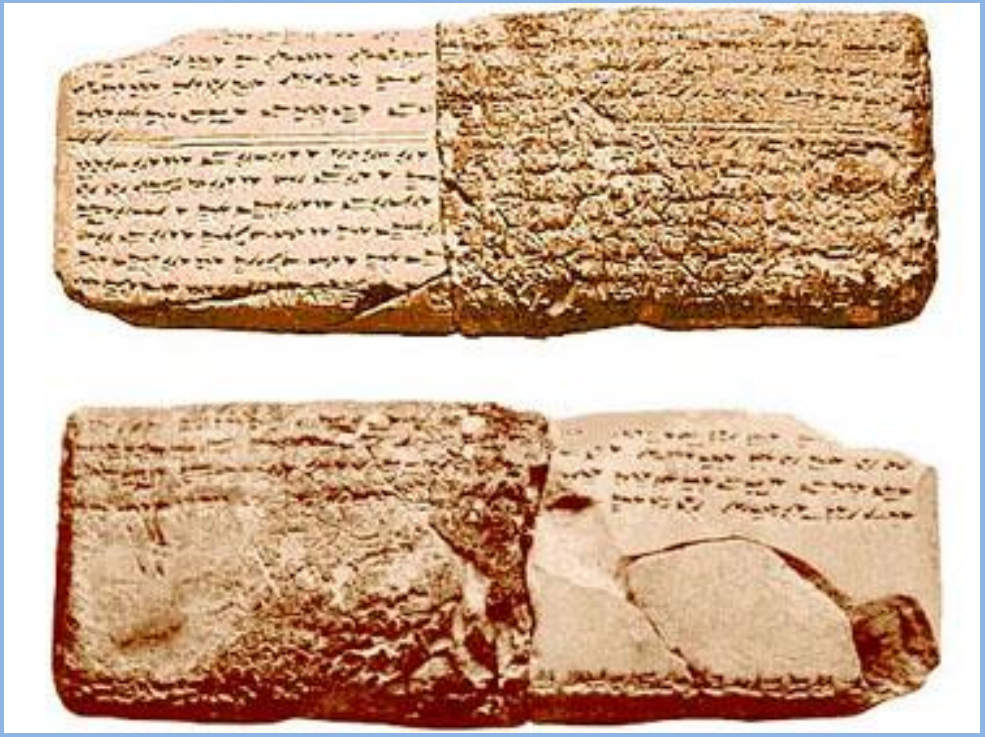
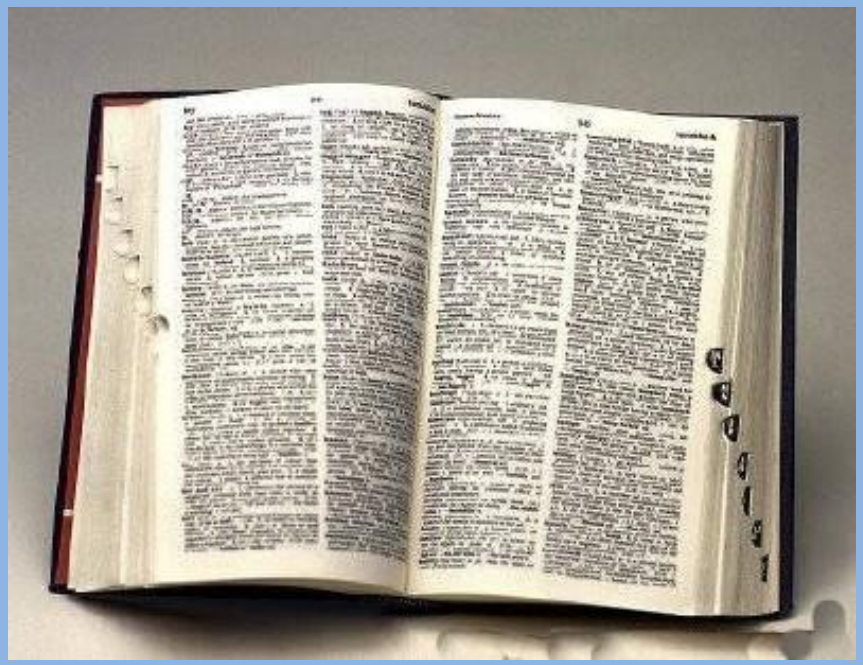
Системы перевода и распознавания текста

Компьютерные словари

Словари необходимы для перевода текстов с одного языка на другой. Первые словари были созданы около 5 тысяч лет назад в Шумере и представляли собой глиняные таблички, разделенные на две части.

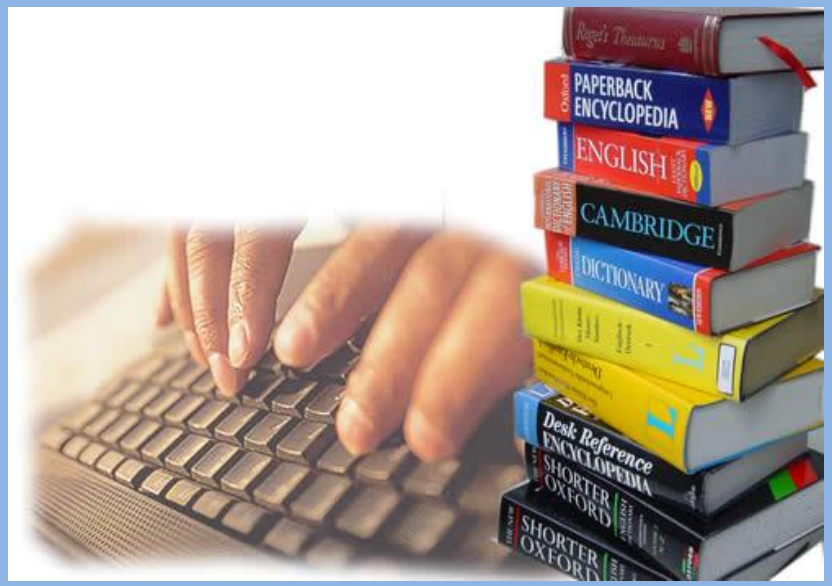


В настоящее время существуют тысячи словарей для перевода между сотнями языков (англо-русский, немецко-французский и так далее), причем каждый из них может содержать десятки тысяч слов. В бумажном варианте словарь представляет собой толстую книгу объемом в сотни страниц, где поиск нужного слова является достаточно трудоемким процессом.



Компьютерные словари предоставляют пользователю дополнительные возможности:

- выбор языков и направлений перевода;
- содержание десятков специализированных словарей по областям знаний (техника, медицина, информатика и др.);
- обеспечение быстрого поиска словарных статей
- прослушивание слов в исполнении дикторов, носителей языка.

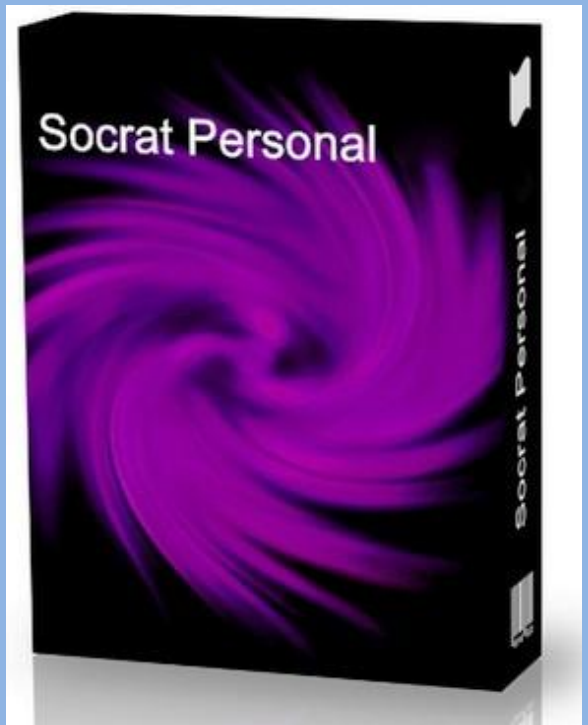


Системы машинного перевода

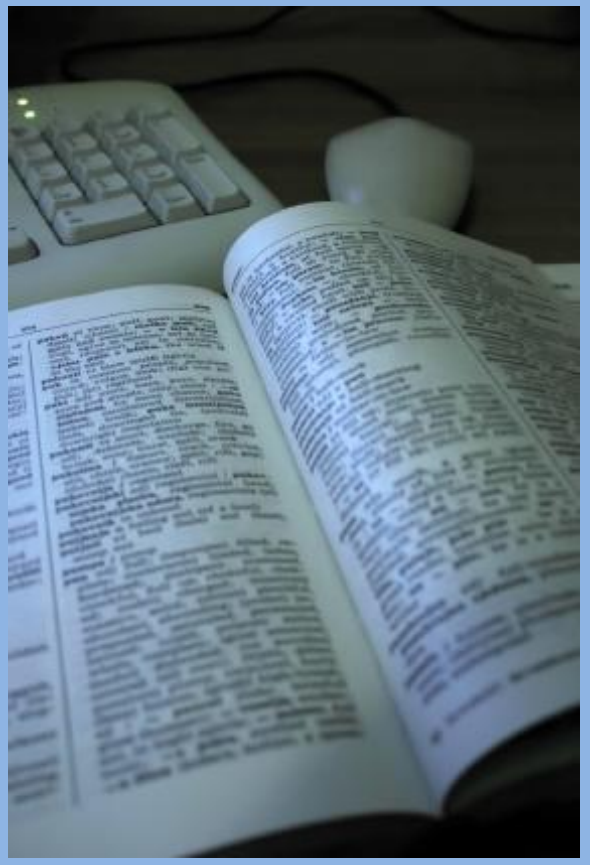
Происходящая в настоящее время глобализация нашего мира приводит к необходимости обмена документами между людьми и организациями, находящимися в разных странах мира и говорящими на различных языках.

В этих условиях использование традиционной технологии перевода «вручную» тормозит развитие межнациональных контактов. Перевод многостраничной документации вручную требует длительного времени и высокой оплаты труда переводчиков. Перевод полученного по электронной почте письма или просматриваемой в браузере Web-страницы необходимо осуществить немедленно, и нет возможности и времени пригласить переводчика.

Системы машинного перевода позволяют решить эти проблемы. Они, с одной стороны, способны переводить многостраничные документы с высокой скоростью (одна страница в секунду) и, с другой стороны, переводить Web-страницы «на лету», в режиме реального времени. Лучшими среди российских систем машинного перевода считаются PROMT и «Сократ».

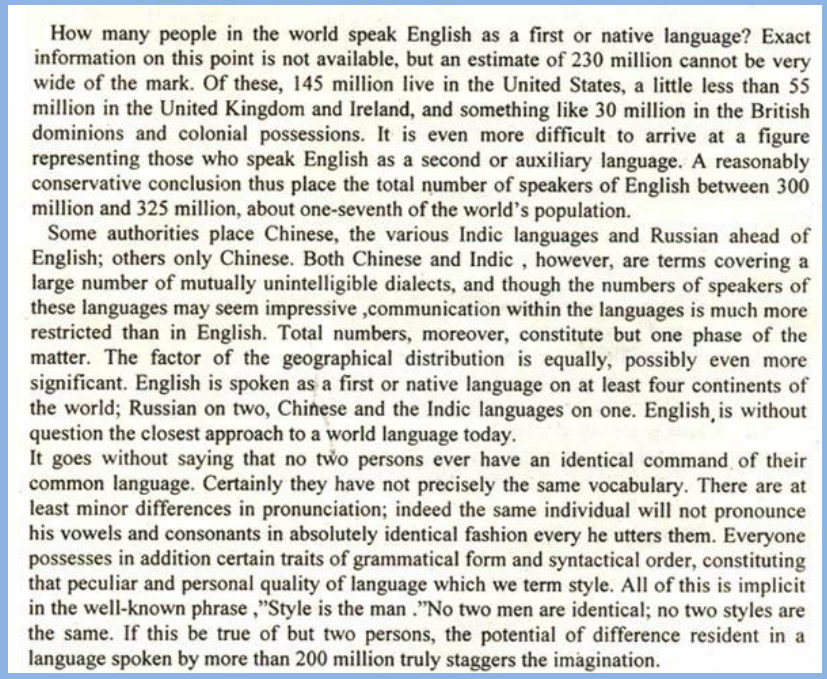


Современные системы машинного перевода позволяют достаточно качественно переводить техническую документацию, деловую переписку и другие специализированные тексты. Однако они неприменимы для перевода художественных произведений, так как не способны адекватно переводить метафоры, аллегории и другие элементы художественного творчества человека.



Системы распознавания текста

С помощью сканера достаточно просто получить изображение страницы текста в графическом файле. Однако работать с таким текстом невозможно: как любое сканированное изображение, страница с текстом представляет собой графический файл - обычную картинку.

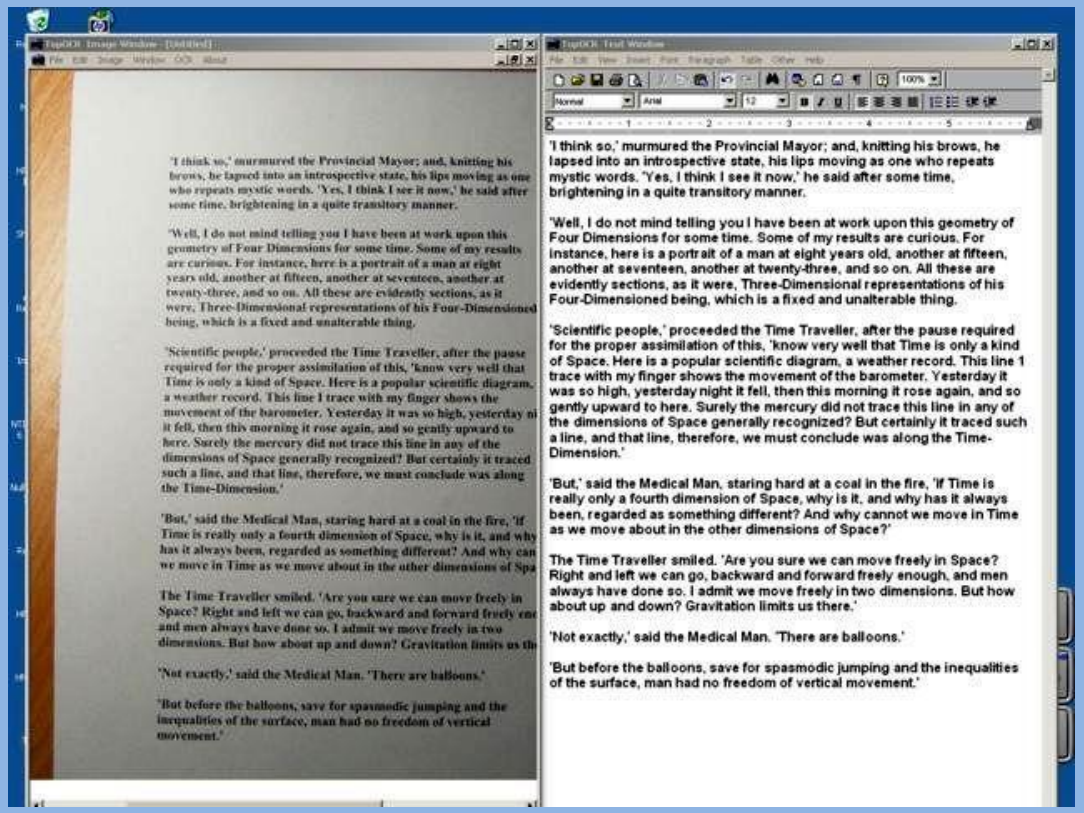


How many people in the world speak English as a first or native language? Exact information on this point is not available, but an estimate of 230 million cannot be very wide of the mark. Of these, 145 million live in the United States, a little less than 55 million in the United Kingdom and Ireland, and something like 30 million in the British dominions and colonial possessions. It is even more difficult to arrive at a figure representing those who speak English as a second or auxiliary language. A reasonably conservative conclusion thus place the total number of speakers of English between 300 million and 325 million, about one-seventh of the world's population.

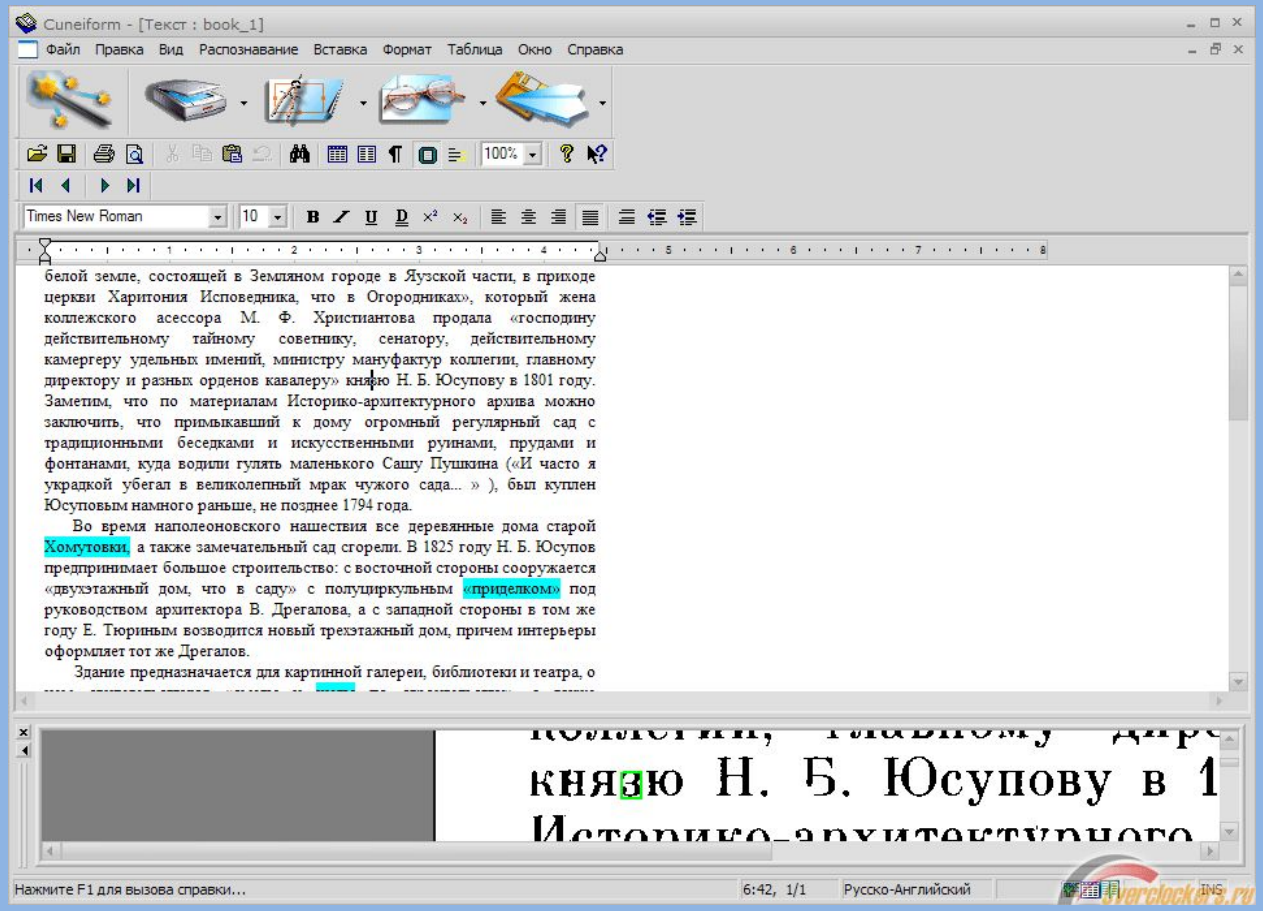
Some authorities place Chinese, the various Indic languages and Russian ahead of English; others only Chinese. Both Chinese and Indic , however, are terms covering a large number of mutually unintelligible dialects, and though the numbers of speakers of these languages may seem impressive ,communication within the languages is much more restricted than in English. Total numbers, moreover, constitute but one phase of the matter. The factor of the geographical distribution is equally, possibly even more significant. English is spoken as a first or native language on at least four continents of the world; Russian on two, Chinese and the Indic languages on one. English, is without question the closest approach to a world language today.

It goes without saying that no two persons ever have an identical command of their common language. Certainly they have not precisely the same vocabulary. There are at least minor differences in pronunciation; indeed the same individual will not pronounce his vowels and consonants in absolutely identical fashion every he utters them. Everyone possesses in addition certain traits of grammatical form and syntactical order, constituting that peculiar and personal quality of language which we term style. All of this is implicit in the well-known phrase ,"Style is the man ."No two men are identical; no two styles are the same. If this be true of but two persons, the potential of difference resident in a language spoken by more than 200 million truly staggers the imagination.

Текст можно будет читать и распечатывать, но нельзя будет его редактировать и форматировать. Для получения документа в формате текстового файла необходимо провести распознавание текста, то есть преобразовать элементы графического изображения в последовательности текстовых символов.



Преобразованием графического изображения в текст занимаются специальные программы распознавания текста (Optical Character Recognition - OCR).



Современная OCR должна уметь:

- распознавать тексты, набранные не только определенными шрифтами, но и рукописные;
- корректно работать с текстами, содержащими слова на нескольких языках, распознавать таблицы;
- корректно распознавать не только четко набранные тексты, но и такие, качество которых, очень плохое; (Например, текст с пожелтевшей газетной вырезки или третьей машинописной копии)
- сохранение результата в файле популярного текстового (или табличного) формата (например, формат Microsoft Word).

Наиболее распространенные системы оптического распознавания символов: FineReader, CuneiForm, используют как растровый, так и структурный методы распознавания. Кроме того, эти системы являются «самообучающимися» (для каждого конкретного документа они создают соответствующий набор шаблонов символов) и поэтому скорость и качество распознавания многостраничного документа постепенно возрастают.


Cuneiform
Файл Вид Справка
Мастер распознавания: Изображение
Выберите источник получения изображения
Открытое изображение
Открыть с диска
C:\Documents and Settings\centurion\Рабо
Изменить...
Отсканировать
Драйвер не установлен
Изменить...
< Назад
Далее >
Отмена
Справка


Интеллектуальное распознавание текстов
Непревзойденная точность преобразования документов и PDF-файлов
ABBYY
Максимальная точность
format retention
OCR and PDF conversion
за минимальное время
ABBYY
FineReader OCR 9.0
Professional Edition