванням граматичних правил. Для кожного документа можна задати набір словників, які переглядаються у певному порядку до першого виявлення слова для перекладу. Програмою PROMT для перекладу передбачено три типи словників:

- **генеральний словник** (містить загальновживану лексику і побутове значення слів). Він використовується завжди, причому останнім з усіх словників. Зміна цього словника неможлива;
- **спеціалізовані словники** (містять терміни з різних областей). Редагувати ці словники не можна, але їх можна підключати й відключати під час перекладу. Базове постачання програми не містить додаткових словників і їх необхідно встановлювати окремо;
- **словник користувача** (створюється користувачем) До нього додаються слова, яких немає в інших словниках, а також уточнені переклади тих або інших слів Як правило, цей словник переглядають насамперед. Словник користувача можна редагувати.

Список словників, що використовуються під час перекладу, відображається у вікні інформаційної панелі. Підключення словників здійснюється за допомогою відповідної команди програми PROMPT.

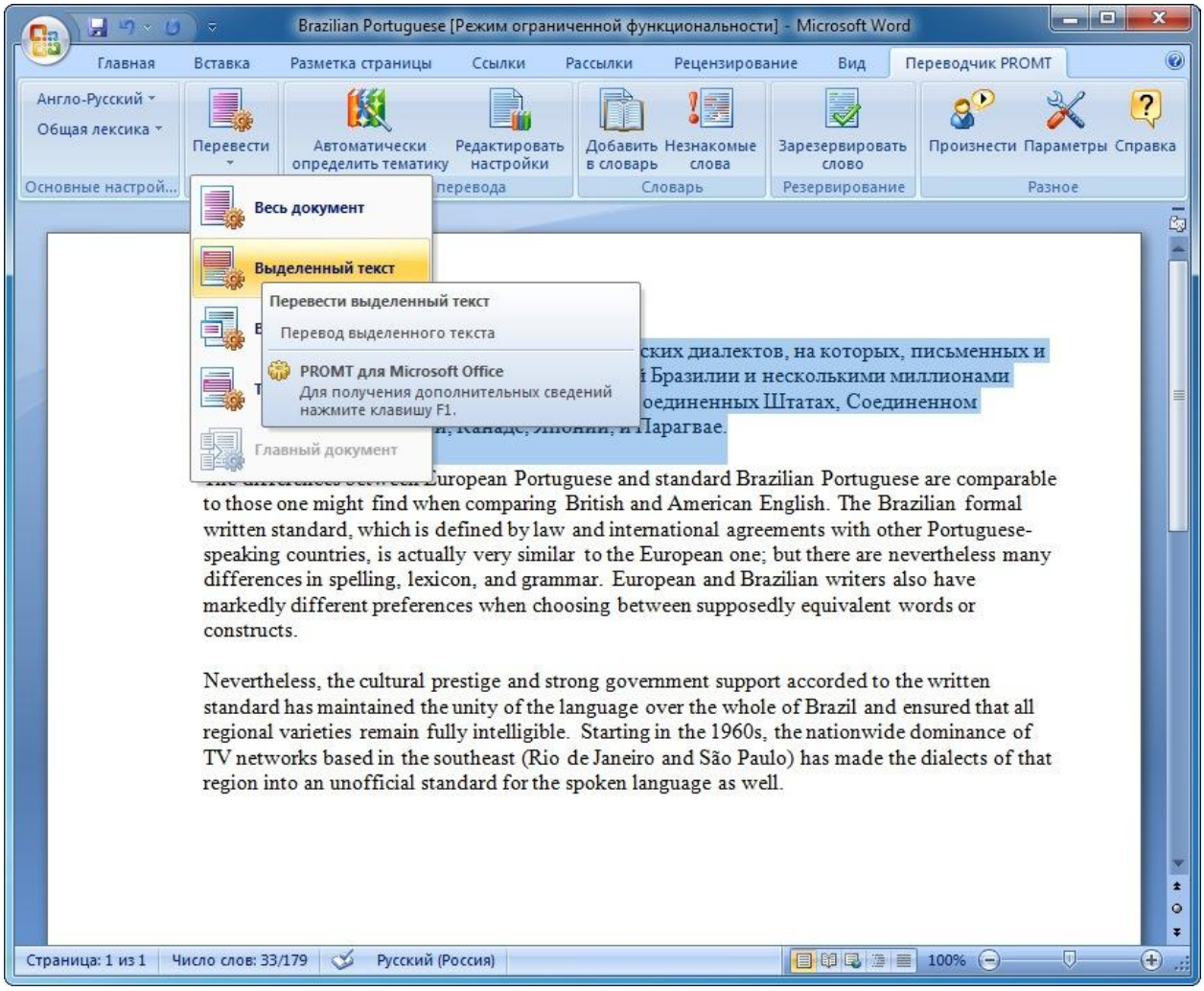
4. **Переклад документа**. Переклад документа починається після вибору користувачем відповідної команди з меню Перевод. Перекладений документ заноситься в область перекладу. Невідомі слова виділяються червоним кольором, а зарезервовані - зеленим. Список невідомих і зарезервованих слів відображається на інформаційній панелі у відповідних вкладках. У разі необхідності невідомі слова можна занести в словник користувача. Початковий текст і переклад можна редагувати, форматувати та перекладати повторно.
5. **Збереження результатів**. Після завершення робіт із текстами, оригінал і переклад можна зберегти в одному з форматів, що підтримуються програмою, використовуючи стандартні команди збереження файлу.

Програма PROMT забезпечує ряд додаткових можливостей, які розглянемо окремо.

1. Сумісна робота з програмою розпізнавання текстів. Якщо до комп'ютера підключено сканер і встановлено програму оптичного розпізнавання текстів, наприклад, FineReader, то її можна запустити безпосередньо з програми перекладу PROMT. Використовуючи сканер, програма FineReader забезпечить перетворення

надрукованого на папері тексту на електронну форму і передасть його до програми для перекладу й редагування.

2. Сумісна робота з пакетом Microsoft Office. Програму перекладу PROMT можна інтегрувати з Word і Microsoft Excel. Це дає змогу перекладати відкриті в цих додатках документи, не виходячи з програм.
3. Переклад Web-сторінок. До складу програми PROMT входить додаткова програма WebView, яка забезпечує підключення користувача до Web-вузлів, пошук інформації у мережі Internet й автоматичний переклад Web-сторінок з англійської, німецької, французької мов на російську і навпаки. Запустити цю програму можна з панелі Інтегратора PROMT .



Вікно програми Promt Professional 9.5

Розпізнавання графічної та текстової інформації за допомогою програми FineReader

Автоматичне розпізнання тексту

Після обробки документа сканером виходить графічне зображення документа. Але графічний вигляд не являється текстом документа. Людині досить подивитись на листок паперу з текстом, щоб зрозуміти, що на ньому написано. З точки зору комп'ютера, документ після сканування перетворюється в набір різнокольорових точок, а не в текстовий документ. Проблема розпізнання тексту в складі точкового графічного зображення являється дуже складною. Подібні задачі вирішуються за допомогою спеціальних програмних засобів, називаються вони засоби розпізнання зображень.

Реальний технічний прорив в цій області пройшов лише в останні роки. До того розпізнання тексту було можливо лише шляхом порівняння знайдених конфігурацій точок із стандартним зразком. Автори програми критерій “схожості” використовуваний при ідентифікації символів. Такі системи називаються OCR(OptikalCharacted Recognition-оптичне розпізнання символів) і оперались на спеціально вироблені шрифти. З часом наукові дослідження в області розпізнання зображень буквально перевернули представлення при оптичному розпізнані символів. Сучасні програми можуть ставитись з різноманітними шрифтами без перенастройки. Багато розпізнають навіть малюнковий.

Програми розпізнання текстів

Оскільки потреба в розпізнані тексту відсканованих документів достатньо велика, невипадково, що є велика кількість програм, призначена для такої цілі. Так, як різні наукові методи розпізнання тексту розвивалась незалежно один від одного, багато із цих програм використовують різні алгоритми. Ці алгоритми можуть давати різні результати на різні документи. Наприклад, система OCR здібна розпізнати тільки стандартний спеціально підготовлений шрифт і дають на цьому шрифті найкращі результати, які не можуть перевершити ні одна із універсальних програм.

Сучасні алгоритми розпізнання тексту не орієнтуються на конкретний шрифт, ні на конкретний алфавіт. Більшість програм розпізнають текст на декількох мовах. Один і той же алгоритм можна використовувати для розпізнання російського, латинського, арабського і других алфавітів і навіть змішаних текстів. Розуміється програма повинна знати про який алфавіт іде мова. Нас перш за все інтересують програми здатні розпізнавати текст, написаний на російській мові. Такі програми випускаються вітчизняними виробниками. Найбільш широко відомі і розповсюджені програми Fine Reader і Cunei Form. Програма Fine Reader забезпечує високу якість розпізнання і вигоду застосування.

Розпізнання документів в програмі Fine Reader

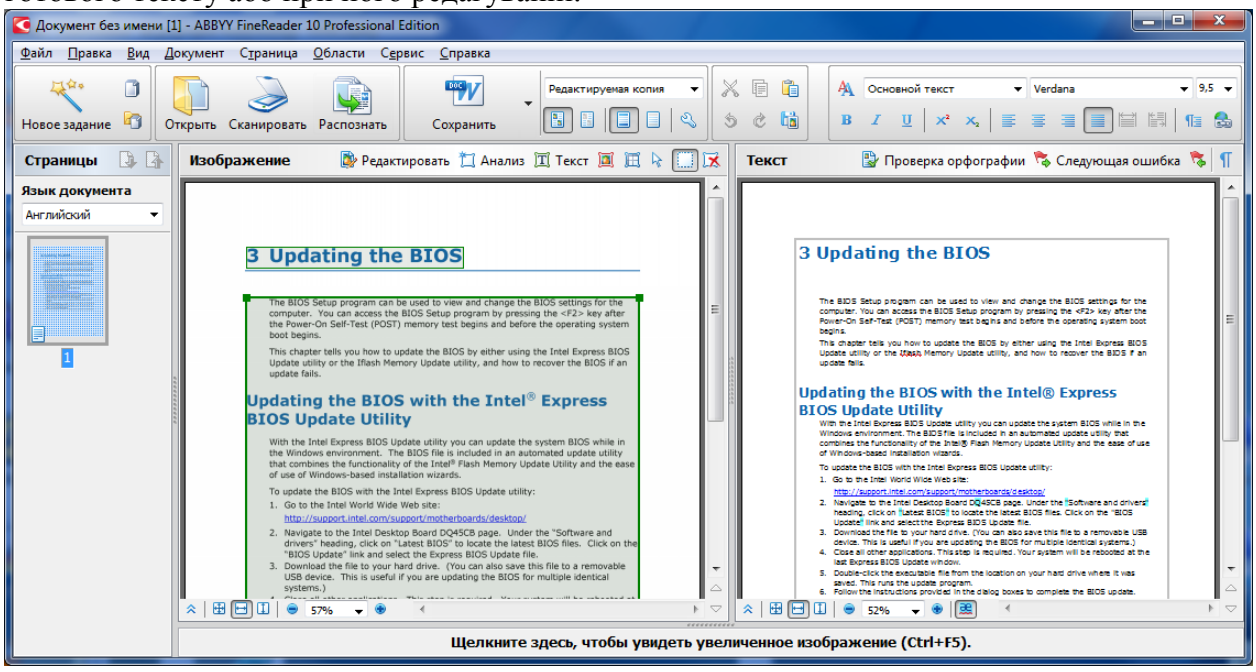
Програма Fine Reader виготовляється вітчизняною компанією АВВУУ Software(w.w.w. bitsoft.ru.). Ця програма призначена для розпізнання текстів на російському, англійському, німецькому, українському, французькому і багатьох інших мовах, а також для розпізнання змішаних двох мовних текстів. Програма має ряд можливостей. Вона дозволяє об’єднати сканування і розпізнання в одну операцію, працювати з пакетами документів і бланками. Програму можна навчити для кращої якості розпізнання неправильно надрукованих текстів і складних шрифтів. Вона дозволяє редагувати текст і провіряти його орфографію. Fine Reader працює з різними моделями сканерів. Програма дотримується стандарту TWAIN. Ми розглянемо програму на прикладі версії 4.0 одну із основних версій.

Вікно програми

Після включення програми Fine Reader в меню програми головного меню появляються пункти, забезпечуючи роботу з нею. Вікно програми має типовий для Windows 9x вид і має стрічку меню, ряд панелей інструментів і робочу область.

1. В лівій частині робочої області розміщується панель Пакет, містить список графічних документів які повинні бути перетворені в текст. Ці графічні файли розглядаються, як частинки одного документа. Результати її обробки в подальшому об’єднуються в єдиний текстовий файл. Форма значка, відмічає початковий файл і вказує чи було проведено розпізнання.
2. Панель в нижній частині робочої області має фрагмент графічного документа в збільшиному виді. З його допомогою можна оцінити якість розпізнання. Цю панель також використовують для “навчання” програми в ході розпізнання тексту.
3. А всю іншу частину робочої області займають вікна документів. Тут розміщується вікно графічного документа, а також вікно текстового документа після розпізнання.
4. У верхній частині вікна під стрічкою меню розміщується панель інструментів.

5. Панель інструментів Стандартна містить кнопки для відкриття документа і для операції з буфером обміну. Інші кнопки цієї панелі служать для зміни представлення документа.
6. Панель Scan Read містить кнопки, які відповідають всім етапам перетворення паперового документа в електронний текст. Перша кнопка дозволяє виконати таке перетворення в рамках єдиної операції. Остальні кнопки відповідають відокремленим етапам роботи і містять відкриваючи меню службові для управління відповідною операцією.
7. Панель Розпізнання дозволяє вказати мову документа і вид шрифта. Остані вимагаються роботи тільки в тих випадках, коли документ має не достатню кількість друку.
8. Панель Інструменти використовується при роботі з вихідними зображеннями. Вона дозволяє управляти сегментацією документів. З допомогою елементів управління цієї панелі задають послідовність фрагментів текстів в заключному документі.
9. Елементи управління панелі Формативна використовується для зміни представленні готового тексту або при його редагуванні.



Порядок розпізнання текстових документів

Перетворення паперового документа в електронний проходить в три етапи. Кожний із цих етапів програми Fine Readerможе виконувати, як автоматичний так і під контролем користувача. Якщо всі етапи проходять автоматично, то перетворення документа проходить за один прийом.
1. Перший етап роботи – сканування. На цьому етапі завжди використовується сканер. Однак зображення з листка паперу може бути перетворена в цифрову форму і з допомогою других засобів таких, як наприклад цифрові фотоапарати і цифрові відеокамери.
2. Другий етап роботи – сегментація тексту. Діло в тому, що в паперових документах, на сторінках книжки чи журналу, текст не завжди розміщується в зазначеному порядку. Він може розміщуватись в декількох колонках. Містить малюнки (підписи до них). Доповнюючі вирізки і дані представлені в таблиці, а також можуть заплутати порядок тексту. Тому перш за все, як включити текст документа його розбивають на блоки, вміст фрагментів. Блоки розпізнають послідовно. Отриманий текст включається в документів порядку номера блока.
3. Останній етап роботи програми-розпізнання. Цей етап не потребує втручання користувача, за винятком тих випадків, коли розпізнання супроводжується ”навчанням”.

Розпізнання тексту відображається у окремому вікні у виді форматованого тексту. Він “втрачає зв’язок” з вихідним зображенням і може редагуватися і форматуватися незалежно від нього. програма виділяє кольором ті символи, які вона сама розглядає, як неоднозначно розпізнання. Це спрощує пошук помилок засобами програми в отримані тексту, можна також провести перевірку граматики.

4. Отриманий текст можна зберегти у виді форматованого документа. Передбачено також можливість прямої передачі отриманого тексту в програму Word чи Excel, а також в буфер обміну Windows.

Сканування документа

Сканування – це технічна операція , яку виконує пристрій для сканування. Задача програми Fine Readerна цьому етапі складається з того, щоб приняти получену інформацію і прийняти значки від сканування сторінок на панелі Пакет. Так сторінки готуються до розпізнання.

1. Для того щоб провести сканування за допомогою програми Fine Reader, необхідно запустити цю програму і включити сканер. Просканові сторінки проходять по клацані на кнопки Сканувати на панель інструментів або при написанні комбінації клавіш Ctrl+K.

2. Програма здатна працювати із сканером, як безпосередньо так і через протокол. При безпосередній взаємодії із сканером можливість сканування кольорових зображень не використовується так, як текст являється в любому випадку одноколірним.

3. Програма використовується для сканування, яке задано по зменшенню. Для того щоб вибрати таке обладнання чи змінити його настройку, потрібно клацнути на відкриваючій кнопці поруч з кнопкою Сканувати і вибрати у відкритому меню пункт Опції – відкривається діалогове вікно Опції.

4. Якщо до комп’ютера підключено декілька сканерів або сканер був підключений після установки програми, слід клацнути на кнопці Вибрати сканер. В тому випадку програма проведе пошук підключених до комп’ютера сканерів і дозволити вибрати потрібний.

5. Для зміни настройки сканера використовують кнопку Настройки сканера.

6. Коли сканер вибраний появляється два флажка, в нижній частині діалогового вікна. Якщо поставити флажок Показувати діалог TWAIN-драйвера сканера, то сканування проходить через протокол з відображенням діалогового вікна. В протележному випадку програма працює напряму із сканером. Використати протокол має значення тільки в тому випадку, коли робота напряму неможлива або дає неякісні результати.

7. Флажок Показати опції перед початком сканування застосовують лише в тому випадку, коли паперові сторінки документа сильно відрізняється одна від другої. Це може бути викликано, наприклад, розмірами паперу або тим, що різні сторінки друкувались в різний час і різними засобами. В такому випадку перед скануванням кожної сторінки відкривається діалогове вікно настройки сканера, щоби користувач міг відрегулювати якість процесу.

Сам процес сканування проходить в автоматичному режимі. Якщо потрібно проробити багато сторінок, то краще всього спочатку їх усіх просканувати, а уже потім проводити розпізнання. Це зв’язано з тим, що сканування потребує присутності користувача через управління сканером, а розпізнання може проводитися в автоматичному режимі.

Сегментація документа

Під порядком розпізнання тексту розуміється послідовне розпізнання сторінок зліва на право. Якщо текст розбитий на декілька колонок або має вирізки, підмалюнкові підписи, примітки і другі елементи форматування, його розпізнання в послідовному порядку неможливо. В таких випадках програма розбиває текст на блоки, кожний з яких представляє собою фрагмент тексту, розпізнаний в послідовному порядку. Таке розбиття документа називається сегментацією.

Автоматична сегментація – це проста задача для програми. Програма шукає проміжки між стрічками, а також зони початку і кінця стрічок. Якщо послідовність

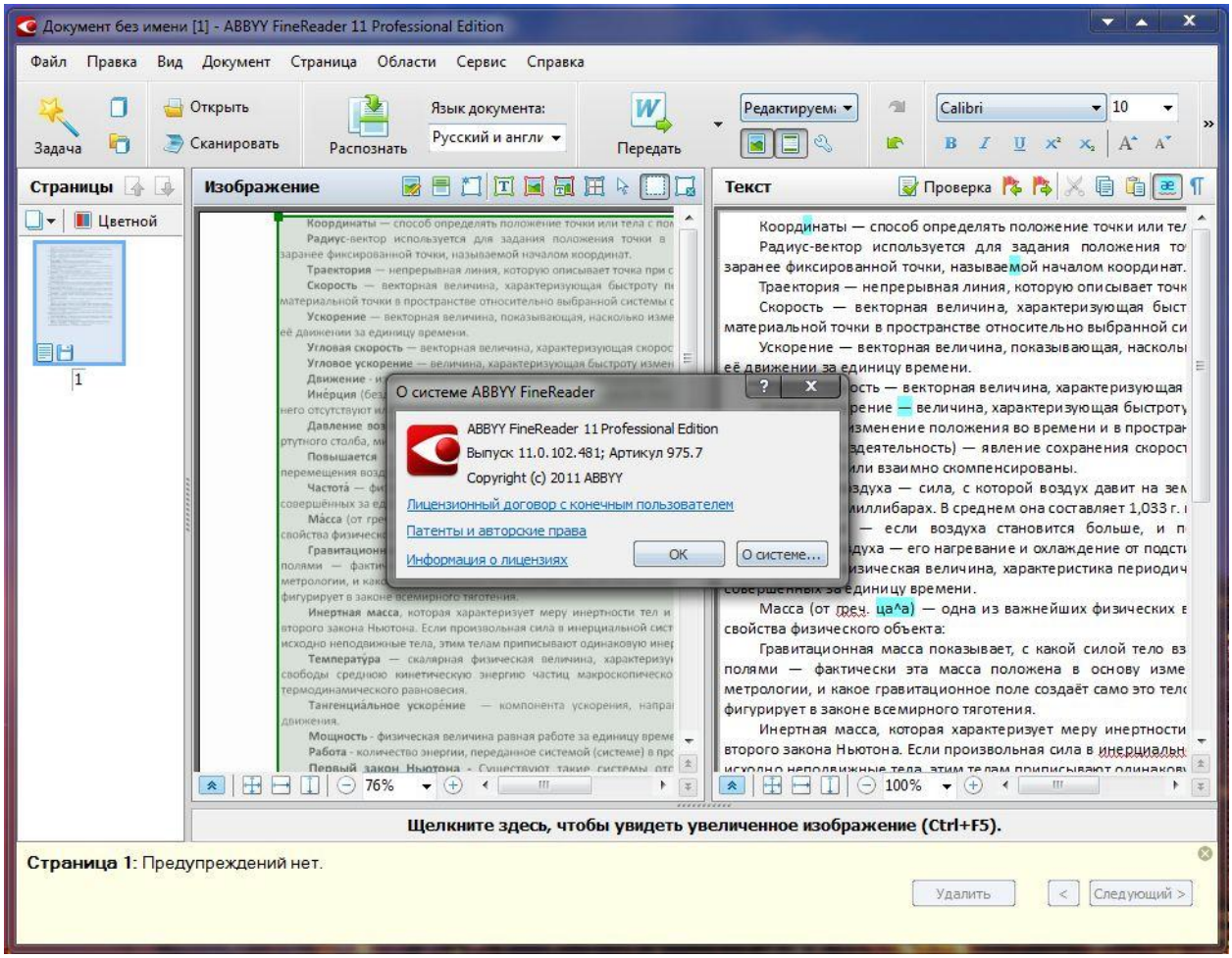
стрічок, ідуть підряд, має одинакові зони початку і кінця, то програма розглядає таку область як текстовий блок.

Якщо проміжки між стрічками взагалі існують, то по всій можливості, мова іде про ілюстрацію. Якщо знайдеться велике число вертикальних і горизонтальних фрагментів, які відображають правильну структуру, то напевне в текст включена таблиця. Якщо клацнути на кнопці

Сегментація виділені сторінки, то сегментація сторінки проходить автоматично. правда зображення документа має невисоку якість, то сегментація може бути виконано не правильно, що виявляється у великій кількості малих блоків. В таких випадках можна вручну вказати границі блоків або змінити автоматичне розбиття. Нові прямокутні блоки створюються методом перетягуванням миші. При перетягуванні створюється блок і виділяється пунктирною рамкою, яка в момент створення блока перетворюється в єдину зелену лінію. Якщо сторінка має простий стандартний вид, то простіше вручну створити одиничний блок, який охоплює всю сторінку, чим представити це програмі, ризикуючи можливістю появлення помилок. Один із створених блоків являється поточним. Він обведений великою жирною лінією, а його вершини помічені маркерами. Переніс цих маркерів можна редагувати границі блока.

Для створення блока прямокутної форми або зміна послідовності блоків використовують кнопки панелі інструментів Інструменти. Всі кнопки цієї панелі використовують на етапі сегментації.

Програма Fine Reader розпізнає декілька блоків які працюють по різному. Такі блоки виділяються різними кольорами. Текстові блоки обводяться зеленою лінією. Щоб змінити тип блока потрібно клацнути в зоні блока ПКМ і вибрати потрібний тип в меню. Тип блока контекстного меню.



Розпізнання документа

Після сегментації і встановлення порядку текстових блоків виконує останній етап роботи – розпізнання. Якщо документ надрукований не стандартним шрифтом, який добре відсканований, по клацанню кнопки Розпознать відкриту сторінку досить, щоб документ був розпізнаний. Якщо паперовий документ має нестандартний шрифт, то процес розпізнання ускладнюється. В такому випадку програма може не справитись з розпізнанням символів і допускати однотипні помилки. В таких випадках для великих документів перш за все спочатку треба провести навчання програми з особливостями даного документа. Це досить великий процес, але він все таки простіший, ніж ручний ввід багато сторінкового документа.

Настройку розпізнання починають із створення еталону в, якому зберігаються особливості даного документа. Для цього потрібно виконати команду Сервіс – Редактор еталонів, клацнути у відкритому діалоговому вікні Еталони на кнопці Нові еталони і ввести ім'я створення еталона.

1. Для підключення еталона при розпізнані, треба клацнути на відкриваючі кнопці поруч з кнопкою Розпізнати відкриту сторінку і вибрати пункт Опції. У відкритому діалоговому вікні в групі Обучение слід вибрати тільки, що створений еталон. Якщо розпізнання документа відповідає еталону, який був створений і настроєний раніше, то вибрати не новий, а старий еталон.
2. Для "навчання" еталона слід встановити прапорець Розпізнання з навчанням.
3. Режим розпізнання в такому випадку змінюється. Коли програма не може розпізнати символ, то вона видає діалогове вікно "навчання еталона". У верхній частині цього вікна проводиться збільшення зображень розпізнання стрічки. Текучий символ обведений рамкою.
4. В полі із списком Символ, який розуміє програма, знаходиться в рамці.
5. Необхідно переконатись, що символ в полі вказаний правильно і замінити його у випадку необхідності. Після цього треба клацнути на кнопці "навчання".
6. Якщо неправильно вказані границі символу, то кнопки Зсунути вліво і Зсунути вправо, дозволять поправити положення рамки.
7. Якщо правильно розмістити рамку не вдається або в тексті зустрічається незнайомий символ, який правильно перекласти не можливо, слід клацнути на кнопці "Пропустити".

Особливості настройки програми Fine Reader

Як і більшість других додатків Windows, програму Fine Reader можна настроїти відповідно з потребою конкретного користувача. Всі настройки виконуються за допомогою діалогового вікна Опції, які відкриваються з допомогою любої відкриваючої стрілки на панелі інструментів, чи через меню Сервіс. Якщо використана панель інструментів, то діалогове вікно відкривається на вкладці, відповідно до використаної кнопки панелі інструментів.

Вкладка Сканування служить для вибору і настройки сканера, а також для визначення способу доступу до нього.

1. Вкладка Сегментація дозволяє настроювати деякі параметри для автоматичної сегментації . тут задаються параметри автоматичного розбиття таблиці і настроюють режим автоматичної сегментації багато стовпчикового тексту.
2. Засоби вкладки Форматування дозволяє задавати спосіб форматування розпізнаної сторінки і вибрати потрібний шрифт.
3. Вкладка Розпізнавати визначає параметри розпізнання документа. Вона дозволяє задати мову документа і особливості початкового шрифту, а також настроїти режим розпізнання з навчанням. Тут також задається метод кольорового виділення не надійно розпізнаного символу.
4. Елементами управління вкладки Перевірка задають метод перевірки орфографії і спосіб виділення знайдених помилок чи незрозумілих місць.

5. Вкладка Установки розпізнає всі настройки програми. Тут задають мову інтерфейсу і настроюють використані одиниці вимірювання.
6. Прапорці панелей Показувати розпізнають спосіб представлення вікна програми і відкритих документів.
7. панель Кольори дозволяє вибрати колір різноманітних елементів документа. В нижній частині вікна можна задати доповнюючи параметри.