



ABBYY®

FineReader® OCR XIX

Based on FineReader 7.0

First Omnifont OCR for Fraktur and Old European Language Recognition

ABBYY FineReader XIX is a special version of the award-winning FineReader optical character recognition (OCR) software for recognising “fraktur” or “black letter” texts from the period between 1800 and 1938. It is designed to convert scans of old documents, books, and papers into text for the purpose of digital archiving and publishing, and it is the first omnifont OCR software for Fraktur.

The Challenge: Digitising Old Texts

Until recently, the limitations of technology and the unique characteristics of text written in a variety of old-fashioned fonts and scripts have made it difficult to automate the process of recording this information via computer. Sophisticated OCR dictionaries, language models used for analysing and verifying text written during this time period, have not existed. Computer systems capable of reading old texts have required many hours of systematic training to recognise fonts and characters that are no longer used in modern printing.

Black letter fonts, also known as “Gebrochene Schriften” or broken scripts, first emerged in as early as the 12th

century, and evolved over the years to host a variety of derivations and font types. The Fraktur typeface, dominant in Germany, was created on behalf of the German Emperor Maximilian and soon became popular in many parts of Europe. Common characteristics and peculiarities of the type include the elongated s and ligatures, “joined” letters for certain letter combinations. The frequency of its application makes understanding of Fraktur essential for studying text and developing recognition technologies for the period between 1800 and 1938.

The Solution: First Omnifont OCR for Fraktur

ABBYY FineReader XIX is the first omnifont OCR for Fraktur, giving users a solution for scanning and converting old documents with minimal training and dictionary work. This was achieved by combining extremely intelligent technology with dedicated linguistic study:

OCR systems work by analysing a text image and making a hypothesis about which letter or word an image represents. The hypotheses are analysed in context and verified by use of sophisticated OCR dictionaries made up of Language Models (LMs). Language Models (LM) are computer databases that describe the vocabulary of a language. The problem is that modern OCR systems do not have LMs for older text fonts and older text spellings. The solution for Fraktur text recognition was achieved through the development of OCR dictionaries specifically for this time period. Special language models were created for five European languages.

The Fraktur language models were created with the help of ABBYY partner, ATAPY software. Through development process, 10 different dictionaries and more than 105 books published between 1808 and 1930 were analysed. Linguists reviewed word stock, identified words that have phased out through the evolution of the

languages, and identified the correct paradigm assignments for synchronising the language models with the appropriate grammar usage for the time period. More than 500.000 word entries were manually compared with existing FineReader dictionaries. Grammatical paradigms and word evolutions were reviewed to add 159 historic grammar paradigms that were missing from the contemporary language models. Language models were then compiled and tested on a control group of testing documents featuring old text.

To recognise the Fraktur style fonts, ABBYY development teams created special classifiers, or alphabets, capable of recognising the Fraktur symbols. As part of this effort, ABBYY development teams collected a symbol image base with an average of 2500 symbol samples for each symbol, a new alphabet pattern, and collected and input a sample test base representing 31000 pages of text from different sources. Using the sample text, the recognition engine was “fine-tuned” to work with the subtle features of the Fraktur alphabet (such as the ligatures, or connected letters). The new alphabet was then added to the FineReader system and interface and tested extensively.

Created in cooperation with major archiving institutions

ABBYY FineReader XIX was also developed with the needs of universities and research center in mind. The product was developed through a cooperation with the worldwide METAe Project. METAe is a consortium of libraries and digitisation companies from across Europe who are working together to create the METAe Engine, a software package specifically designed for organising the work flow of the archiving and conversion of historical materials such as books, journals, magazines and news-

papers. ABBYY FineReader XIX will provide a key component for archiving some of Europe’s most priceless historical documents. Partners in the METAe project include: the University of Innsbruck (Austria), University of Florence (Italy) Bibliothèque Nationale de France, the National Library of Norway, the Friedrich-Ebert-Foundation (Germany), CCS Compact Computer Systeme (Germany), and Cornell Library University (USA).

Key Features and Functionality:

ABBYY FineReader XIX is based on ABBYY's FineReader Corporate Edition OCR software. It features a user-friendly interface with a Scan&Read Wizard that guides users through the process of scanning and converting a document. Converted text is easy to edit and save into a variety of popular file formats, including Microsoft Word, regular text, and searchable PDF file formats. Texts are converted accurately and the format of the original document is maintained so that columns, placement of pictures, and tables, all appear as they do in the original document.



On top of FineReader's basic OCR functions, FineReader XIX is capable of reading old texts that feature elaborate type prints. This includes text with ornamental curls that break the continuous line of the word and roman type characters no longer in use such as the elongated "s" used in early English or French text. FineReader XIX support for Fraktur includes:

Languages:

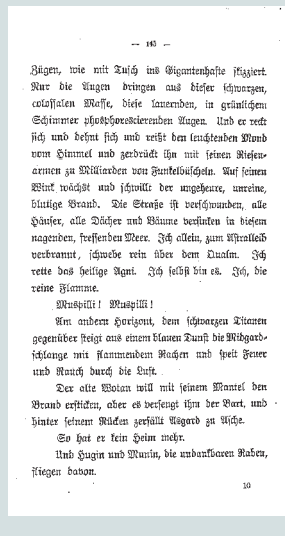
German, English, French, Italian, and Spanish

Fonts:

Fraktur, Schwabacher, and a majority of Textura (Gothic) fonts

Output/Saving to formats:

PDF, MS Word (.DOC), MS Excel (.XLS), WordPro, WordPerfect, RTF, HTML, DBF, CSV, TXT and MS Word XML



Additional Support for Developers:

ABBYY FineReader XIX is also available for use with ABBYY FineReader development tools including the ABBYY FineReader 7.0 Scripting Edition API and ABBYY FineReader 7.1 Engine Software Development Kit. Insitutional developers and service providers can either customise the FineReader application or insert ABBYY OCR for Fraktur into other applications such as document archiving and retrieval systems.

For additional information on the use of ABBYY FineReader XIX with FineReader development tools, please contact the ABBYY Europe sales team at sales@abbyeu.com.

Trial version:

ABBYY offers a fully-functional trial version of FineReader XIX. The trial version is limited in time and the number of pages which can be processed. Please contact ABBYY sales team to get yourself a free trial version.

Specifications

Interface Languages

FineReader XIX supports 17 interface languages: Bulgarian, Czech, Dutch, English, Estonian, French, German, Hungarian, Italian, Lithuanian, Polish, Portuguese, Russian, Slovak, Spanish, Swedish and Ukrainian.

System Requirements

- PC with Intel® Pentium®/Celeron®/Xeon™, AMD K6/Athlon™/ Duron™ or compatible processor with a minimum of 200 MHz
- Microsoft Windows 2003, Windows XP, Windows 2000, Windows NT 4.0 (SP6 or later), Windows Me/98 (to work with localized interface, corresponding language support is required)
- 64 MB RAM for Windows 2003/XP/2000/NT 4.0; 32 MB RAM for Windows Me/98. An additional 16 MB of RAM is required for each additional processor in a multi-processor system.
- 230 MB hard-disk space for typical installation, 70 MB hard-disk space for program operation
- 100% TWAIN-compatible scanner, digital camera, or fax modem
- Video card and monitor (min. resolution 800x600)
- Keyboard, mouse or other input device

Input/Output Formats

Supported Image Input Formats:

- BMP: black and white, gray, color
- PCX, DCX: black and white, gray, color
- JPEG: gray, color
- JPEG 2000, part 1: gray, color
- PNG: black and white, gray, color
- TIFF: black and white, gray, color, multi-image. Methods of compression: Unpacked, CCITT Group 3, CCITT Group 3 FAX(2D), CCITT Group4, PackBits, JPEG, ZIP
- PDF

Document Saving Formats:

- Microsoft Word Document (*.DOC)
- Rich Text Format (*.RTF)

- Microsoft Word XML Document (*.XML) (Microsoft Office Word 2003 only)
- Adobe Acrobat Format (*.PDF)
- HTML. FineReader supports various code pages (Windows, DOS, Mac, ISO) and Unicode (UTF-8) encoding.
- Microsoft PowerPoint Format (*.PPT)
- Comma Separated Values File (*.CSV). FineReader supports various code pages (Windows, DOS, Mac, ISO) and Unicode (UTF-16, UTF-8) encoding.
- Plain Text (*.TXT). FineReader supports various code pages (Windows, DOS, Mac, ISO) and Unicode (UTF-16, UTF-8) encoding.
- Microsoft Excel Spreadsheet (*.XLS)
- DBF. FineReader supports various code pages (Windows, DOS, Mac, ISO).

Recognition Languages

- 34 Main Languages, for which FineReader provides dictionary support and spelling check system: Armenian (Eastern, Western, Grabar), Bulgarian, Catalan, Croatian, Czech, Danish, Dutch (Netherlands and Belgium), English, Estonian, Finnish, French, German (new and old spelling), Greek, Hungarian, Italian, Latvian, Lithuanian, Norwegian (Nynorsk and Bokmal), Polish, Portuguese (Portugal and Brazil), Romanian, Russian, Slovak, Spanish, Swedish, Tatar, Turkish and Ukrainian.
- 5 FineReader XIX Languages, for recognition of old European document printed in 17-19th centuries: English, French, German, Italian and Spanish.
- 133 Additional Languages with Latin, Cyrillic or Greek characters (please see a full list of supported languages at www.ABBYY.com)
- 4 Artificial Languages: Esperanto, Interlingua, Ido and Occidental.
- 6 Programming Languages: Basic, C/C++, COBOL, Fortran, JAVA and Pascal.
- Simple chemical formulas.
- Digits.

Barcode Types

- 1D: Check Code 39, Check Interleaved 25, Code 128, Code 39, EAN 13, EAN 8, Interleaved 25, CODABAR (without checksum), UCC Code 128, Code 2 of 5 (Industrial, IATA, Matrix), Code 93, UPC-A, UPC-E and Postnet.
- 2D: PDF 417



Worldwide:

ABBYY Software House
P.O. Box 54, Moscow, 129301
Russia
Tel.: +7-095-783-3700
Fax: +7-095-783-2663
E-mail: sales@abbyeu.com
Internet: www.ABBYY.com

Western Europe:

ABBYY Europe GmbH
Anglerstr. 6, 80339 Munich
Germany
Tel.: +49-(0)89-511159-0
Fax: +49-(0)89-511159-59
E-mail: sales@abbyeu.com
Internet: www.ABBYY.com

Eastern Europe:

ABBYY Ukraine
P.O. Box 23, 02002 Kiev,
Ukraine
Tel.: +380-44-4909999
Fax: +380-44-4909461
E-mail: sales@abbyeu.ua
Internet: www.ABBYY.ua

North/Central America:

ABBYY USA
47221 Fremont Boulevard,
Fremont, CA 94538, USA
Tel.: +1-510-2266717
Fax: +1-510-2266069
E-mail: sales@abbyeu.com
Internet: www.ABBYYUSA.com