Die Abbyy-OCR-Engine für Linux im Test

Richtig gelesen?

Fehler sind bei optischer Schrifterkennung meist garantiert. Ob sich das mit der seit Anfang März für Linux-Anwender verfügbaren OCR-Engine Finereader ändert, prüft dieser Test. Peter Kreußel

© Sean_Nel, 123RF.com

Bei der Mustererkennung verweist das menschliche Gehirn – anders als beim Rechnen mit mathematischer Formeln – den schnellsten Rechner mühelos in die Schranken. OCR (Optical Character Recognition) erreicht bis heute keine hundertprozentige Erkennungsgenauigkeit, auch wenn sie diesem Ideal inzwischen nahe kommt. Doch eingescannte Bitmaps in Text zu verwandeln bringt handfeste Vorteile: Textdateien lassen sich durchsuchen und maschinell auswerten.

Seit März bietet Abbyy [1] seine renommierte OCR-Engine Finereader als Einzelplatz-Lizenz für Linux an [2]. Wie schon das seit einigen Jahren verfügbare Linux-SDK mit Volumenlizenzen für Entwickler enthält es nur ein Kommandozeilen-Tool, aber keine grafische Scansoftware, wie sie von Windows her bekannt ist.

Die Abbyy-Software enthält einen Seitenzähler, und so schwanken die Lizenzgebühren je nach erlaubter Seitenzahl pro Jahr zwischen 150 Euro und 3500 Euro. Der Kasten „Abbyy Finereader Engine CLI für Linux" zeigt den Leistungsumfang und die Lizenzgebühren im Detail. Mit seinem Angebot zielt Abbyy offensichtlich vorwiegend auf den Unternehmensbereich ab. Aber wer kommerziell mit eingescannten Texten arbeitet, rechnet Lizenzgebühren, Software-Aktivierung und das für Linux-Anwender gewöhnungsbedürftige Abschalten der Software beim Erreichen des Seitenzahllimits nüchtern gegen den Nutzwert auf.

Freiheitsliebend

Beim Abwägen des Kosten-Nutzen-Faktors liegt es nahe, die Abbyy-Software mit den führenden freien OCR-Lösungen zu vergleichen. Aber so viele gibt es da gar nicht: Der einzige ernsthafte Mitbewerber ist Cuneiform (siehe Kasten „Cuneiform für Linux"). Nur dieser einst kommerzielle Konkurrent von Abbyy, der inzwischen als quelloffene Software mit Linux-Port für die Kommandozeile vorliegt, beherrscht das Einlesen von mehrspaltigen Texten ausreichend gut.

Das Ocropus-Projekt, eine weitere OCR-Lösung aus dem universitären Umfeld mit Layout-Erkennung auf Basis des ebenfalls ehemals kommerziellen Tesseract, befindet sich noch in einem frühen Entwick-

Abbyy Finereader Engine CLI für Linux

Kategorie: OCR-Software
Hersteller: Abbyy [http://www.abbyy.com]
Preis, Lizenz: Die Version 8.0 kostet 150 Euro für 12 000 Seiten/Jahr, 3500 Euro für 500 000 Seiten. Eine Version ohne Seitenzahlbeschränkung (ohne Preisangabe) ist erhältlich. Abschaltung nach Erreichen des Limits.
SDK: Seit Mai 2010 auch für Linux in Version 9.0 erhältlich, ab 4900 Euro für drei Entwickler. Unterzeichnen eines NDA erforderlich.
Lauffähig: Nach Herstellerangaben Fedora 10 bis 12, SLES 10 bis 11, Open Suse 10.3 bis 11.1, Debian 4 und 5, Ubuntu 6.06 bis 8.04, im Test auch auf Ubuntu 9.10.
Merkmale: Robuste Texterkennung bei Vorlagen in schlechter Qualität, Layout-Erkennung für Magazinseiten und Formulare geeignet.
Sandwich-PDFs: Aus gescannten Bitmaps und exakt darüber platziertem, unsichtbarem OCR-Text (durchsuchbar, indizierbar).

Cuneiform für Linux

Kategorie: OCR-Software
Hersteller: Cognitive Technologies [http://www.cuneiform.ru/eng/]
Linux-Port: [https://launchpad.net/cuneiform-linux]
Lizenz: Simplified BSD License
Merkmale: Gute Ergebnisse bei qualitativ hochwertigen Scans, Layout-Erkennung bewältigt Magazinseiten ohne Tabellen.
Sandwich-PDFs: Über Exactimage [3]

lungsstadium. Gocr [4] und Ocrad [5], beide von Anfang an Teile der Open-Source-Welt, bieten bisher keine ausreichend leistungsfähige Layout-Erkennung und schneiden bei der Erkennungsgenauigkeit viel schlechter ab [6].

Sieg durch K.o.

Den Einstieg für den Test bildet eine optimale OCR-Vorlage – ein einspaltiger Text auf glattem Druckpapier. Der Scan hat eine Auflösung von 300 dpi in Graustufen, ist gerade und weist einen hohen Kontrast auf (Abbildung 1). Abbyy konvertiert dieses Bitmap fehlerfrei. Auch die freie Cuneiform-Engine schlägt sich ganz gut: Lediglich eine geschweifte statt einer runden Klammer und ein überzähliges Leerzeichen trüben das Bild. Die Erkennung erfolgte bei beiden Engines mit Beschränkung der Erkennungssprache auf Deutsch und HTML als Ausgabeformat. Anders als bei bloßem Text geben beide Engines die Kursivschrift korrekt wieder. Die Abbyy-Engine wählt außerdem wie die Vorlage eine Serifenschrift.

Der Test mit einer einspaltigen Vorlage gibt aber den typischen Büroalltag nicht wirklich wieder. Der Text der zu verarbeitenden Dokumente fließt meist in Spalten, die Seiten enthalten Bilder, das eine oder andere Blatt weist einen Knick oder eine Verschmutzung auf. Die Schriftarten wechseln ebenso wie die Schrift- und Hintergrundfarben.

Der nächste Test, der Farbscan einer Linux-Magazin-Seite in 300 dpi Auflösung, trägt dem Rechnung. Die Seite enthält farbigen Hintergrund, das Druckpapier sorgt für einiges Rauschen im Scan-Ergebnis (Abbildung 2, Ausschnittsvergrößerung). Außer dem Spaltensatz erschwert vor allem die enthaltene Tabelle die Texterkennung.

Abbildung 1: Eine einspaltige Vorlage in optimaler Scanqualität mit kontrastreichen, scharfen Buchstabenkanten bewältigt Abbyy ganz ohne, das freie Cuneiform mit zwei Erkennungsfehlern.

Die Erkennungsergebnisse von Abbyy und Cuneiform unterscheiden sich stark: Die Abbyy-Engine bleibt bis auf einige Überläufe der Tabellenzellen fehlerfrei (Abbildung 2, Mitte). Diese entfallen bei reinem Text oder PDF als Ausgabeformat. Cuneiform dagegen kann mit dem Text vor hellblauem Hintergrund in der Tabelle nichts mehr anfangen (Abbildung 2, rechts). Auch die Reihenfolge der Textblöcke im Spaltensatz ist durcheinandergeraten, der Seitenaufbau hat die Layout-Erkennung völlig überfordert.

Abstiegsgefährdet

Gut bleibt bei der freien Software die Texterkennung in den Bereichen außerhalb der Tabelle mit blauem Hintergrund. Hochglanzqualität müssen die Seiten für eine zufriedenstellende Erkennungsleistung also ebenso wenig wie bei der kommerziellen Engine aufweisen.

Auch gewöhnlichen dreispaltigen Textsatz bewältigt Cuneiform, wie weitere Tests gezeigt haben. Enthält dieser aber Bilder in nicht rechteckiger Form, so quittiert dies die quelloffene Software mit in den erkannten Text eingebettetem Buchstabensalat, während sich die Abbyy-Engine dadurch nicht aus dem Lesefluss bringen lässt.

Außer der Layout-Erkennung, bei der die Abbyy-Engine klar für sich entscheidet, gilt es noch, die Erkennungsleistung bei qualitativ schlechten Vorlagen wie dem zerknitterten Zeitungsausschnitt in Abbildung 3 zu testen. Der Scan ist zudem leicht verkantet, was sich in der Praxis oft nur mit großem Zeitaufwand vermeiden lässt. Auch hier zeigt sich die Abbyy-Software überlegen: Lediglich zwei Fehler schleichen sich bei der Erkennung ein: Die Software „korrigiert“ die historische Kleinschreibung eines Substantivs im Original. Gibt der Anwender »German« als einzige Option für die Erkennungssprache an, fehlt der Akzent auf dem Wort „succès“, das korrigiert aber die Einstellung »German French«.

Cuneiforms Erkennungsleistung spart bei der schlechten Vorlagequalität im Vergleich zum Abtippen kaum noch Arbeit. In derselben Liga wie die kommerzielle Software spielt die freie auch in diesem Test also keinesfalls. ▶

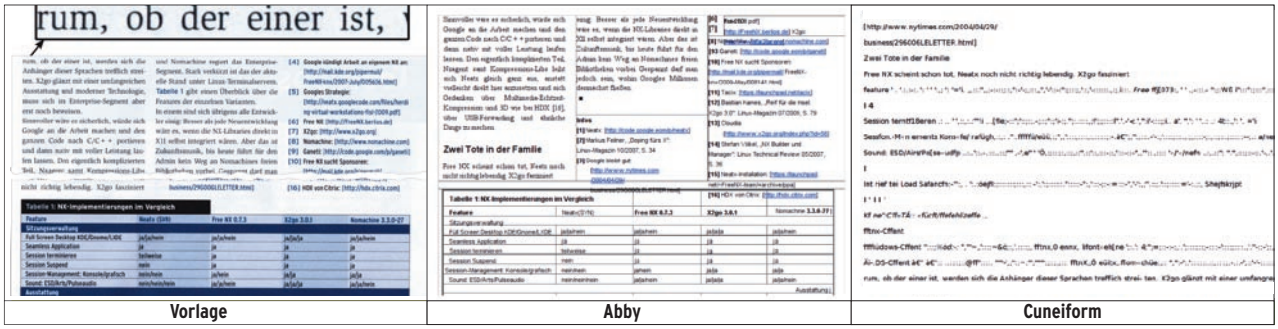

Abbildung 2: Bei der eingebetteten Tabelle scheidet sich die Spreu vom Weizen: Mit dem Tabellentext vor blauem Hintergrund ist Cuneiform restlos überfordert.

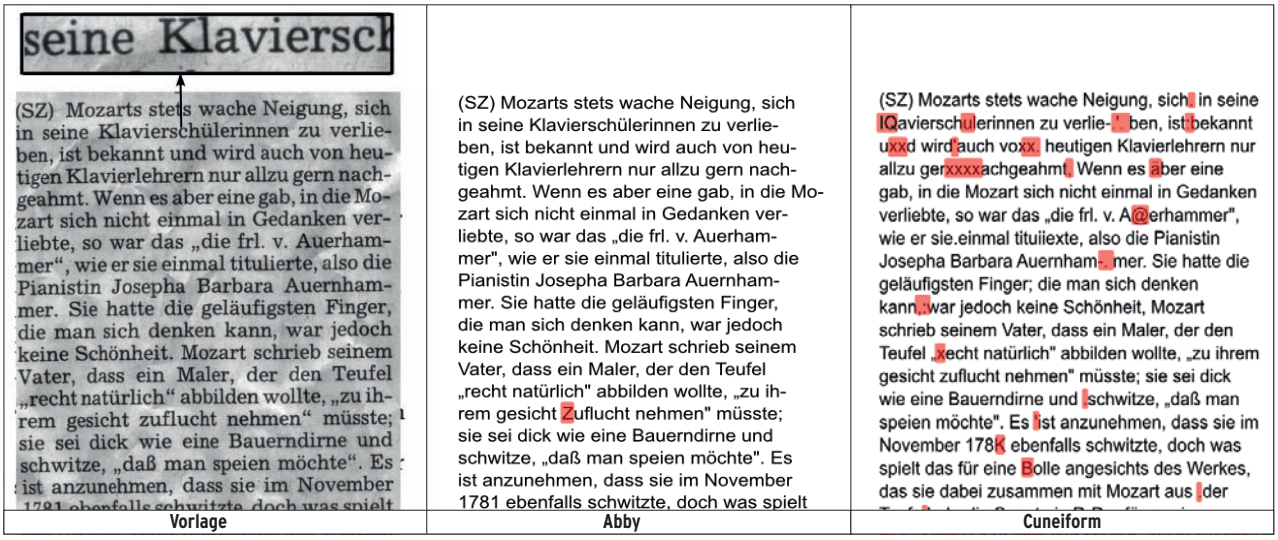

Abbildung 3: Abbyys einzige Abweichungen vom Originaltext der faltigen und kontrastarmen Vorlage sind ein nach heutiger Rechtschreibung zu Unrecht klein geschriebenes Wort und ein fehlender französischer Akzent, wenn Deutsch als einzige Erkennungssprache gewählt ist. Cuneiform hat seine liebe Not mit dem Text.

Eine ebenfalls wichtige Rolle, vor allem im gewerblichen Bereich, spielen die plattformübergreifende Verfügbarkeit kostenloser Betrachtersoftware und die entsprechende Standardisierung. All dies prädestiniert das PDF-Format für die Dokumentarchivierung. Besonders praktisch sind dabei so genannte Sandwich-PDFs. Bei ihnen liegt über dem Scan der Vorlage ein unsichtbarer Textlayer aus der OCR. So zeigt der PDF-Viewer ein unverfälschtes Bild des Originals, das sich dennoch nach Schlagwörtern durchsuchen lässt (Abbildung 4).

Abbyy erstellt solche Sandwich-PDFs ohne Mehraufwand. Der Anwender muss auf der Kommandozeile nur die Optionen »-f PDF -pem ImageOnText« angeben. Fehlt »-pem ImageOnText«, bettet die Software als Grafik erkannte Dokumentbereiche nur als Bitmap in das PDF ein. Zwar platziert Abbyy den Text exakt wie in der Vorlage, Details wie Hintergrundfarbe und Schriftart bleiben dann aber nicht erhalten. Ist sich die Software hinsichtlich des erkannten Texts nicht sicher, blendet sie in den Textfluss Schnipsel des Originals ein (Abbildung 5).

Dies garantiert zwar, dass der Text auch bei unvollständiger Erkennung lesbar bleibt, stört aber den Lesefluss. Wer nicht mit Plattenplatz geizen muss, ist deshalb mit der Sandwich-Variante besser bedient: Abbys OCR-Engine platziert Text und Bild so exakt übereinander, dass der Anwender den Eindruck hat, direkt im sichtbaren Bild zu arbeiten.

Cuneiform dagegen beherrscht PDF als Ausgabeformat nicht selbst. Allerdings kennt die Software das HOCR-Dateiformat. Dabei handelt es sich um HTML-Dateien mit zusätzlichen, im Browser nicht sichtbaren Informationen für eine genaue Platzierung der Elemente. Das für viele Distributionen verfügbare Grafik-Paket Exactimage [3] enthält jedoch das Tool »hocr2pdf«, das die HOCR-Dateien in Sandwich-PDFs umwandelt [7].

Patchwork

Zunächst fällt auf, dass das Duo Cuneiform/Hocr2pdf den Text deutlich weniger exakt übereinanderlegt (Abbildung 4). Im von Cuneiform erzeugten PDF scheint die Markierung mitten im Wort zu enden, obwohl der Tester ganze Wörter ausgewählt hatte. Normalerweise reicht dies aus, um Textstellen mit der Suchfunktion des PDF-Reader zu finden, doch die Illusion, den sichtbaren Text direkt in der grafischen Reproduktion des Originals auszuwählen, ist gestört.

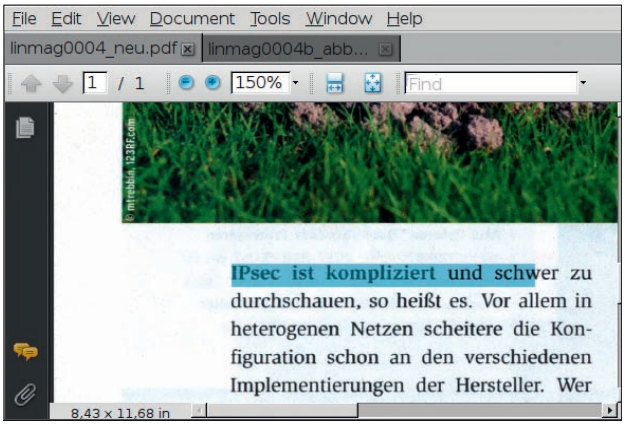

Abbildung 4: In Sandwich-PDFs liegen OCR-Text und Bitmap übereinander. Cuneiform erstellt sie via Exactimage, patzt aber bei der Platzierung des Texts.

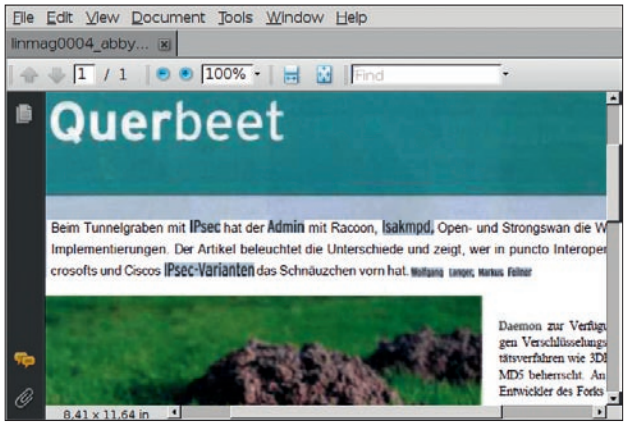

Abbildung 5: Auf Wunsch stellt Abbyy in PDFs den Text auch als Vektorschrift dar und blendet lediglich Bilder und unsicher erkannte Textstellen als Bitmap ein.

Vorlage	Abby	Cuneiform
Dies ist ein Text mit vielen unterschiedlichen Schriftarten **Nicht alle Schriftarten eignen sich *gleich gut* für die automatische Charaktererkennung.** MANCHE SCHRIFTEN SETZEN DER OCR-ENGINE ERHEBLICH MEHR WIDERSTAND ENTGEGEN. Mache stellen sie vor geradezu unlösbare Aufgaben. **Bei einem** *Schriftengemisch* kommt **selbst** die beste *Software* an ihre Grenzen. Wie steht es mit sehr großen Schriften? Und wie steht es mit sehr kleinen?	Dies ist ein Text mit vielen unterschiedlichen Schriftarten Nicht alle Schriftarten eignen sich gleich gut für die automatische Charaktererkennung. Manche Schriften setzen der OCR-Engine erheblich mehr Widerstand entgegen. Mache stellen sie vor geradezu unlösbare AuPgaben. Bei einem Schriftengemischkomrnt selbst die beste software an ihre Grenzen. Wie steht es mit sehr großen Schriften? Und wie steht es mit sehr kleinen?	Dies ist ein Text mit vielen unterschiedhchen Schriftarten Nicht alle Schriftarten eignen sich gleich gut fur die automatische Charaktererkennung. MANCHE SCHRIFTEN SETZEN DER OCR-ENGINE ERHEBLICH MEHR WIDERSTAND ENTGEGEN. Mache stellen sie vor ger adezu unlösbare AuFgaben. Sei einem Schri ftengemisch kommt SelbSt die beste stiaware an ihre Grenzen. Wie steht es mit sehr großen Schriften? Und wie steht es mit sehr kleinen?

Abbildung 6: Bei willkürlich wechselnden Schriftarten kommt selbst die beste OCR-Software an ihre Grenzen.

Außerdem komprimiert Hocr2pdf das Bitmap vor dem Einbetten in das PDF nicht. Um auf vernünftige Dateigrößen wie die etwa 130 KByte für eine Linux-Magazin-Seite zu kommen, muss der Anwender neben der hochauflösenden Scanvorlage noch eine komprimierte Version des Bitmap erstellen. Bei Abbyy, das diese Arbeitsschritte mit einem Kommandozeilen-Aufruf erledigt, hat der Anwender auch noch die Möglichkeit, Auflösung und Kompressionsqualität einzustellen.

Bunte Mischung

Die Abbyy-Software kommt auch mit unterschiedlichen Schriftarten klar. Sie hat keinerlei Probleme mit Kursivschrift und bewältigt auch die Fonts Aachen und Engravers Gothic (**Abbildung 6**, Absatz zwei und drei). Das erste Mal strauchelt sie bei dem f gegen Ende des vierten Absatzes, das sie für ein P hält. Wer sich die Letternform ansieht, erkennt, dass das f in der Schriftart Balker tatsächlich einem beschädigten P ähnelt. Bei dekorativen Schriften aus Plakaten oder Printprodukten mit ungewöhnlicher Buchstabenform steigt also die Wahrscheinlichkeit von Erkennungsfehlern.

Im fünften Absatz mit wechselnden Schriften fehlt im erkannten Text ein Leerzeichen, ein m wird zu einem nr, ein großes S zu einem kleinen. Gut, dass solche Schriftenmischungen in der Praxis selten vorkommen. Die letzen beiden Absätze zeigen, dass Abbyy starke Schriftgrößenunterschiede bewältigt.

Cuneiform schneidet bei dem Schriftenmix mit sieben statt vier Erkennungsfehlern schlechter ab. Doch so vollständig wie beim Tabellentext auf blauen Hintergrund streckt die Software die Waffen nicht: Die meisten Erkennungsfehler (vier) unterlaufen in dem Satz mit gemischten Schriften, eine Schwäche, die sich in der Praxis selten auswirkt. Schade jedoch, dass Cuneiform schon im ersten Satz in kursiver URW Palladio patzt, denn dabei handelt es sich um einen gewöhnlichen Schrifttyp. Das f in der ungewöhnlich schlank und rund geschnittenen Schriftart Balker erkennt auch Abbyy falsch (**Abbildung 6**, vierter Absatz gegen Ende).

Hoch entwickelt

OCR ist eine Technik, die sich nur mit großem Entwicklungsaufwand bewältigen lässt. Die freien Engines Gocr und Ocrad hinken weit hinterher. Auch den ehemals kommerziellen, nun freien Anwendungen Cuneiform und Tesseract merkt man an, dass die Entwicklung vor der Freigabe als Open Source einige Jahre ruhte. Sie kommen mit guten Vorlagen zufriedenstellend zurecht, Cuneiform auch mit Seitenlayouts, die eine gewisse Komplexität nicht übersteigen. Mit typischem Magazin-Layout oder gar mit Formularen ist die Layout-Erkennung der freien Engine aber überfordert.

Die Abbyy-Engine schlägt sich in allen Disziplinen spürbar besser. Besonders die unbeirrbare Layout-Erkennung und die ohne Umwege erzeugten Sandwich-PDFs erweisen sich im Büroalltag als wertvoll. Wer dagegen Dokumente ohne Tabellen in guter Scanqualität verarbeitet und ohnehin Korrektur liest, fährt mit dem freien Cuneiform nicht schlecht und kann sich Lizenzgebühren und Gängelung durch einen Seitenzähler sparen.

Kurz vor Redaktionsschluss hat Abbyy ein Linux-SDK von Version 9.0 seiner OCR-Engine vorgestellt, eine Lizenz für drei Entwickler kostet 4900 Euro und verpflichtet zum Unterzeichnen eines NDA (Non Disclosure Agreement). Eine Lizenz für eine damit erstellte Anwendung gibt's (einmalig) ab 360 Euro/Stück **[8]**. Auch die gegenwärtig noch in Entwicklung befindliche Version 10 plant Abbyy unter Linux als SDK und Kommandozeilen-Programm zur Verfügung zu stellen, allerdings erst 2011.

In Version 9 soll eine Dokumentstruktur-Analyse das OCR-Ergebnis konsistenter formatieren, das Preprocessing gleicht Drehungen bis 20 Grad und Verzerrungen aus. Neu sind auch eine Unterstützung des für Archivzwecke ISO-zertifizierten PDF/A- sowie der MS-Office-2007-Formate. Für geringere Dateigrößen kann der Anwender nun die Komprimierung unterschiedlicher Dokumentbereiche getrennt einstellen, was laut Abbyy bis zu zehnmal kleineren Dateien ergibt. Eine Testversion gibt es unter **[9]**. *(mfe)* ■

Infos

[1] Abbyy: [http://www.abbyy.com]
[2] Abbyy-Engine für Linux: [http://www.ocr4linux.com]
[3] Exactimage: [http://www.exactcode.de/site/open_source/exactimage]
[4] Gocr: [http://jocr.sourceforge.net]
[5] Ocrad: [http://www.gnu.org/software/ocrad]
[6] Peter Kreußel, „Freie Schrifterkennungs-Software": Linux-Magazin 12/06, [http://www.linux-magazin.de/Heft-Abo/Ausgaben/2006/12/Abc-Schuetzen]
[7] Sandwich-PDF mit Exactimage: [http://wiki.ubuntuusers.de/Cuneiform-Linux#PDF-Erstellung]
[8] SDK-Lizenz: [http://www.abbyyeu.com/sdk/en:business:pricing]
[9] Abbyy-Testversion: [http://www.ocr4linux.com/:de:trial]