



"I've got FineReader XIX installed here on my computer. The Frakturschrift recognition is very good. Even though old text recognition is not a large and growing market, I am sure all the service bureaus here in Germany will be ordering 1 or 2 copies and have it run 7x24"

Johannes Stöpetie
CEO,
ABBYY Europe GmbH

ABBYY
FineReader
Based on FineReader 7.0 **OCR XIX**

Meeting the Challenge of Time

ATAPY Software participates in the development of ABBYY FineReader XIX - an OCR system for reading old European books

Meta-E (<http://meta-e.uibk.ac.at>) is a collaborative initiative undertaken by a consortium of 14 universities from 7 European countries and the US, co-funded by the European Union. The project is focused on providing technology basis for digitization and web-publishing of valuable old printed sources spanning several centuries of European history. For this purpose, an OCR system was required, capable of recognizing historical texts for the period 1800-1938, including those printed with Frakturschrift (an old-styled black-letter typeface prevalent at that time). At that point no omnifont-Frakturschrift systems were available: all OCR products had to be trained on each individual book before processing it. Meta-E coordinators started looking for a high quality OCR package to be augmented according to their requirements. ABBYY FineReader was chosen due to its unrivalled recognition accuracy, support for 176 modern languages, and user-friendliness. ABBYY Software House, the international manufacturer of FineReader product line, took up the project as a direct contractor to carry out the development of the omnifont part (introducing the Frakturschrift graphics to FineReader). The linguistic part of the project was subcontracted to ATAPY Software, ABBYY's long-term partner in OCR and computer linguistics development.

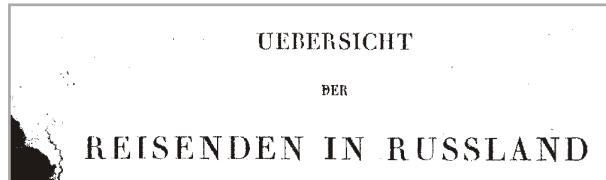
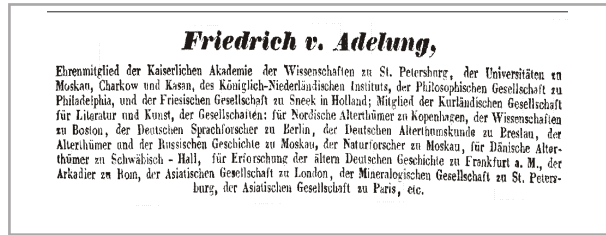
ATAPY's role in the Meta-E project was constructing Old Language Models (Lms) for 5 European languages: English, French, German, Italian, and Spanish. LM is a computer database that describes the vocabulary of a language. FineReader uses LMs during recognition for building OCR hypotheses and spellchecking. LMs are not just full lists of words in all possible grammar forms: such a database would be enormous in size and hardly manageable. FineReader LMs store only stems of each word, and describe the grammar as a set of flexing rules (paradigms). Each stem is assigned a list of paradigms; applying them to the stem produces all possible forms of the word. ATAPY was to study a large amount of authentic dictionaries and original old European texts dating back to the targeted time span, review the word stock, add the words that got phased out of the languages, and correct the paradigm assignments to synchronize the LMs with the actual grammatical practice used at that time.

To complete this task, ATAPY's linguists carefully selected 10 dictionaries reflecting the state of the 5 languages, published between 1808 and 1930. ATAPY had also thoroughly analyzed 105 authentic books of that period, comprising more than 50 MB of text. The next step was to build FineReader LMs. ATAPY's linguists manually compared the information from authentic dictionaries and texts - about 500,000 entries in total - to the existing FineReader vocabularies. This work turned up a total of 458,767 words, from which 61% remained unchanged, and 36% were added to the vocabularies from the analyzed sources. About 3% of the words had their paradigms corrected towards the XVIII-early XX century grammar rules. To carry out such correction, the linguists had to add 159 historic grammar paradigms that were missing in the contemporary models.


Finally, the LMs were compiled and tested on the control text corpus. They manifested 98.91% vocabulary coverage for Old English, 99.16% for Old French, 96.58% for Old German, 98.58% for Old Italian, and 98.79% for Old Spanish languages.

To illustrate the above, let's look at a few samples. A regular FineReader package, or any other contemporary OCR system, will make a lot of mistakes here. For example, "Alterthumskunde" may become "Allerlhumskunde" on the first fragment; on the second fragment, "UEBERSICHT" ("Übersicht" in modern German) gets recognized as two words "UEBERSICHT", etc. These mistakes occur because of two factors. The first is the low printing quality, but there is nothing that can be done about it. The second is the old spelling used in those incorrectly-recognized words. All existing OCR systems are targeted at modern texts and therefore only know modern spelling.

Once the five LMs were merged into FineReader 7 shell, ABBYY was able to offer a specialized product that "knows" the spelling specifics of old European languages - ABBYY FineReader XIX. This product has a much lesser chance of making mistakes in places similar to those shown above. In effect, users will be able to OCR old texts with higher quality, saving much of the time which previously had to be spent on error correction.



FineReader XIX (<http://www.frakturschrift.com>) became a powerful tool to assist the Meta-E consortium in its large-scale digitization work. Moreover, the product was the industry's first box OCR product to recognize Renaissance and Late Medieval sources, a product specially targeted at European libraries and public organizations engaged in preservation and publishing of cultural assets, and at service bureaus helping them fulfill this mission.



ABBYY Europe GmbH is a European department of ABBYY Software House based in Munich, Germany. ABBYY Software House is the manufacturer of software products in the fields of artificial intelligence, document recognition and applied linguistics. One of the most notable products by ABBYY Software House is the optical character recognition package ABBYY FineReader.

©2010 ATAPY Software. All rights reserved.
 ABBYY, ABBYY FineReader and FineReader XIX are registered trademarks of ABBYY Software House.
 All the other trademarks are the property of their respective owners.



ATAPY Software

630090, Enginernaya Street, 4a, 522
 Novosibirsk, Russia
 Tel. +7 383 33 56 569 Fax +7 383 33 56 561
www.atapy.com office@atapy.com

ABBYY Europe Software House

80687 Munich, Germany
 Elsenheimerstrasse 49
 Tel. +49 89 511 159 0 Fax +49 89 511 159 59
www.abbyeu.com info@abbyeu.com