Readiris™ PRO 10

ASIAN OCR ADD-ON FOR READIRIS

I.R.I.S.

Document to Knowledge™

OCR technology by I.R.I.S.
Connectionist and Linguistic technology by I.R.I.S.

© 2005 I.R.I.S. All rights reserved

Introducing the Asian OCR Add-on

Congratulations on acquiring the Asian OCR add-on!

By installing it, Readiris offers the recognition of four extra **Asian languages**: Japanese, Simplified and Traditional Chinese and Korean. (It goes without saying that a large number of Asian languages such as Malay, Tagalog etc. are supported by the "standard" Readiris software because they use the Latin alphabet.)

This extra CD-ROM "Asian OCR Add-On" complements your installed Readiris license - the "add-on" only works on version 10 and later of the Readiris software. This technical note explains all you need to know to successfully recognize Asian documents.

A Word about These Asian Languages

Contrary to the Latin, Greek and Cyrillic alphabets, these languages use small icons ("**ideograms**"), not letters, to represent reality.

Japanese

The Japanese writing system actually uses a mix of alphabets: the phonetic character sets are called Hiragana and Katakana, and the ideograms, Chinese in origin, are called Kanji. In normal Japanese writing, Hiragana and Kanji are used, while Katakana is used for words borrowed from the (non-Chinese) foreign languages.

すべての人間は、生まれながらにして自由であり、かつ、尊厳と権利とについて平等である。人間は、理性と良心とを授けられており、互いに同胞の精神をもって行動しなければならない。

An educated person can read about 10,000 Kanji symbols; the government has published a list of some 2,000 symbols that it considers basic.

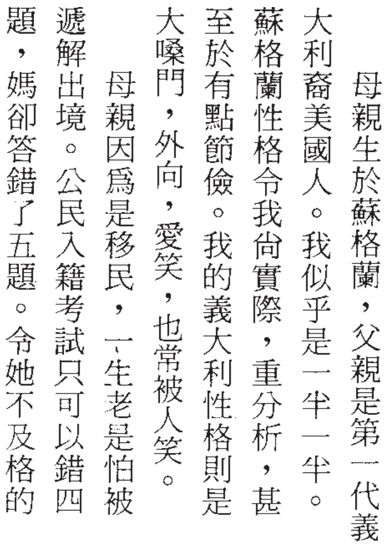
Japanese is generally written vertically beginning on the right, but many texts today are written horizontally to permit the inclusion of English words, Arabic numerals and mathematical and chemical formulae.

Simplified and Traditional Chinese

As a script, Chinese is derived from picture writing. It is written with thousands of distinctive characters called “ideograms” or “pictograms” which have no relation to the sound of a word. The earliest Chinese characters were pictographs, such as a crescent for the moon, or a circle with a dot in the center to represent the sun. Gradually, these gave way to non-pictorial ideographs which, in addition to standing for tangible objects, also represented abstract concepts.

The majority of Chinese characters consist of two elements: a “signific”, which indicates the meaning of a word, and a “phonetic”, which indicates the sound.

In a large dictionary there are 40,000 to 50,000 characters (many of which are archaic or obscure), while the telegraphic code book contains nearly 10,000 symbols. Some 3,000 symbols are used on a daily basis.



母親生於蘇格蘭，父親是第一代義大利裔美國人。我似乎是一半一半。蘇格蘭性格令我尚實際，重分析，甚至於有點節儉。我的義大利性格則是大嗓門，外向，愛笑，也常被人笑。

母親因爲是移民，一生老是怕被遞解出境。公民入籍考試只可以錯四題，媽卻答錯了五題。令她不及格的

Simplified Chinese is a simplified version of the “traditional” Chinese; the 500 most common symbols were simplified. Simplified Chinese is used on China’s mainland and in Singapore, **Traditional Chinese** is used by Hong Kong, Taiwan, Macau and the overseas Chinese communities.

貝見門馬體豐龍龜

贝见门马体丰龙龟

Every character has exactly the same amount of space, no matter what its shape may be. There are no spaces between characters; the characters which make up multi-syllable words are not grouped together. When reading Chinese, you have to work out which characters belong together!

Chinese can be written vertically and from right to left or horizontally from left to right.

Korean

Korean is not related to Chinese, although it has used the Chinese characters, together with the Korean alphabet, for many centuries. The Korean alphabet, the "Hangul" script, invented in the years 1443-46, is the only true alphabet native to the Far East.

모든 개인과 사회 각 기관이 이 선언을 항상 유념하면서 학습 및 교육을 통하여 이러한
권리와 자유에 대한 존중을 증진하기 위하여 노력하며 , 국내적 그리고 국제적인 점진적
조치를 통하여 회원국 국민들 자신과 그 관할 영토의 국민들 사이에서 이러한 권리와
자유가 보편적이고 효과적으로 인식되고 준수되도록 노력하도록 하기 위하여 , 모든
사람과 국가가 성취하여야 할 공통의 기준으로서 이 세계인권선언을 선포한다.

System Requirements

The Asian OCR "add-on" does not modify the system requirements of Readiris.

It complements version 10 and later of the Readiris software - it does not work on *earlier* versions.

This extra module takes about 50 MB of hard disk space.

However, you will need a localized Asian version of the Windows **operating system** to make good use of the Asian texts. Alternatively, you can use Word 2003, Word 2002 and Word 2000 to view and edit such documents: Microsoft Office 2003 System, Office XP and Office 2000 were specifically designed to cope with documents in many different languages.

Use an Asian version of the Adobe Reader software to view and edit **Asian PDF documents**. Know above all that the Asian versions of Adobe Reader can be found on the Readiris CD-ROM!

Installing the Asian OCR "Add-on"

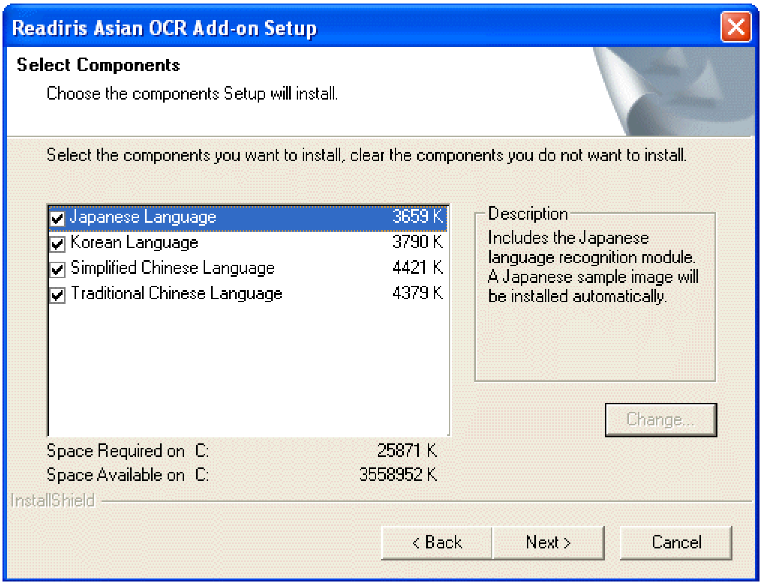
As is the "basic" software Readiris, the Asian OCR "add-on" software is delivered exclusively on an **autorunning CD-ROM**. To install, simply insert the

CD-ROM in your CD-ROM drive and wait for the installation program to start running.

Should the installation not begin to run when the CD-ROM is inserted in your CD-ROM drive, run the setup program SETUP.EXE to install the software.

Users of Windows XP, Windows 2000 and Windows NT 4.0 must ensure that they have the necessary **access rights -** contact the system administrator if necessary.

Follow the on-screen instructions. The installer detects automatically where your Readiris software is installed **-** the software will only install if Readiris 10 is present on your PC **-**, but some installation **options** are offered: you can limit the installation to specific languages. A sample image is installed automatically for the selected languages.



Readiris Asian OCR Add-on Setup

Select Components

Choose the components Setup will install.

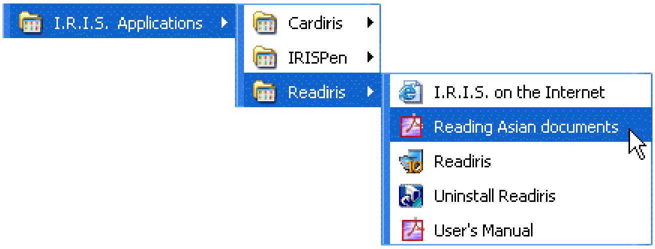
Select the components you want to install, clear the components you do not want to install.

☑ Japanese Language	3659 K
☑ Korean Language	3790 K
☑ Simplified Chinese Language	4421 K
☑ Traditional Chinese Language	4379 K

Description
Includes the Japanese language recognition module. A Japanese sample image will be installed automatically.

Space Required on C: 25871 K
Space Available on C: 3558952 K

As this "add-on" software smoothly complements your installed Readiris software, no new submenu or application is added by the installation program. But you will find a shortcut to this electronic document added to the submenu "I.R.I.S. Applications - Readiris".



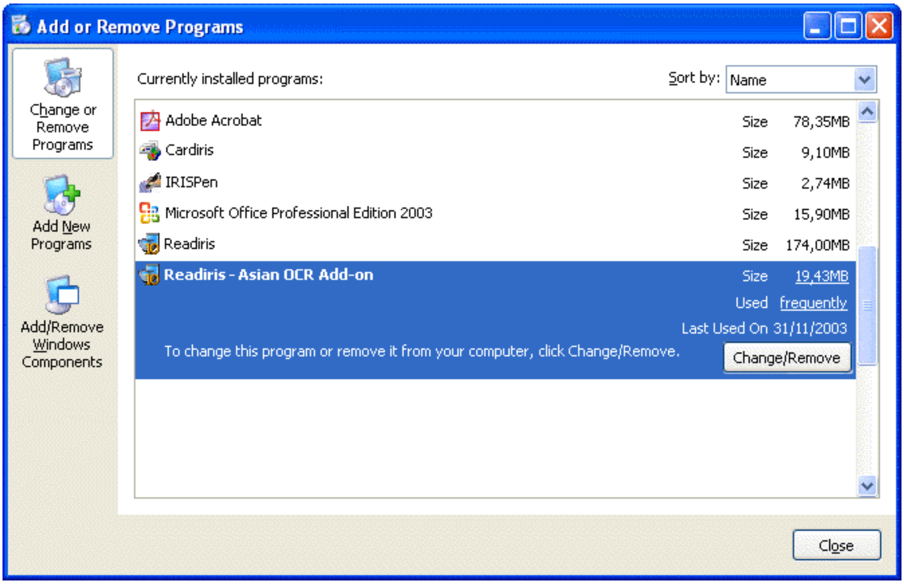
Uninstalling the Asian OCR "Add-on"

There's only one correct way of uninstalling Readiris - by using the Windows (un)install wizard. You are strongly recommended *not* to uninstall this Readiris software module by manually erasing the program files.

Execute the following steps to make use of the Windows (un)install wizard.

❑ Click "Settings" under the "Start" menu of Windows and go to the "Control Panel".

❑ Click the icon "Add/Remove Programs" under the control panel.



Add or Remove Programs
Change or Remove Programs
Add New Programs
Add/Remove Windows Components
Currently installed programs:
Sort by: Name
Change/Remove
Close

Adobe Acrobat	Size	78,35MB
Cardiris	Size	9,10MB
IRISPen	Size	2,74MB
Microsoft Office Professional Edition 2003	Size	15,90MB
Readiris	Size	174,00MB
Readiris - Asian OCR Add-on	Size	19,43MB
	Used	frequently
	Last Used On	31/11/2003

To change this program or remove it from your computer, click Change/Remove.

❑ Follow the on-screen instructions to remove the Asian OCR "add-on" module.

CONFIGURING AN "ASIAN" WORKING ENVIRONMENT

So much for the installation of the Asian OCR software. But we must also ensure that your computer system handles the ideograms of these Asian languages correctly.

It is not necessary to install a localized Asian version of the **Windows** operating system to make good use of such recognized texts. You can also use **Word 2003**, **Word 2002** and **Word 2000** to view and edit such documents: Microsoft Office 2003 System, Office XP and 2000 were specifically designed to cope with documents in many different languages.

Refer to the documentation supplied with your Windows or Office software to learn how to set up and use your Asian-enabled environment.

Use an Asian version of the Adobe Reader or Adobe Acrobat software to view and edit **Asian PDF documents**. (The Asian versions of Adobe Reader can be found on the Readiris CD-ROM!)

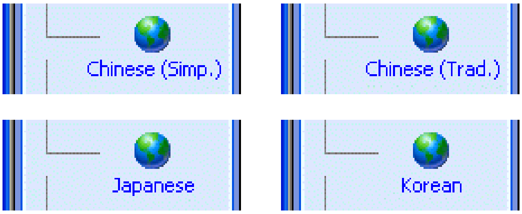
Taking these steps ensures that your computer system copes with the symbols ("ideograms") of these Asian languages. If your operating system is not "Asian-enabled", you will inevitably generate illegible output whenever you try to display text in one of these languages. This phenomenon is not caused by the Readiris software, but by the setup of Windows and Adobe Acrobat: Readiris does recognize the ideograms of these Asian languages, but when you open the text file with your wordprocessor or with Adobe Reader, your computer system does not *represent* them correctly on your computer screen.

Recognizing Asian Documents

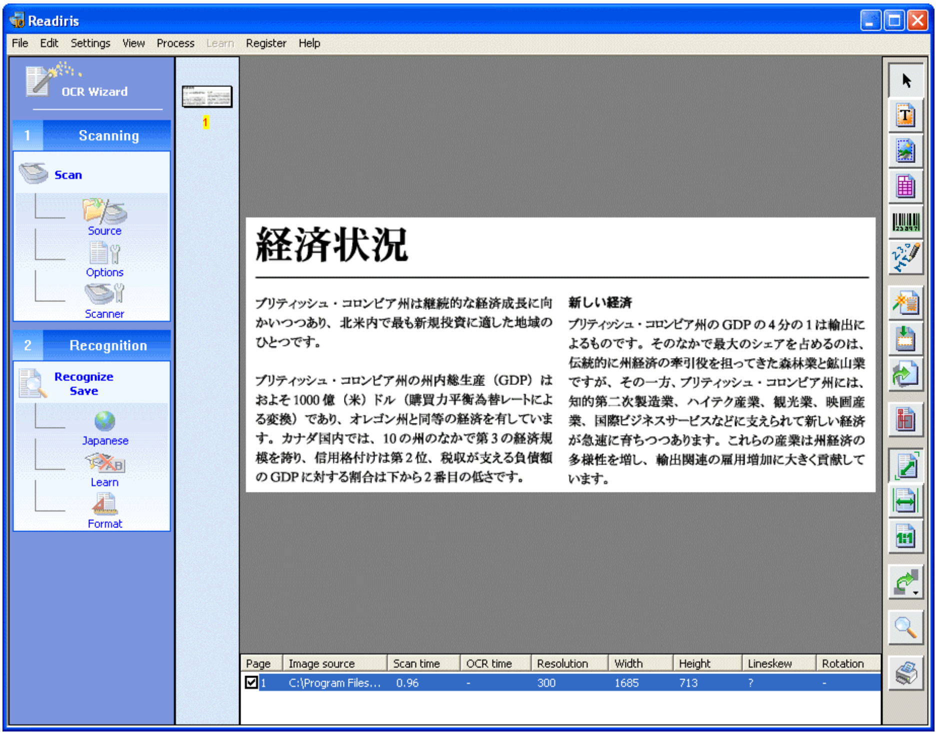
Assuming that your environment is set up correctly, we will now turn to the recognition of these languages.

When you start up your Readiris software, there are but few new elements in the user interface. We'll study these in detail.

Evidently, you can now select the languages Japanese, Simplified Chinese, Traditional Chinese and Korean with the "Language" button on the main toolbar.



The language setting influences the **page analysis**. In other words, indicate the language *before* you execute the page analysis! (Should you have forgotten this, select the appropriate language afterwards and the page analysis is re-executed automatically...)



経済状況

ブリティッシュ・コロンビア州は継続的な経済成長に向かいつつあり、北米内で最も新規投資に適した地域のひとつです。

ブリティッシュ・コロンビア州の州内総生産（GDP）はおよそ1000億（米）ドル（購買力平衡為替レートによる変換）であり、オレゴン州と同等の経済を有しています。カナダ国内では、10の州のなかで第3の経済規模を誇り、信用格付けは第2位、税収が支える負債額のGDPに対する割合は下から2番目の低さです。

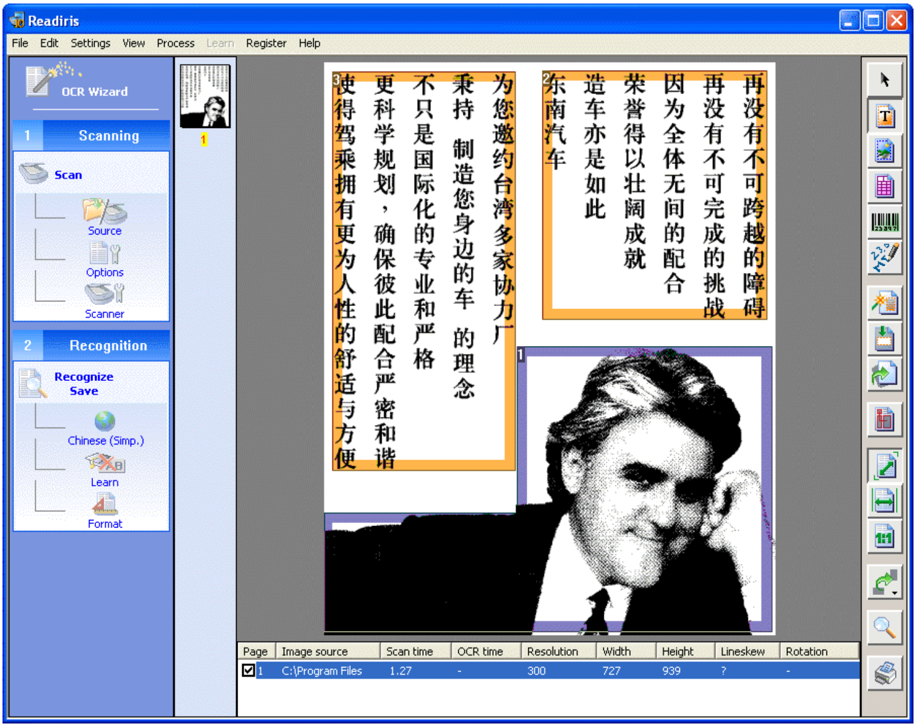
新しい経済

ブリティッシュ・コロンビア州のGDPの4分の1は輸出によるものです。そのなかで最大のシェアを占めるのは、伝統的に州経済の牽引役を担ってきた森林業と鉱山業ですが、その一方、ブリティッシュ・コロンビア州には、知的第二次製造業、ハイテク産業、観光業、映画産業、国際ビジネスサービスなどに支えられて新しい経済が急速に育ちつつあります。これらの産業は州経済の多様性を増し、輸出関連の雇用増加に大きく貢献しています。

Page	Image source	Scan time	OCR time	Resolution	Width	Height	Lineskew	Rotation
☑ 1	C:\Program Files...	0.96	-	300	1685	713	?	-

Specialized analysis routines are used for these languages. The interline spacing is in most cases bigger than with Western texts, and the text is less "dense": the words are made up of small icons ("ideograms") that could be seen as graphic zones in Western documents. (The maximal character size of 72 points also holds for the Asian languages.)

Thirdly, the text orientation may be different: the text may run from top to bottom, from right to left. Readiris adapts itself by sorting the text blocks from right to left!



再没有不可跨越的障碍
再没有不可完成的挑战
因为全体无间的配合
荣誉得以壮阔成就
造车亦是如此
东南汽车

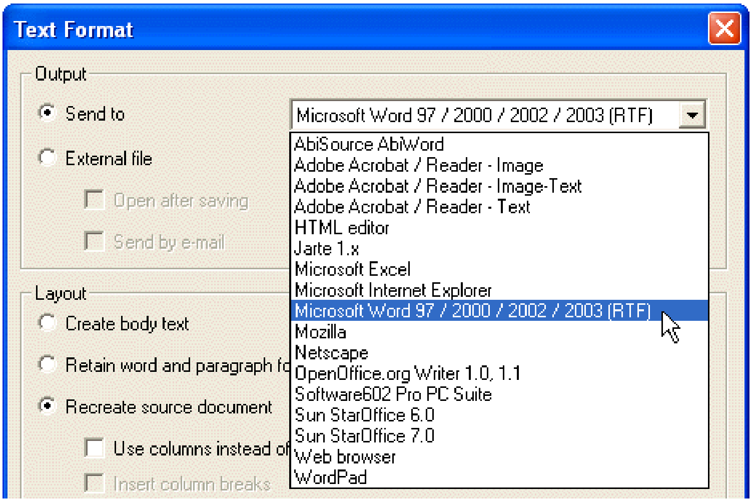
为您邀约台湾多家协力厂
秉持 制造您身边的车 的理念
不只是国际化的专业和严格
更科学规划，确保彼此配合严密和谐
使得驾乘拥有更为人性的舒适与方便

Page	Image source	Scan time	OCR time	Resolution	Width	Height	Lineskew	Rotation
☑1	C:\Program Files	1.27	-	300	727	939	?	-

Proceed as usual by clicking the button "Recognize-Save" on the main toolbar. Don't try to set the document characteristics such as font type or character pitch - these don't apply to Asian documents anyway.

Indicate the **formatting** options before you execute the OCR. Similar to Western documents, you can apply "autoformatting" and recognize tables! However, the number of text filters adapted to the Asian alphabets is somewhat smaller than the text formats available for Western documents.

Generally speaking, the formatting options shouldn't bother you: if an unavailable option is currently enabled and you enable an Asian language, Readiris prompts you to select an available option first.


Text Format

Output

● Send to: Microsoft Word 97 / 2000 / 2002 / 2003 (RTF)

○ External file

□ Open after saving

□ Send by e-mail

Layout

○ Create body text

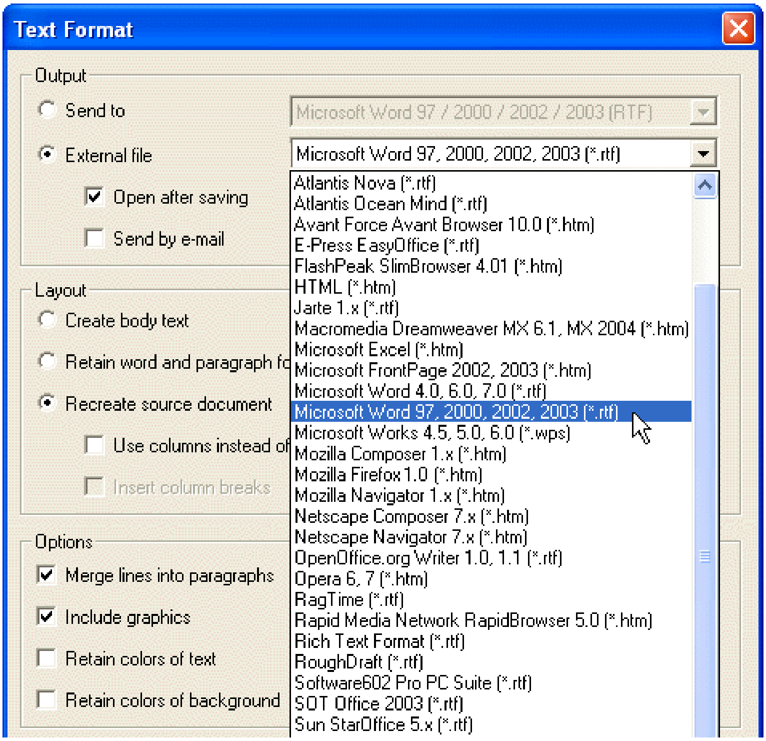
○ Retain word and paragraph fo

● Recreate source document

□ Use columns instead of

□ Insert column breaks

AbiSource AbiWord
Adobe Acrobat / Reader - Image
Adobe Acrobat / Reader - Image-Text
Adobe Acrobat / Reader - Text
HTML editor
Jarte 1.x
Microsoft Excel
Microsoft Internet Explorer
Microsoft Word 97 / 2000 / 2002 / 2003 (RTF)
Mozilla
Netscape
OpenOffice.org Writer 1.0, 1.1
Software602 Pro PC Suite
Sun StarOffice 6.0
Sun StarOffice 7.0
Web browser
WordPad



Text Format

Output

○ Send to — Microsoft Word 97 / 2000 / 2002 / 2003 (RTF)

● External file — Microsoft Word 97, 2000, 2002, 2003 (*.rtf)

☑ Open after saving

☐ Send by e-mail

Layout

○ Create body text

○ Retain word and paragraph fo

● Recreate source document

☐ Use columns instead of

☐ Insert column breaks

Options

☑ Merge lines into paragraphs

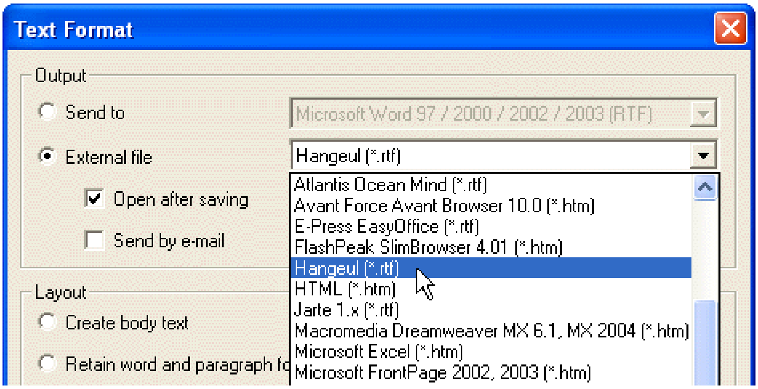
☑ Include graphics

☐ Retain colors of text

☐ Retain colors of background

Atlantis Nova (*.rtf)
Atlantis Ocean Mind (*.rtf)
Avant Force Avant Browser 10.0 (*.htm)
E-Press EasyOffice (*.rtf)
FlashPeak SlimBrowser 4.01 (*.htm)
HTML (*.htm)
Jarte 1.x (*.rtf)
Macromedia Dreamweaver MX 6.1, MX 2004 (*.htm)
Microsoft Excel (*.htm)
Microsoft FrontPage 2002, 2003 (*.htm)
Microsoft Word 4.0, 6.0, 7.0 (*.rtf)
Microsoft Word 97, 2000, 2002, 2003 (*.rtf)
Microsoft Works 4.5, 5.0, 6.0 (*.wps)
Mozilla Composer 1.x (*.htm)
Mozilla Firefox 1.0 (*.htm)
Mozilla Navigator 1.x (*.htm)
Netscape Composer 7.x (*.htm)
Netscape Navigator 7.x (*.htm)
OpenOffice.org Writer 1.0, 1.1 (*.rtf)
Opera 6, 7 (*.htm)
RagTime (*.rtf)
Rapid Media Network RapidBrowser 5.0 (*.htm)
Rich Text Format (*.rtf)
RoughDraft (*.rtf)
Software602 Pro PC Suite (*.rtf)
SOT Office 2003 (*.rtf)
Sun StarOffice 5.x (*.rtf)

Also note that support of a special application for **Korean** documents, the Hanguel wordprocessor, was added!



Text Format

Output

○ Send to — Microsoft Word 97 / 2000 / 2002 / 2003 (RTF)

● External file — Hangeul (*.rtf)

☑ Open after saving

□ Send by e-mail

Atlantis Ocean Mind (*.rtf)
Avant Force Avant Browser 10.0 (*.htm)
E-Press EasyOffice (*.rtf)
FlashPeak SlimBrowser 4.01 (*.htm)
Hangeul (*.rtf)
HTML (*.htm)
Jarte 1.x (*.rtf)
Macromedia Dreamweaver MX 6.1, MX 2004 (*.htm)
Microsoft Excel (*.htm)
Microsoft FrontPage 2002, 2003 (*.htm)

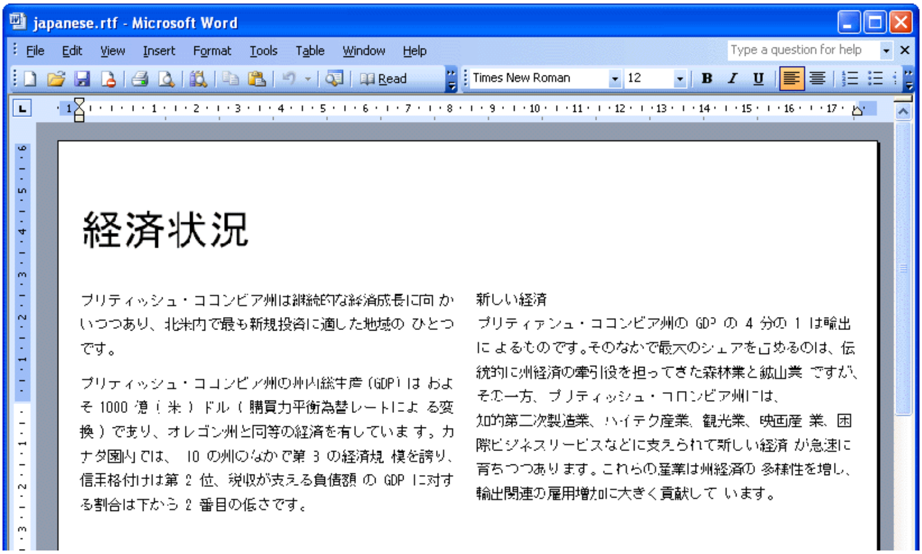
Layout

○ Create body text

○ Retain word and paragraph fo

Learning is disabled as soon as you activate an Asian language, so you'll never enter the interactive phase at the end of the recognition. Learning hardly makes sense for these languages which use thousands of different symbols, and you'd have to be able to enter the ideograms, not an easy task when using a Western keyboard!

Let's see what our Japanese text looks like.



経済状況

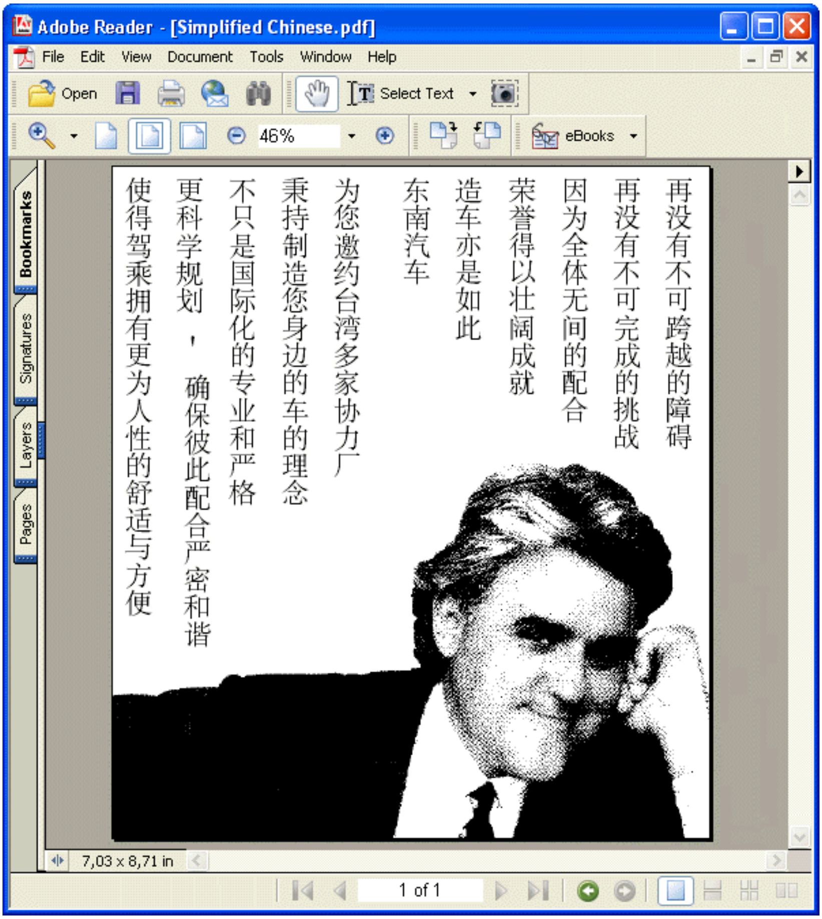
ブリティッシュ・ココンビア州は継続的な経済成長に向 かいつつあり、北米内で最も新規投資に適した地域の ひとつです。

ブリティッシュ・ココンビア州の州内総生産（GDP）は およそ 1000 億（米）ドル（購買力平衡為替レートによ る変換）であり、オレゴン州と同等の経済を有していま す。カナダ国内では、 10 の州のなかで第 3 の経済規 模を誇り、信用格付けは第 2 位、税収が支える負債額 の GDP に対する割合は下から 2 番目の低さです。

新しい経済

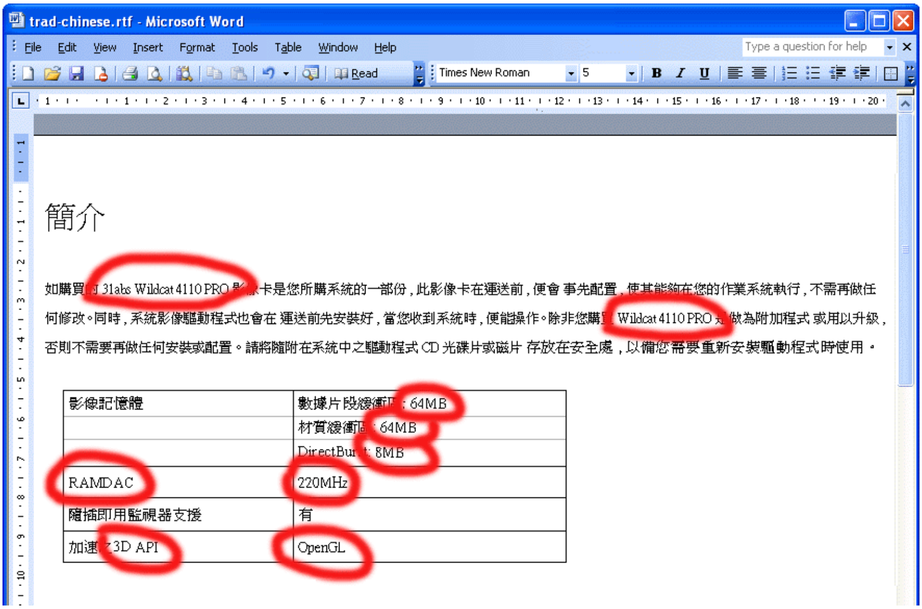
ブリティアンュ・ココンビア州の GDP の 4 分の 1 は輸出によるものです。そのなかで最大のシェアを占めるのは、伝統的に州経済の牽引役を担ってきた森林業と鉱山業 ですが、その一方、ブリティッシュ・コロンビア州には、
知的第二次製造業、ハイテク産業、観光業、映画産 業、国際ビジネスサービスなどに支えられて新しい経済 が急速に育ちつつあります。これらの産業は州経済の 多様性を増し、輸出関連の雇用増加に大きく貢献して います。

The orientation as applied to your source document is maintained across the recognition.



再没有不可跨越的障碍
再没有不可完成的挑战
因为全体无间的配合
荣誉得以壮阔成就
造车亦是如此
东南汽车
为您邀约台湾多家协力厂
秉持制造您身边的车的理念
不只是国际化的专业和严格
更科学规划，确保彼此配合严密和谐
使得驾乘拥有更为人性的舒适与方便

Also notice that these Asian texts may contain Western symbols - numbers, untranscribable proper names etc. As with Greek-English and the Cyrillic-English language settings, Readiris in fact uses a mixed alphabet to encode these documents.


trad-chinese.rtf - Microsoft Word
File Edit View Insert Format Tools Table Window Help
Type a question for help
Read
Times New Roman
5
B I U

簡介

如購買的 3Dlabs Wildcat 4110 PRO 影像卡是您所購系統的一部份，此影像卡在運送前，便會 事先配置，使其能夠在您的作業系統執行，不需再做任何修改。同時，系統影像驅動程式也會在 運送前先安裝好，當您收到系統時，便能操作。除非您購買 Wildcat 4110 PRO 是做為附加程式 或用以升級，否則不需要再做任何安裝或配置。請將隨附在系統中之驅動程式 CD 光碟片或磁片 存放在安全處，以備您需要重新安裝驅動程式時使用。

影像記憶體	數據片段緩衝區：64MB
	材質緩衝區：64MB
	DirectBurst: 8MB
RAMDAC	220MHz
隨插即用監視器支援	有
加速之3D API	OpenGL