Succeed List of Tools

												1st criterion for exclusion: If there is no trial version available to test a tool within Succeed, the tool will be discarded		2nd criterion for exclusion: If there is no technical documentation or support available, the tool will be discarded		3rd criterion for exclusion: If there is no information about the tool being used in other projects, no information about existing benchmarks or no information from users about the tool, the tool will be discarded		Relevance for libraries: A rating from 0-5 indicates how relevant a tool is for libraries (0 = not relevant; 5 = very relevant). The description provides further insights on how the consortium came to this rating.	
Name of the tool	Group	Type	Subtype	Description	Link to the tool/website	Entry author	Type of license	Language support	Tech. context	Time and effort for installation	Trial version available (Yes/No)	Further Description	Documentation/S available? (Yes/No)	Further Description (e.g. available support, activity status/last update)	Information assuring tool performance available? (Yes/No)	Further Description (Applicability to mass digitisation, quality and robustness)	Rating (0-5)	Further Description
Aletheia	Evaluation	Layout	GT production	GUI-based document layout and text ground truthing system: a comprehensive tool for semi-automated production of ground truth and annotation of document images on page level	http://www.digitisation.eu/tools/evaluation-toolkit/aletheia/	Sebastian	commercial (although web-version coming in second half 2013 will be open source)		C++ / Javascript	Minimal. Packed EXE.	Yes	http://www.primaresearch.org/tools.php	Yes	University of Salford	Yes	Used by service providers in IMPACT to create ground-truth (approx. 50k pages)	4	
Evaluation Tool for OCR	Evaluation	OCR (text)		This tool evaluates the performance of an optical character recognition system on character and word level.	http://www.digitisation.eu/tools/browse/evaluation/evaluation-tool-for-ocr/	Sebastian	unknown	Not applicable	C++, MSI Installer	Minimal	Yes	Based on ISRI evaluation tool source code: https://code.google.com/p/isri-ocr-evaluation-tools/	Yes	Yes (IMPACT deliverable document - missing on digitisation.eu website?)	Yes	Stable for single column, plain text files based evaluation. Supports batch processing. Web service available.	4	
GEDI Ground Truthing Environment	Evaluation	OCR (text)	GT production	GEDI is a generic annotation tool that assists you in ground truthing scanned text documents. Its basic structure involves two types of files, an Image file, and a corresponding .xml file in GEDI Format	http://lampsrv02.umiacs.umd.edu/projdb/project.php?id=53	Katrien&	Own license		Java		Yes		Yes	research group http://lamp.cfar.umd.edu/contact.htm	No	last updated 2011	0	
Ground Truth Maker	Evaluation	OCR (text)	GT production	An application based on the lexicon defined by historians NaviDoMass. This application allows you to create the ground truth associated to an image to test various tools.	http://navidomass.univ-lr.fr/gtm.html	Katrien&	unknown		.NET		No		No				0	
GTText	Evaluation	OCR (text)	GT production	OCR free software and Ground Truthing tool for Color Images with Text: The gttext project helps to create fast and quality Ground Truthed data-sets from color text images.	https://code.google.com/p/gttext/	Katrien&	GPLv2		C++		Yes		Yes	single researcher (David Torne Berga); result master thesis	No		0	
ISRI Tools	Evaluation	OCR (text)	evaluation	Images and Ground Truth text and zone files for several thousand English and some Spanish pages that were used in the UNLV/ISRI annual tests of OCR accuracy between 1992 and 1996. Source code of OCR evaluation tools used in the UNLV/ISRI annual tests of OCR Accuracy.	https://code.google.com/p/isri-ocr-evaluation-tools/	Katrien&	ASL 2.0		C		Yes		Yes	community	Yes	32 sets of images and txt files for OCR evaluation + tool, combined with wrapper script to make the tool work with UTF8. Further info: http://www.stephenvrice.com/images/AT-1993.pdf	4	
Layout Evaluation	Evaluation	Layout		Performance evaluation tool for layout analysis and segmentation methods based on detailed metrics (types of errors such as merges, splits, missed regions, etc.) and use scenarios	http://www.digitisation.eu/tools/evaluation-toolkit/evaluation-tool-for-segmentation/	Sebastian	unknown				Yes	http://www.primaresearch.org/tools.php	Yes		Yes		4	
MILE ocr-performance-evaluator	Evaluation	OCR (text)	evaluation	A desktop application used for performance evaluation of Optical Character Recognizers (OCR). Implemented using Eclipse SWT and runs on Windows & Linux.	https://code.google.com/p/ocr-performance-evaluator/	Katrien&	ASL 2.0		Java		Yes		Yes	community	No	Used for Indian languages, project from MILE Lab, Indian Institute of Science, Bangalore.	0	
Abbyy Binarisation and Colour Reduction	Image Processing	Image Processing and Enhancement		Use this toolkit when building your own OCR workflow out of various tools from various vendors.	http://www.digitisation.eu/tools/image-enhancement-toolkit/binarisation-and-colour-reduction/	Sebastian	commercial	Many...(add link)	CL tool / Web Service / SDK	Depending on chosen platform. SDK requires compilation using MS Visual Studio.	Yes	Directly from Abbyy homepages	Yes	active	Yes	Recognition Server product for mass digitisation (allthough also possible with SDK)	1	only interesting for libraries who are doing their own OCR
Agora	Image Processing	Image Processing and Enhancement	Image enhancement	Analysis and indexation of Historical and Degraded Documents: pre-processing, layout analysis and character recognition	http://www.rfai.li.univ-tours.fr/PagesPerso/jyramel/gb/work1.html	Tomasz	unknown		C		Yes		No		No	Seems to be very old	0	
Block Segmentation	Image Processing	Image Segmentation		Before characters and words can be recognised by an OCR engine, the print space of the image has to be identified, and from there paragraphs and lines. This tool can be used to identify blocks on a scanned document.	http://www.digitisation.eu/tools/browse/segmentation/block-segmentation	Sebastian	commercial		SDK	Major, needs implementation of executable via SDK. SDK ships demo apps.	No	Request from Abbyy	Yes	active	Yes		0	
Border Detection and Removal	Image Processing	Image Processing and Enhancement		This tool detects and removes noisy black borders as well as noisy text regions. Moreover, it detects the optimal page frames of double page document images.	http://www.digitisation.eu/tools/image-enhancement-toolkit/border-detection-and-removal/	Sebastian	commercial	Not applicable	CL tool	Minimal. MSI-installer	Yes	Available through contacting NSCR	Yes	active	Yes	Not suitable for mass digitisation. Tests revealed issues with multi-columns, illustrations, initials.	2	relevant, but tool has some limitations
Character Segmentation	Image Processing	Image Segmentation		The developed methodology takes as input isolated words and separates them into characters.	http://www.digitisation.eu/tools/browse/segmentation/character-segmentation	Sebastian	commercial		CLI tool / DLL	Minimal for EXE, DLL requires integration with 3rd party tool.	Yes	Available through contacting NSCR	Yes	active	Yes	Not really, more suitable for integration into OCR engines.	0	
Document Deskewer	Image Processing	Image Processing and Enhancement		generic skew detection and correction (for the full range 0-360 degrees) for documents printed using Roman scripts	http://www.iais.fraunhofer.de/diensteplattform-technologien.html	Sebastian	commercial		CL tool / Web Service		Yes	Free to use in Succeed	Yes	active (2012)	Yes	Used in various research and industry projects	3	might need som e adjustments for "difficult" images
Geometric Correction: Arbitrary Warping	Image Processing	Image Processing and Enhancement		Software for correction of arbitrary local distortions in scans of historical documents	http://www.digitisation.eu/tools/image-enhancement-toolkit/geometric-correction-arbitrary-warping/	Sebastian	commercial	Not applicable	CL tool	Minimal. Packed EXE.	Yes	Available through USAL	Yes	unknown	Yes	Not suitable for mass digitisation. Prototype implementation of algorithm.	1	prototype based on PhD
Geometric Correction: Page Curl	Image Processing	Image Processing and Enhancement		This tool rectifies document images which suffer from warping and perspective distortions	http://www.digitisation.eu/tools/image-enhancement-toolkit/geometric-correction-page-curl/	Sebastian	commercial	Not applicable	CL tool	Minimal. MSI-installer	Yes	Available through contacting NSCR	Yes	active	Yes	Not suitable for mass digitisation. Tests revealed issues with multi-columns, illustraions, initials.	2	relevant, but tool has some limitations
GIMP	Image Processing	Image Processing and Enhancement		GIMP is the GNU Image Manipulation Program. It is a freely distributed piece of software for such tasks as photo retouching, image composition and image authoring.	http://www.gimp.org/	Clemens	GPL	Not applicable. UI is translated into most languages.	C	Minimal on any Linux system. For Windows, seperate installers are provided (.exe).	Yes		Yes	community, April 2012 (stable)	Yes	Yes. Stable and widely used. Can be highly automated via command-line or gimp-fu scripts. High quality achievable with appropriate configuration effort.	4	very powerful, used by KB
Hectography Foreground Extractor	Image Processing	Image Processing and Enhancement		foreground-background separation in solor (3 channel) scans of hectographic copies, allowing an order of magnitude improvement in OCR quality	http://www.iais.fraunhofer.de/diensteplattform-technologien.html	Sebastian	commercial		CL tool / Web Service		Yes	Free to use in Succeed	Yes	active (2012)	No		0	
Hot Metal Font Enhancer	Image Processing	Image Processing and Enhancement		font enhancement of prints produced hot metal typesetting allowing higher OCR accuracy	http://www.iais.fraunhofer.de/diensteplattform-technologien.html	Sebastian	commercial		CL tool / Web Service		Yes	Free to use in Succeed	Yes	active (2012)	No		0	
ImageMagick / GraphicsMagick	Image Processing	Image Processing and Enhancement		ImageMagick is a software suite to create, edit, compose, or convert bitmap images. GraphicsMagick is the swiss army knife of image processing. It has been derived from ImageMagick 5.5.2	http://www.imagemagick.org/ / http://www.graphicsmagick.org/	Sebastian	Apache License v2 / MIT		CL tool (cross platform)		Yes	Open Source	Yes	documentation and community, active (2013)	Yes		5	
Leptonica	Image processing	Image Processing and Enhancement	toolbox	Leptonica is a pedagogically-oriented open source site containing software that is broadly useful for image processing and image analysis applications.	http://www.leptonica.com/	Tomasz	Own license (similar to ASL)		C		Yes		Yes	community	Yes		1	
Line and Word Segmentation	Image Processing	Image Segmentation		Segmentation of text regions into text lines and words independent of text recognition (OCR).	http://www.digitisation.eu/tools/segmentation-toolkit/line-and-word-segmentation/	Sebastian	commercial		CLI tool	Minimal. Packed EXE.	Yes	Available through USAL	Yes	active	Yes		3	
NCSR Binarisation and Colour Reduction	Image Processing	Image Processing and Enhancement		Perform image binarisation using an algorithm developed at NCSR.	http://www.digitisation.eu/tools/image-enhancement-toolkit/binarisation-and-colour-reduction/	Sebastian	commercial	Not applicable	CL tool / Web Service	Minimal. MSI-installer	Yes	Available through contacting NSCR	Yes	actively developed	Yes	Partly suitable for mass-digitisation. Using it for binarisation of 10 million images (issues with processing batches greater than 300k pages)	1	only interesting for libraries who are doing their own OCR
Scan Tailor	Image Processing	Image Processing and Enhancement		Scan Tailor is an interactive post-processing tool for scanned pages. It performs operations such as page splitting, deskewing, adding/removing borders, and others.	http://scantailor.sourceforge.net/	Sebastian	GPL v3		C++ standalone GUI tool for Windows	10 min	Yes	Open Source	Yes	documentation and forum, active (2012)	Yes	University Library Bratislava	4	already used by various Libraries

Succeed List of Tools

											1st criterion for exclusion: If there is no trial version available to test a tool within Succeed, the tool will be discarded		2nd criterion for exclusion: If there is no technical documentation or support available, the tool will be discarded		3rd criterion for exclusion: If there is no information about the tool being used in other projects, no information about existing benchmarks or no information from users about the tool, the tool will be discarded		Relevance for libraries: A rating from 0-5 indicates how relevant a tool is for libraries (0 = not relevant; 5 = very relevant). The description provides further insights on how the consortium came to this rating.	
Name of the tool	Group	Type	Subtype	Description	Link to the tool/website	Entry author	Type of license	Language support	Tech. context	Time and effort for installation	Trial version available (Yes/No)	Further Description	Documentation/S available? (Yes/No)	Further Description (e.g. available support, activity status/last update)	Information assuring tool performance available? (Yes/No)	Further Description (Applicability to mass digitisation, quality and robustness)	Rating (0-5)	Further Description
tifftool	Image Processing	Image Processing and Enhancement		Tifftool is a high-performance tool to clean scanned documents in preparation for onscreen display or for OCR	http://sourceforge.net/projects/tifftool/	Sebastian	GPL v2		CL tool (Linux)		Yes	Open Source	Yes	community	No	not many downloads	0	
Unpaper	Image Processing	Image Processing and Enhancement		Unpaper is a post-processing tool for scanned sheets of paper, especially for book pages that have been scanned from previously created photocopies. The main purpose is to make scanned book pages better readable on screen after conversion to PDF. Additionally, unpaper might be useful to enhance the quality of scanned pages before performing optical character recognition (OCR).	http://unpaper.berlios.de/	Clemens	GPL	Not applicable.	C	Minimal on Linux system. Not supported on Windows.	Yes		Yes	community, December 2012	Yes	Limited testing done in IMPACT revealed very good results with some configuration effort. Supports only PBM/PGM/PNM image formats, thus there is extra complexity in pre-/post-conversion	4	produced good results at KB, but doesn't support standard image formats
Document layout analysis tools	Layout Analysis			Intented to be used in Mapa76 processing pipeline for detecting the clusters of text in a PDF file to correctly perform NE dectection to the body of text, excluding other unrelated text lines (like page numbers, titles, footnotes, etc)	https://github.com/munshkr/layout-analysis	Sebastian	unknown		Ruby		Yes		No				0	
Fraunhofer Newspaper Segmenter	Layout Analysis			Award-winning (e.g. ICDAR'09,'11) page and article segmentation for scanned documents featuring complex layouts (e.g. (historical) newspapers, contemporary magazines, text books, etc.)	http://www.iais.fraunhofer.de/diensteplattform-technologien.html	Sebastian	commercial		CL tool / Web Service	web service	Yes	Free to use in Succeed	Yes	active (2012)	Yes	Used in a large newspaper project with the National Library in Berlin	4	
Functional Extension Parser	Layout Analysis			The Functional Extension Parser (FEP) is a Document Understanding Software tool capable of decoding layout elements of books. Based on the output of Optical Character Recognition, layout elements such as page numbers, running titles, headings, and footnotes are detected and annotated.	http://www.digitisation.eu/tools/ocr-post-correction-and-enrichment/functional-extension-parser/	Sebastian	SLA		C++ / Ajax	No installation possible, hosted service / SOAP endpoint	Yes	Available from University of Innsbruck	Yes	actively developed	Yes	Yes. Integrated into the Ebooks on Demand service of the University and has been used for 10 million pages dissertations.	3	very powerful, but software library, might be less suitable for libraries
O2	Layout Analysis		Framework	Library with methods developed for document analysis and recognition	http://www.imglab.org/p/O2/	Katrien&	Own license		C	No binaries available	Yes		No	no more activity since 2009; version 2.0 is last version			0	
Olena	Layout Analysis			A platform dedicated to image processing and pattern recognition. Its core component is a generic and efficient C++ library called Milena. Milena provides a framework to implement simple, fast, safe, reusable and extensible image processing tool chains.	http://www.lrde.epita.fr/cgi-bin/twiki/view/Olena/Download	Sebastian	GPLv2		C++		Yes		Yes		Yes	http://olena.lrde.epita.fr/demos/historical_docume	4	
abbot	Metadata Processing		Format conversion (XML)	Abbot is a tool for undertaking large-scale conversion of XML document collections in order to make them interoperable with one another. Java technology.	https://github.com/CDRH/abbot	Tomasz	https://github.com/CDRH/a		Java		Yes		Yes				4	
Augmented SIP Creator (ASC)	Metadata Processing			The ASC uses XSL scripts to transform Metadata from a source to a target XML format. It can be used to normalize and validate input metadata from heterogenous sources.	http://www.iais.fraunhofer.de/5196.html?&L=1	Sebastian	commercial		Java CL tool	none	Yes		Yes		Yes	Used in the German Digital Library	4	
jmet2ont	Metadata Processing		Format transformation (XML)	A tool that makes it possible to transform metadata from a traditional XML-based schema to RDF/OWL. Mappings are described with XML. Existing mappings used in SYNAT transform traditional library/museum formats to the CIDOC CRM/FRBRoo ontology.	http://fbc.pionier.net.pl/pro/jmet2ont/	Tomasz	GPL		Java		Yes		Yes	PSNC support	Yes	by PSNC	4	
MapForce	Metadata Processing			Altova MapForce® 2013 is an award-winning any-to-any graphical data mapping, conversion, and integration tool that maps data between any combination of XML, database, flat file, EDI, Excel, XBRL, and/or Web service, then transforms data instantly or autogenerates royalty-free data integration code for the execution of recurrent conversions.	http://www.altova.com/mapforce.html	Sebastian	commercial		Windows		Yes		Yes		Yes		0	
OxGarage	Metadata Processing		Format transformation	OxGarage is an web, and RESTful, service to manage the transformation of documents between a variety of formats. The majority of transformations use the Text Encoding Initiative format as a pivot format	https://github.com/sebastianrahtz/oxgarage	Tomasz	unknown		Java		Yes		Yes		Yes		2	
Pandoc	Metadata Processing		Format transformation (XML)	conversion engine	http://johnmacfarlane.net/pandoc/index.html	Tomasz	GNU GPL		C	Installer available	Yes		Yes		Yes	Active	2	
BlackLight	Miscellaneous Utilities		discovery interface	Blacklight is an open source Ruby on Rails gem that provides a discovery interface for any Solr index.	http://projectblacklight.org/	Tomasz	Creative Commons Attribution-Share Alike 3.0 United States License.		Ruby on Rails		Yes		Yes		Yes		2	
Color Target Quality Checker	Miscellaneous Utilities			Fully automatic color target detection from digitized printed material and quality assurance	http://www.iais.fraunhofer.de/diensteplattform-technologien.html	Sebastian	commercial		CL tool / Web Service		Yes	Free to use in Succeed	Yes	active (2012)	No		0	
digilib	Miscellaneous Utilities		creating presentation version	Digilib is a web based client/server image viewing environment for the internet	http://digilib.berlios.de/	Tomasz	GNU GPL				Yes		Yes		No	Quite old	0	
DigitLab	Miscellaneous Utilities		toolset helping with digitisation activities	DigitLab (http://digitlab.psnc.pl) is an especially adapted operating system based on Linux Ubuntu. The main aim of its creation was to create a complete system which can be used for collections digitisation with the usage of free and widely available tools. DigitLab is a perfect solution for both everyday work and hands-on trainings. It allows to work with images, textual content (OCR included) and audio-visual collections. Gives access to three example digital libraries based on DSpace, dLibra and Greenstone.	http://digitlab.psnc.pl/	Tomasz	free				Yes		Yes		Yes		3	
DjVu tools	Miscellaneous Utilities		DjVu toolset	Suit of open source tools and utilities related to the DjVu format	https://bitbucket.org/jsbien/ndt/wiki/wyniki	Tomasz	unknown				Yes		Yes		Yes	Not sure	2	
File-Analyzer	Miscellaneous Utilities			The application allows a user to analyze the contents of a file system or external drive and generates statistics about the contents of the contained directories.	https://github.com/usnationalarchives/File-Analyzer	Sebastian	unknown		Java stanalone GUI tool		Yes		Yes	samples and documentation	No		0	
FromThePage	Miscellaneous Utilities		Transcription	FromThePage is an open-source tool that allows volunteers to collaborate to transcribe handwritten documents.	https://github.com/benwbrum/fromthepage/wiki	Tomasz	AGPL		Ruby		Yes		Yes	community	Yes	Evaluated in http://opus4.kobv.de/opus4-fhpotsdam/files/331/master (German) and http://manuscripttranscription.blogspot.nl/2013/05/choosing-crowdsourced-transcription.html	4	

Succeed List of Tools

Name of the tool	Group	Type	Subtype	Description	Link to the tool/website	Entry author	Type of license	Language support	Tech. context	Time and effort for installation	1st criterion for exclusion: If there is no trial version available to test a tool within Succeed, the tool will be discarded: Trial version available (Yes/No)	Further Description	2nd criterion for exclusion: If there is no technical documentation or support available, the tool will be discarded: Documentation/S available? (Yes/No)	Further Description (e.g. available support, activity status/last update)	3rd criterion for exclusion: If there is no information about the tool being used in other projects, no information about existing benchmarks or no information from users about the tool, the tool will be discarded: Information assuring tool performance available? (Yes/No)	Further Description (Applicability to mass digitisation, quality and robustness)	Relevance for libraries: A rating from 0-5 indicates how relevant a tool is for libraries (0 = not relevant; 5 = very relevant). The description provides further insights on how the consortium came to this rating.: Rating (0-5)	Further Description
hOCR tools	Miscellaneous Utilities			hOCR is a format for representing OCR output, including layout information, character confidences, bounding boxes, and style information. It embeds this information invisibly in standard HTML. By building on standard HTML, it automatically inherits well-defined support for most scripts, languages, and common layout options. Furthermore, unlike previous OCR formats, the recognized text and OCR-related information co-exist in the same file and survives editing and manipulation. hOCR markup is independent of the presentation.	https://code.google.com/p/hocr-tools/	Katrien&	ASL 2.0		Python		Yes		Yes	community (Thomas Breuel)	Yes	used for Tessaract	3	tesseract/google format, might be limited to Google partners
Islandora	Miscellaneous Utilities		Transcription	Javascript based TEI Transcription Editor	https://github. com/Islandora/islandora_tei_editor	Tomasz	unknown		Javascript		Yes		Yes	community			3	
Lightbox	Miscellaneous Utilities		image comparison	The Virtual Lightbox is a software tool for comparing images online.	http://mith.umd.edu/lightbox/	Tomasz	GPL		Java		Yes		Yes		No	Quite old	0	
Metadata Extraction Tool	Miscellaneous Utilities		metadata extraction	The Metadata Extraction Tool was developed by the National Library of New Zealand to programmatically extract preservation metadata from a range of file formats like PDF documents, image files, sound files Microsoft office documents, and many others.	http://meta-extractor.sourceforge.net/	Tomasz	Apache License 2.0				Yes		Yes		No	Not much activity there	0	
METS page turner	Miscellaneous Utilities		creating presentation version	Pure XSLT solution for the display of image files along with selected Descriptive, Administrative and Structural metadata elements of a digital object serialized into an xml-encoded METS document. This application evolved from METSFramesSX.xsl, incorporating a frames-based page turner with search functionality using XPATH.	http://dlib.nyu.edu/metstools/metsviewer/	Tomasz	unknown				Yes		No		No	I guess no	0	
pyBossa	Miscellaneous Utilities		Transcription	Open-source crowd-sourcing (microtasking) platform with a focus on volunteer contribution and making it super-easy to create a crowd-sourcing app.	https://github.com/PyBossa/pybossa	Tomasz	GPLv3		Python		Yes		Yes	community	Yes	Actively in use by the Open Knowledge Foundation http://blog. okfn. org/2012/06/08/introducing-	3	Relatively small set
Scribe	Miscellaneous Utilities		Transcription	Scribe is a framework for generating crowd sources transcriptions of image based documents. It provides a system for generating templates which combined with a magnification tool guide a user through the process of transcribing an asset (an image).	https://github.com/zooniverse/Scribe	Tomasz	ASL 2.0		Ruby		Yes		Yes	community			3	
tb-transcription-desk	Miscellaneous Utilities		Transcription	MediaWiki based environment for a distributed, collaborative transcription effort.	http://code.google.com/p/tb-transcription-desk/	Tomasz	GPLv2		PHP		Yes		Yes	community			3	Relative small set
Textlab	Miscellaneous Utilities		Transcription	An innovative image and text mark-up tool, TextLab is based on the protocols of fluid text editing of revision. Here, "revision sites" are any areas of interest on a manuscript leaf or print page that indicates evidence of revision.	http://mel.hofstra.edu/textlab.html	Tomasz	unknown		Ruby		Yes		Yes	community			3	Very extensive toolset
Alchemy API	Text Processing	NLP Tools	Keyword Extraction	AlchemyAPI is capable of extracting topic keywords from your HTML, text, or web-based content. We employ sophisticated statistical algorithms and natural language processing technology to analyze your data, extracting keywords that can be utilized to index content, generate tag clouds, and more!	http://www.alchemyapi.com/api/keyword/	Bob	Commercial	English, French, German, Italian, Potuguese, Russian, Spanish, Swedish			Yes		Yes	Commercial support	Yes		4	
Alchemy API	Text Processing	NLP Tools	NER	AlchemyAPI provides the world's most popular natural language processing service via an easy-to-use SaaS API. Integrate advanced text mining and analytics functionality into your application, service, or data-processing pipeline.	http://www.alchemyapi. com/products/products-overview/	Bob	Commercial	8	SDKs in all major programming languages		No	Only after receiving a demo license	Yes		Yes		4	
Alchemy API	Text Processing	NLP Tools	Sentiment Mining	AlchemyAPI provides easy-to-use mechanisms to identify positive / negative sentiment within any document or web page. AlchemyAPI Sentiment Analysis APIs are capable of computing document-level sentiment, user-targeted sentiment, entity-level sentiment, and keyword-level sentiment. Multiple modes of sentiment analysis provide for a variety of use cases ranging from social media monitoring to trend analysis.	http://www.alchemyapi.com/api/sentiment/	Bob	Commercial	English, German			Yes		Yes	Commercial support	Yes		4	
Alchemy API	Text Processing	NLP Tools	Text Classification	AlchemyAPI is capable of categorizing your HTML, or web-based content. We employ sophisticated statistical algorithms and natural language processing technology to analyze your information, assigning the most likely topic category (news, sports, business, etc.).	http://www.alchemyapi.com/api/categ/	Bob	Commercial	English, French, German, Italian, Potuguese, Russian, Spanish, Swedish			Yes		Yes	Commercial support	Yes		4	
AlchemyAPI	Text processing	NLP Tools	NLP toolset and resources	AlchemyAPI uses natural language processing technology and machine learning algorithms to extract semantic meta-data from content, such as information on people, places, companies, topics, facts, relationships, authors, and languages.	http://www.alchemyapi.com/	Bob	Commercial		webservices	1d	Yes	Upon request	Yes		yes		4	
Apache openNLP	Text processing	NLP Tools	POS Tagger	The Part of Speech Tagger marks tokens with their corresponding word type based on the token itself and the context of the token. A token might have multiple pos tags depending on the token and the context. The OpenNLP POS Tagger uses a probability model to predict the correct pos tag out of the tag set. To limit the possible tags for a token a tag dictionary can be used which increases the tagging and runtime performance of the tagger.	http://opennlp.apache.org/documentation/1. 5.3/manual/opennlp.html#tools.postagger	Bob	Apache License 2		Linux		Yes		Yes		yes		4	
Apache openNLP	Text Processing	NLP Tools	Tokenizer	The OpenNLP Tokenizers segment an input character sequence into tokens. Tokens are usually words, punctuation, numbers, etc.	http://opennlp.apache.org/documentation/1. 5.3/manual/opennlp.html#tools.tokenizer	Bob	Apache License 2		Linux		Yes		Yes		yes		4	
Apache OpenNLP	Text Processing	NLP Tools	NLP toolset and resources	The Apache OpenNLP library is a machine learning based toolkit for the processing of natural language text.	http://opennlp.apache.org/	Sebastian	Apache License v2		Java library / CL tools		Yes	Open Source	Yes	documentation and community support, active (2013)	No		4	
Apache openNLP	Text Processing	NLP Tools	NER	The Name Finder can detect named entities and numbers in text. To be able to detect entities the Name Finder needs a model. The model is dependent on the language and entity type it was trained for. The OpenNLP projects offers a number of pre-trained name finder models which are trained on various freely available corpora. They can be downloaded at our model download page. To find names in raw text the text must be segmented into tokens and sentences. A detailed description is given in the sentence detector and tokenizer tutorial. Its important that the tokenization for the training data and the input text is identical.	http://opennlp.apache.org/documentation/1. 5.3/manual/opennlp.html#tools.namefind	Bob	Apache License 2	Any	Linux	1d	Yes		Yes		Yes		4	
Apache Stanbol	Text Processing	NLP Tools	NE linking	Apache Stanbol provides a set of reusable components for semantic content management	http://stanbol.apache.org/	Sebastian	Apache License v2		Java web application / REST web service		Yes	Open Source	Yes	documentation and community support, active (2013)	No		4	

Succeed List of Tools

Name of the tool	Group	Type	Subtype	Description	Link to the tool/website	Entry author	Type of license	Language support	Tech. context	Time and effort for installation	1st criterion for exclusion: If there is no trial version available to test a tool within Succeed, the tool will be discarded: Trial version available (Yes/No)	Further Description	2nd criterion for exclusion: If there is no technical documentation or support available, the tool will be discarded: Documentation/S available? (Yes/No)	Further Description (e.g. available support, activity status/last update)	3rd criterion for exclusion: If there is no information about the tool being used in other projects, no information about existing benchmarks or no information from users about the tool, the tool will be discarded: Information assuring tool performance available? (Yes/No)	Further Description (Applicability to mass digitisation, quality and robustness)	Relevance for libraries: A rating from 0-5 indicates how relevant a tool is for libraries (0 = not relevant; 5 = very relevant). The description provides further insights on how the consortium came to this rating.: Rating (0-5)	Further Description
ASV Toolbox	Text Processing		NLP toolset and resources	ASV Toolbox is a modular collection of tools for the exploration of written language data. They work either on word lists or text and solve several linguistic classification and clustering tasks. The topics covered contain language detection, POS-tagging, base form reduction, named entity recognition, and terminology extraction. On a more abstract level, the algorithms deal with various kinds of word similarity, using pattern based and statistical approaches. The collection can be used to work on large real world data sets as well as for studying the underlying algorithms. The ASV Toolbox can work on plain text files and connect to a MySQL database.	http://wortschatz.uni-leipzig.de/~cbiemann/software/toolbox/	Clemens	ASL 2.0	German	Java	Available as Java Framework and binaries (jar). Given an existing JVM, installation is a matter of extracting the ZIP file. Also available as SOAP web services.	Yes		Yes	CLARIN-D	Yes	Yes. More than 6 billion requests served over 9 years, integrated into CLARIN-D as well as several German research projects (AQUA, eTraces)	1	
Brevity	Text Processing	NLP Tools	Summerizer	Businesses and other organizations often deal with hundreds or even hundreds of thousands of documents. Knowing the content of these documents can be difficult. While you can discern the content of a graphical image at a glance, with text documents you have to read through each to discern it's content. Reading through an entire document takes time - time you don't have to waste. The traditional solution to this problem has been to assign people to read the documents and write a brief abstract for each one. Unfortunately many organizations simply don't have the resources to assign people to summarize hundreds or even thousands of documents. Brevity provides you with a solution. Brevity easily generates document summaries for you. The summaries can be as long or as short as you wish. You can also use Brevity to highlight key sentences or words in your document.	http://www.lextek.com/brevity/	Bob	Commercial	1			No				No		4	
Chaos	Text Processing	NLP Tools	Morphological Analysis	CHAOS: A robust syntactic parser for Italian and for English. The system implements a modular and lexicalised approach to the syntactic parsing problem. It is based on the notion of eXtended Dependency Graph (XDG) that has been seen as a useful representation mechanism in a shallow parsing approach. The system offers a collection of modules for designing parsing architectures. The pool of modules consists of:	http://art.uniroma2.it/external/chaosproject/	Bob	Unclear	Italian, English	Java		Yes	upon request	Yes	Not active	No		0	
Chaos	Text Processing	NLP Tools	Parser	CHAOS: A robust syntactic parser for Italian and for English. The system implements a modular and lexicalised approach to the syntactic parsing problem. It is based on the notion of eXtended Dependency Graph (XDG) that has been seen as a useful representation mechanism in a shallow parsing approach. The system offers a collection of modules for designing parsing architectures. The pool of modules consists of:	http://art.uniroma2.it/external/chaosproject/	Bob	Unclear	Italian, English	Java		Yes	upon request	Yes	Not active	No		0	
Chaos	Text Processing	NLP Tools	POS tagger	CHAOS: A robust syntactic parser for Italian and for English. The system implements a modular and lexicalised approach to the syntactic parsing problem. It is based on the notion of eXtended Dependency Graph (XDG) that has been seen as a useful representation mechanism in a shallow parsing approach. The system offers a collection of modules for designing parsing architectures. The pool of modules consists of:	http://art.uniroma2.it/external/chaosproject/	Bob	Unclear	Italian, English	Java		Yes	upon request	Yes	Not active	No		0	
CiceroLite	Text Processing	NLP Tools	NER	Language Computer's CiceroLite recognizes hundreds of different types of named entities in English, Arabic, and Chinese texts with nearly 90% precision and recall. It is available as one of many plug-in NLP components which operate within the Cicero On-Demand server.	http://www.languagecomputer.com/products/text-annotation/cicerolite.html	Bob	Commercial	7	All		Yes	Live demo	No	Only on request	Yes		0	depends if the evaluating libraries use it, more a research tools
CLAWS part-of-speech tagger for English	Text Processing		PoS tagger		http://ucrel.lancs.ac.uk/claws/	Tomasz	http://ucrel.lancs.ac.uk/claws/purc	English			Yes		No		No	Old project	0	
Conjecture	Text Processing	Core Text Recognition	Framework	Conjecture is a modular, extensible, open-source C++ framework for Optical Character Recognition (OCR). Conjecture is not a single OCR, but rather is an extensible collection of OCRs that can be explored, analyzed, compared, extended, modified, and merged within a unified environment.	http://conjecture.sourceforge.net/	Katrien&	GPL		C		Yes		No	version 0.06 was put on sourceforge in 2006; no more activity, links not working (eg. link to mailinglist for support)			0	
Corpus Based Lexicon Tool (CoBaLT)	Text Processing	NLP Tools	Lexicon building	Corpus Based Lexicon Tool (CoBaLT). A tool for corpus-based lexicon construction. Users can upload a text dataset (corpus) for use in creating an attestation-based lexicon. This tool is used to manually correct the automatically lemmatized corpus text. Verified lemmatized words plus the context in which they appear will be stored in the Information Retrieval Lexicon. The tool can handle plain text and various XML formats, among which the IMPACT Page XML format and TEI. An important requirement of the tool is that it should be fit to quickly process large quantities of data, that it is a web application that can be run from any computer in the local network, that frequent input actions can be performed with the keyboard, and that the information is presented in such a way that quick evaluation is possible.	http://www.digitisation.eu/tools/toolbox-for-lexicon-building/corpus-based-lexicon-tool-cobalt/	Bob	ASL 2.0				Yes		Yes	IMPACT team members involved	Yes	Yes. Part of IMPACT	4	

Succeed List of Tools

Name of the tool	Group	Type	Subtype	Description	Link to the tool/website	Entry author	Type of license	Language support	Tech. context	Time and effort for installation	1st criterion for exclusion: If there is no trial version available to test a tool within Succeed, the tool will be discarded		2nd criterion for exclusion: If there is no technical documentation or support available, the tool will be discarded		3rd criterion for exclusion: If there is no information about the tool being used in other projects, no information about existing benchmarks or no information from users about the tool, the tool will be discarded		Relevance for libraries: A rating from 0-5 indicates how relevant a tool is for libraries (0 = not relevant; 5 = very relevant). The description provides further insights on how the consortium came to this rating.	
											Trial version available (Yes/No)	Further Description	Documentation/S available? (Yes/No)	Further Description (e.g. available support, activity status/last update)	Information assuring tool performance available? (Yes/No)	Further Description (Applicability to mass digitisation, quality and robustness)	Rating (0-5)	Further Description
cue.language	Text Processing		NLP toolset and resources	cue.language is a small library of Java code and resources that provides the following basic natural-language processing capabilities	https://github.com/jdf/cue.language	Tomasz	Apache License	Arabic, Catalan, Croatian, Czech, Dutch, Danish, English, Esperanto, Farsi, Finnish, French, German, Greek, Hebrew, Hindi, Hungarian, Italian, Latin, Norwegian, Polish, Portuguese, Romanian, Russian, Slovenian, Slovak, Spanish, Swedish, Turkish			Yes		Yes	To some extent	No	Last updated 2 years ago	0	
Dates Recognizer	Text Processing		dates recognizer	Accepts a string thought to contain a date (or a date range, or a period) and parses it, returning a date range.	TBD	Tomasz	free service				No	Will be available	Yes		No	New tool	0	
DBPedia spotlight	Text Processing	NLP Tools	NE linking	DBpedia Spotlight is a tool for automatically annotating mentions of DBpedia resources in text, providing a solution for linking unstructured information sources to the Linked Open Data cloud through DBpedia. DBpedia Spotlight recognizes that names of concepts or entities have been mentioned (e.g. "Michael Jordan"), and subsequently matches these names to unique identifiers (e.g. dbpedia: Michael_I._Jordan, the machine learning professor or dbpedia:Michael_Jordan the basketball player). It can also be used for building your solution for Named Entity Recognition, Keyphrase Extraction, Tagging, etc. amongst other information extraction tasks.	https://github.com/dbpedia-spotlight/dbpedia-spotlight/wiki	Bob	Free		Java / Python		Yes		Yes		Yes		4	Might be relevant to try out tools, not for productive environments
Expervision WebOCR	Text Processing	Core Text Recognition	Web service	In 1999, Expervision released WebOCR (Online OCR) 1.0, providing her users with flexible and easy modes of OCR application. WebOCR (OnlineOCR) 2.0 updated later is able to provide 4 kinds of Web OCR (Online OCR) application modes based on different business environment and processing requirements of her users.	http://www.expervision.com/ocr-software/webocr-onlineocr	Katrien&	Own license		?		Yes		Yes	commercial	No		0	
FM-SBLEX	Text Processing	NLP Tools	Morphological Analysis	FM-SBLEX consists of three computational morphology tools for modern Swedish (SALDO), for 19th century Swedish (Dalin), and for Old Swedish. FM-SBLEX has been developed using the Functional Morphology library.	http://spraakbanken.gu.se/eng/research/swefn/fm-sblex	Bob	GPL3	1	Linux / Unix		Yes		Yes	Active development	No		0	might be interesting, already used in some polish libraries
FreeLing	Text Processing	NLP Tools	Lemmatizer	This module is somehow different of the other modules, since it doesn't enrich the given text. It compares the given text with available models for different languages, and returns the most likely language the text is written in. It can be used as a preprocess to determine which data files are to be used to analyze the text.	http://nlp.lsi.upc.edu/freeling/doc/userman/html/node18.html	Bob	GPL	Any			Yes		Yes		yes		4	
Freeling	Text processing	NLP Tools	NLP toolset and resources	FreeLing is a library providing language analysis services, oriented to satisfy the needs of Natural Language Processing. FreeLing is designed to be used as an external library from any application requiring this kind of services. Nevertheless, a simple main program is also provided as a basic interface to the library, which enables the user to analyze text files from the command line. Actually, many users do not develop on FreeLing, but use it as a text processing tool.	http://nlp.lsi.upc.edu/freeling/	Bob	GPL	Any		1d	Yes		Yes		yes		4	
Freeling	Text processing	NLP Tools	Tokenizer	Tokenization rules are regular expressions that are matched against the beggining of the text line being processed. The first matching rule is used to extract the token, the matching substring is deleted from the line, and the process is repeated until the line is empty.	http://nlp.lsi.upc.edu/freeling/doc/userman/html/node20.html	Bob	GPL		Linux	1h	Yes		Yes		yes		4	
FreeLing	Text Processing	NLP Tools	Language Identification	It compares the given text with available models for different languages, and returns the most likely language the text is written in. It can be used as a preprocess to determine which data files are to be used to analyze the text.	http://nlp.lsi.upc.edu/freeling/index.php	Bob	GPL	Asturian, Catalan, English, Galician, Italian, Portuguese, Russian, Spanish, Welsh, expandable to any language	Linux	1h-1d	Yes		Yes	Active user group	Yes		4	
FreeLing	Text Processing	NLP Tools	NER	There are two different modules able to perform NE recognition. They can be instantiated directly, or via a wrapper that will create the right module depending on the configuration file.	http://nlp.lsi.upc.edu/freeling/index.php	Bob	GPL	Asturian, Catalan, English, Galician, Italian, Portuguese, Russian, Spanish, Welsh	Linux	1h-1d	Yes		Yes	Active user group	Yes		4	useful for format conversion, but you have to develop your own plugins for conversion
FreeLing	Text Processing	NLP Tools	POS Tagger	There are two different modules able to perform PoS tagging. The application should decide which method is to be used, and instantiate the right class.The first PoS tagger is the hmm_tagger class, which is a classical trigam Markovian tagger, following [#! brants00!#].The second module, named relax_tagger, is a hybrid system capable to integrate statistical and hand-coded knowledge, following [#!padro98a!#].	http://nlp.lsi.upc.edu/freeling/index.php	Bob	GPL	Asturian, Catalan, English, Galician, Italian, Portuguese, Russian, Spanish, Welsh	Linux	1h-1d	Yes		Yes	Active user group	Yes		4	
FreeLing	Text Processing	NLP Tools	Morphological Analysis	The morphological analyzer is a meta-module which does not perform any processing of its own.It is just a convenience module to simplify the instantiation and call. At instantiation time, it receives a maco_options object, containing information about which submodules have to be created and which files have to be used to create them. to the submodules described in the next sections (from [*] to [*]).At instantiation time, it receives a maco_options object, containing information about which submodules have to be created and which files have to be used to create them.	http://nlp.lsi.upc.edu/freeling/index.php	Bob	GPL	Asturian, Catalan, English, Galician, Italian, Portuguese, Russian, Spanish, Welsh	Linux	1h-1d	Yes		Yes	Active user group	Yes		0	probably not relevant for libraries

Succeed List of Tools

Name of the tool	Group	Type	Subtype	Description	Link to the tool/website	Entry author	Type of license	Language support	Tech. context	Time and effort for installation	1st criterion for exclusion: If there is no trial version available to test a tool within Succeed, the tool will be discarded		2nd criterion for exclusion: If there is no technical documentation or support available, the tool will be discarded		3rd criterion for exclusion: If there is no information about the tool being used in other projects, no information about existing benchmarks or no information from users about the tool, the tool will be discarded		Relevance for libraries: A rating from 0-5 indicates how relevant a tool is for libraries (0 = not relevant; 5 = very relevant). The description provides further insights on how the consortium came to this rating.	
											Trial version available (Yes/No)	Further Description	Documentation/S available? (Yes/No)	Further Description (e.g. available support, activity status/last update)	Information assuring tool performance available? (Yes/No)	Further Description (Applicability to mass digitisation, quality and robustness)	Rating (0-5)	Further Description
FreeLing	Text Processing	NLP Tools	Parser	The dependency parser works in three stages:At the first stage, the <GRPAR> rules are used to complete the shallow parsing produced by the chart into a complete parsing tree. The rules are applied to a pair of adjacent chunks. At each step, the selected pair is fused in a single chunk. The process stops when only one chunk remains. The next step is an automatic conversion of the complete parse tree to a dependency tree. Since the parsing grammar encodes information about the head of each rule, the conversion is straighforward. The last step is the labeling. Each edge in the dependeny tree is labeled with a syntactic function, using the <GRLAB> rules	http://nlp.lsi.upc.edu/freeling/index.php	Bob	GPL	Asturian, Catalan, English, Galician, Italian, Portuguese, Russian, Spanish, Welsh	Linux	1h-1d	Yes		Yes	Active user group	Yes		0	
Frog	Text Processing	NLP Tools	Parser	Frog, formerly known as Tadpole, is an integration of memory-based natural language processing (NLP) modules developed for Dutch. All NLP modules are based on Timbl, the Tilburg memory-based learning software package. Most modules were created in the 1990s at the ILK Research Group (Tilburg University, the Netherlands) and the CLiPS Research Centre (University of Antwerp, Belgium). Over the years they have been integrated into a single text processing tool. More recently, a dependency parser, a base phrase chunker, and a named-entity recognizer module were added.	http://ilk.uvt.nl/frog/	Bob	GPL	Dutch			Yes		Yes	Active development	Yes		0	
GATE	Text Processing			open source software capable of solving almost any text processing problem	https://gate.ac.uk/	Tomasz	free				Yes		Yes		Yes	Mailing list active.	0	
graph-based dependency parser	Text Processing	NLP Tools	Parser	Bernd Bohnet. 2010. Top Accuracy and Fast Dependency Parsing is not a Contradiction. The 23rd International Conference on Computational Linguistics (COLING 2010), Beijing, China.	http://code.google.com/p/mate-tools/	Bob	GPL	English, German, Chinese	Java		Yes		Yes		Yes		0	
Impact Tools	Text Processing	NLP Tools	Lemmatization	IMPACT provides tools for: 1. Reducing historical word forms to one or several possible modern lemma's (lemmatization) 2. Expanding lemma lists with part of speech information to possible ("hypothetical") full forms.	http://www.digitisation.eu/tools/toolbox-for-lexicon-building/tools-for-lemmatization-and-reverse-lemmatization/	Bob	ASL 2.0				Yes		Yes	IMPACT team members involved	Yes	Yes. Part of IMPACT	4	
Impact Tools	Text Processing	NLP Tools	Spelling variations	The spelling of words in historical texts can differ widely from modern spelling. There are two general approaches to match different spellings. First, it is possible to use rewrite rules that transform words in one spelling to another. For historical dictionary which covers a large timespan, and in which variation is not limited to orthography, this approach is not satisfactory. Therefore, the use of statistics is often needed.	http://www.digitisation.eu/tools/toolbox-for-lexicon-building/spelling-variation-tool/	Bob	ASL 2.0		Java		Yes		Yes	IMPACT team members involved	Yes	Yes. Part of IMPACT	4	
IOBBER (chunker)	Text Processing		chunker	IOBBER is a chunker for Polish. Its job is to recognise syntactic phrases (chunks) in Polish text. The name comes from IOB tags that are assigned to tokens to represent chunks (strictly speaking, we use IOB2 representation). Here is an example sentence annotated with NP and VP chunks: * [Dziennikarka]NP [zarzucała]VP [Rutkowskiemu] NP [to]NP, że [całe jego działanie ws. zaginięcia]NP [to]VP [„show"]NP IOBBER is a reimplementation of CRF++ chunker available in Disaster.	http://nlp.pwr.wroc.pl/redmine/projects/iobber/wiki	Tomasz	unknown	Polish			Yes		Yes		Yes		0	
JGAAP	Text Processing		authorship attribution	authorship attibution software	http://evllabs.com/jgaap/w/index.php/Main_Page	Tomasz	GPL?				Yes		Yes		Yes	Not sure - last version from October 2012	0	
LemmaGen	Text Processing	NLP Tools	Stemmer/Lemm	LemmaGen project aims at providing standardized open source multilingual platform for lemmatisation. We started this work as a result of lack of high quality lemmatiser for Slovene language. Currently we have, not only the lemmatiser for Slovene, but also for 11 other European languages and the system which is able to learn lemmatisation rules for new languages by providing it with existing wordform-lemma pair examples.	http://lemmatise.ijs.si/	Bob	free, open source	Slovene,11 more			Yes		No	Probably not	No		0	
Lextek	Text Processing	NLP Tools	Language Identification	For many applications, it is important to be able to correctly identify the language that a document or piece of text is written in. The Lextek Language Identifier enables you to do this. Since some languages may be written in several character encodings, the Lextek Language Identifier will automatically identify what character encoding the text was written in. Supporting approximately 260 different languages and character encodings, the Lextek Language Identifier gives you the ability to automatically recognize more languages and encodings than any other language identifier available. We are adding more languages all the time and work closely with our customers to ensure that their language recognition needs are fully supported.	http://www.lextek.com/langid/	Bob	commercial	260			No				Yes		0	

Succeed List of Tools

Name of the tool	Group	Type	Subtype	Description	Link to the tool/website	Entry author	Type of license	Language support	Tech. context	Time and effort for installation	1st criterion for exclusion: If there is no trial version available to test a tool within Succeed, the tool will be discarded		2nd criterion for exclusion: If there is no technical documentation or support available, the tool will be discarded		3rd criterion for exclusion: If there is no information about the tool being used in other projects, no information about existing benchmarks or no information from users about the tool, the tool will be discarded		Relevance for libraries: A rating from 0-5 indicates how relevant a tool is for libraries (0 = not relevant; 5 = very relevant). The description provides further insights on how the consortium came to this rating.	
											Trial version available (Yes/No)	Further Description	Documentation/S available? (Yes/No)	Further Description (e.g. available support, activity status/last update)	Information assuring tool performance available? (Yes/No)	Further Description (Applicability to mass digitisation, quality and robustness)	Rating (0-5)	Further Description
Liner2 (NER)	Text Processing		NER	Liner2 is a customizable and open-source framework for proper names recognition. The framework consists of several universal methods for sequence chunking which include: dictionary look-up, pattern matching and statistical processing.The statistical processing is performed using Conditional Random Fields and a rich set of features including morphological, lexical and semantic information. We present an application of the framework to the task of recognition proper names in Polish texts (5 common categories of proper names, i.e. first names, surnames, city names, road names and country names) and an extended model to recognize 56 categories of proper names which was used to bootstrap the manual annotation of KPWr corpus.	http://nlp.pwr.wroc.pl/inforex/index.php?page=ner	Tomasz	unknown	Polish			Yes		Yes		Yes		0	
LingPipe	Text processing	NLP Tools	Language Identification	LingPipe's text classifiers learn by example. For each language being classified, a sample of text is used as training data. LingPipe learns the distribution of characters per language using character language models. Character language models provide state-of-the-art accuracy for text classification. Character-level models are particularly well-suited to language ID because they do not require tokenized input; tokenizers are often language-specific.	http://alias-i.com/lingpipe/demos/tutorial/langid/read-me.html	Bob	Free	Any	java	1h	Yes		Yes		yes		4	
LingPipe	Text Processing	NLP Tools	NER	LingPipe is tool kit for processing text using computational linguistics. LingPipe is used to do tasks like: Find the names of people, organizations or locations in news, Automatically classify Twitter search results into categories, Suggest correct spellings of queries	http://alias-i.com/lingpipe/	Bob	Limited version free, production version at a fee	all, in principle	Java		Yes		Yes	Active development / large user group	Yes		4	
LingPipe	Text processing	NLP Tools	NLP toolset and resources	LingPipe is tool kit for processing text using computational linguistics.	http://alias-i.com/lingpipe/	Bob	Free/Comme		java	1d	Yes		Yes		yes		4	
LingPipe	Text Processing	NLP Tools	Tokenizer	Part-of-speech tagging is a process whereby tokens are sequentially labeled with syntactic labels, such as "finite verb" or "gerund" or "subordinating conjunction". This tutorial shows how to train a part-of-speech tagger and compile its model to a file, how to load a compiled model from a file and perform part-of-speech tagging, and finally, how to evaluate and tune models.	http://alias-i.com/lingpipe/demos/tutorial/posTags/read-me.html	Bob	unknown	Any			Yes		Yes		yes		4	
Link Grammar Parser	Text Processing	NLP Tools	Parser	The Link Grammar Parser is a syntactic parser of English, based on link grammar, an original theory of English syntax. Given a sentence, the system assigns to it a syntactic structure, which consists of a set of labeled links connecting pairs of words. The parser also produces a "constituent" representation of a sentence (showing noun phrases, verb phrases, etc.).	http://www.link.cs.cmu.edu/link/	Bob	GPL	English	C		Yes		No	Unclear	No		0	
LX-Parser	Text Processing	NLP Tools	Parser	LX-Parser is a statistical constituency parser for Portuguese. It performs a syntactic analysis of Portuguese sentences in terms of their constituency structure.	http://lxcenter.di.fc.ul.pt/tools/en/LXParserEN.html	Bob	Free	Portuguese			Yes		Yes	no active development	No		0	
LX-Tagger	Text Processing	NLP Tools	POS tagger	Lx-Tagger is a part-of-speech tagger for Portuguese that assigns a single morpho-syntactic tag, from the tagset below, to every token	http://lxcenter.di.fc.ul.pt/tools/en/LXTaggerEN.html	Bob	Proprietary	1	Perl	5m	Yes		No	No active development	No		0	
MALLET	Text Processing		NLP toolset and resources	MALLET is a Java-based package for statistical natural language processing, document classification, clustering, topic modeling, information extraction, and other machine learning applications to text.	http://mallet.cs.umass.edu/index.php	Tomasz	CPL		Java		Yes		Yes		Yes	Mailing list active.	0	
MaltParser	Text Processing	NLP Tools	Parser	MaltParser is a system for data-driven dependency parsing, which can be used to induce a parsing model from treebank data and to parse new data using an induced model. MaltParser is developed by Johan Hall, Jens Nilsson and Joakim Nivre at Växjö University and Uppsala University, Sweden.	http://www.maltparser.org/	Bob	http://www.maltparser.org/license.html	English, French, Swedish, Spanish	Java		Yes		Yes	Active development	Yes	MaltParser 0.4 was used in the multi-lingual track of the CoNLL 2007 Shared Task in the systems that obtained the first and fifth best overall scores.	0	
MBT – Memory-Based Tagger-Generator	Text Processing	NLP Tools	POS tagger	MBT is a memory-based tagger-generator and tagger in one. The tagger-generator part can generate a sequence tagger on the basis of a training set of tagged sequences; the tagger part can tag new sequences. MBT can, for instance, be used to generate part-of-speech taggers or chunkers for natural language processing. It has also been used for named-entity recognition, information extraction in domain-specific texts, and disfluency chunking in transcribed speech.	http://ilk.uvt.nl/mbt/	Bob	GNU3	2	Linux / Unix	1d	Yes		Yes	Active development	No		4	
Minipar	Text Processing	NLP Tools	Parser	MINIPAR is a broad-coverage parser for the English language. An evaluation with the SUSANNE corpus shows that MINIPAR achieves about 88% precision and 80% recall with respect to dependency relationships. MINIPAR is very efficient, on a Pentium II 300 with 128MB memory, it parses about 300 words per second.	http://webdocs.cs.ualberta.ca/~lindek/minipar.htm	Bob	Unclear	English	Windows, Linux		Yes		No	Unclear	No		0	
MontyChunker	Text Processing	NLP Tools	Chunker	Lightning fast regular expression chunker	http://web.media.mit.edu/~hugo/montylingua/index.html	Bob	Free for non-commercial use	1	Java / Python		Yes		Yes	Development not active	No		0	
MontyLemmatiser	Text Processing	NLP Tools	Stemmer/Lemm	Strips inflectional morphology, i.e. changes verbs to infinitive form and nouns to singular form	http://web.media.mit.edu/~hugo/montylingua/index.html	Bob	Free for non-commercial use		Java / Python		Yes		Yes	No active development	No		0	
MontyTagger	Text Processing	NLP Tools	POS tagger	Part-of-speech tagging based on Brill94, enriched with common sense	http://web.media.mit.edu/~hugo/montylingua/index.html	Bob	Free for non-commercial use	1	Java / Python	1d	Yes		No	No active development	No		0	
MontyTokenizer	Text Processing	NLP Tools	Tokenizer	Tokenizes raw English text (sensitive to abbreviations), and resolve contractions, e.g. "you're" ==> "you are"	http://web.media.mit.edu/~hugo/montylingua/index.html	Bob	Free for non-commercial use		Java / Python		Yes		Yes	No active development	No		0	

Succeed List of Tools

Name of the tool	Group	Type	Subtype	Description	Link to the tool/website	Entry author	Type of license	Language support	Tech. context	Time and effort for installation	1st criterion for exclusion: If there is no trial version available to test a tool within Succeed, the tool will be discarded		2nd criterion for exclusion: If there is no technical documentation or support available, the tool will be discarded		3rd criterion for exclusion: If there is no information about the tool being used in other projects, no information about existing benchmarks or no information from users about the tool, the tool will be discarded		Relevance for libraries: A rating from 0-5 indicates how relevant a tool is for libraries (0 = not relevant; 5 = very relevant). The description provides further insights on how the consortium came to this rating.	
											Trial version available (Yes/No)	Further Description	Documentation/S available? (Yes/No)	Further Description (e.g. available support, activity status/last update)	Information assuring tool performance available? (Yes/No)	Further Description (Applicability to mass digitisation, quality and robustness)	Rating (0-5)	Further Description
morphadorner	Text Processing			MorphAdorner is a Java command-line program which acts as a pipeline manager for processes performing morphological adornment of words in a text. Language recognition, lemmatizer, lexicon lookup, etc.	http://morphadorner.northwestern.edu/	Tomasz	http: //morphadorn	English	Java		Yes		Yes		No	Not sure if it is still used - last activity in 2009	0	
Morphette	Text Processing	NLP Tools	Morphological Analysis	Morfette is a tool for supervised learning of inflectional morphology. Given a corpus of sentences annotated with lemmas and morphological labels, and optionally a lexicon, morfette learns how to morphologically analyse new sentences.In the learning stage Morfette fits two separate logistic regression models: one for morphological tagging and one for lemmatization. The predictions of the models are combined dynamically and produce a globally plausible sequence of morphological-tag - lemma pairs for a sentence.	https://sites.google.com/site/morfetteweb/	Bob	Unclear				Yes		No	Unclear	No		0	equivalent to imagemagick/graphicsma
Morphette	Text Processing	NLP Tools	Stemmer/Lemm	In Morfette lemmatization is cast as a classification task where a a lemmatization class corresponds to the specification of the edit operations which are needed to transform the inflected word form into the corresponding lemma. The basic approach is described in (Chrupala et al 2008 and Chrupala 2008). The current version of Morfette uses an averaged perceptron to fit the models, rather than Maximum Entropy training. The lemmatization classes are Edit-Tree-based as described in (Chrupala 2008).	https://sites.google.com/site/morfetteweb/	Bob	Unclear				Yes		No	Unclear	No		0	
NERT	Text Processing	NLP Tools	NER	NERT is a tool that can mark and extract named entities (persons, locations and organizations) from a text file. It uses a supervised learning technique, which means it has to be trained with a manually tagged training file before it is applied to other text. In addition, version 2.0 of the tool and higher also comes with a named entity matcher module, with which it is possible to group variants or to assign modern word forms of named entities to old spelling variants. As a basis for the tool in this package, the named entity re cognizer from Stanford University is used. This tool has been extended for use in IMPACT. Among the extensions is the aforementioned matcher module, and a module that reduces spelling variation within the used data, thus leading to improved performance.	https://www.impact-project. eu/uploads/media/IMPACT_D-EE2. 6_NERT_User_Manual.pdf	Bob	GPLv2		Java		Yes		Yes	IMPACT team members involved	Yes	Yes. Part of IMPACT	4	
NLTK	Text processing	NLP Tools	Tokenizer	Tokenizers divide strings into lists of substrings. For example, tokenizers can be used to find the list of sentences or words in a string.		Bob	Free		Python	1h	Yes		Yes		yes		4	
NLTK	Text Processing	NLP Tools	NER		http://nltk.org/api/nltk.chunk.html#module-nltk.chunk.named_entity	Bob	Free	Any	Python	1d	Yes		Yes		Yes		4	
NLTK	Text Processing		NLP toolset and resources	NLTK is a leading platform for building Python programs to work with human language data. It provides easy-to-use interfaces to over 50 corpora and lexical resources such as WordNet, along with a suite of text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning.	http://nltk.org/	Tomasz	Apache License				Yes		Yes		Yes	Latest version from February 2013	1	
NLTK Classify Package	Text Processing	NLP Tools	Topic Modelling	Classes and interfaces for labeling tokens with category labels (or “class labels”). Typically, labels are represented with strings (such as 'health' or 'sports'). Classifiers can be used to perform a wide range of classification tasks. For example, classifiers can be used...	http://nltk.org/api/nltk.classify.html#module-nltk.classify	Bob	free, open source		Python	1h	Yes		Yes	active development team, large user group	Yes		4	
NLTK Parsers	Text Processing	NLP Tools	Parser	Classes and interfaces for producing tree structures that represent the internal organization of a text. This task is known as “parsing” the text, and the resulting tree structures are called the text’s “parses”. Typically, the text is a single sentence, and the tree structure represents the syntactic structure of the sentence. However, parsers can also be used in other domains. For example, parsers can be used to derive the morphological structure of the morphemes that make up a word, or to derive the discourse structure for a set of utterances.	http://nltk.org/api/nltk.parse.html#module-nltk.parse	Bob	free, open source		Python	1h	Yes		Yes	active development team, large user group	Yes		0	
NLTK Stemmers	Text Processing	NLP Tools	Stemmer/Lemm	Interfaces used to remove morphological affixes from words, leaving only the word stem. Stemming algorithms aim to remove those affixes required for eg. grammatical role, tense, derivational morphology leaving only the stem of the word. This is a difficult problem due to irregular words (eg. common verbs in English), complicated morphological rules, and part-of-speech and sense ambiguities (eg. ceil- is not the stem of ceiling).	http://nltk.org/api/nltk.stem.html#module-nltk.stem	Bob	free, open source		Python	1h	Yes		Yes	active development team, large user group	Yes		4	
NLTK Taggers	Text Processing	NLP Tools	POS Tagger	This package defines several taggers, which take a token list (typically a sentence), assign a tag to each token, and return the resulting list of tagged tokens. Most of the taggers are built automatically based on a training corpus.	http://nltk.org/api/nltk.tag.html#module-nltk.tag	Bob	free, open source		Python	1h	Yes		Yes	active development team, large user group	Yes		4	

Succeed List of Tools

Name of the tool	Group	Type	Subtype	Description	Link to the tool/website	Entry author	Type of license	Language support	Tech. context	Time and effort for installation	1st criterion for exclusion: If there is no trial version available to test a tool within Succeed, the tool will be discarded: Trial version available (Yes/No)	Further Description	2nd criterion for exclusion: If there is no technical documentation or support available, the tool will be discarded: Documentation/S available? (Yes/No)	Further Description (e.g. available support, activity status/last update)	3rd criterion for exclusion: If there is no information about the tool being used in other projects, no information about existing benchmarks or no information from users about the tool, the tool will be discarded: Information assuring tool performance available? (Yes/No)	Further Description (Applicability to mass digitisation, quality and robustness)	Relevance for libraries: A rating from 0-5 indicates how relevant a tool is for libraries (0 = not relevant; 5 = very relevant). The description provides further insights on how the consortium came to this rating.: Rating (0-5)	Further Description
OCRopodium	Text Processing	Core Text Recognition	Framework	As part of the Ocropodium project at KCL's Centre for e-Research we're investigating OCR workflows for digitising historical collections. In the course of experimenting with Ocropus, Tesseract, and other software, we've developed some tools and utilities that might be of interest to others. Currently there's a Django web application for performing batch OCR, a Qt GUI for correcting ground-truth transcripts from Ocropus bookstores, and a viewer for previewing its page segmentation results:	https://code.google.com/p/ocropodium/	Katrien&	ASL 2.0		Python		Yes		No	project finished end of July of 2011; documentation available, some bugfixing done, but not active	No	Wiki-info: The OCRopodium WEB APP is an experimental front-end for OCR'ing using OCRopus and Tesseract. It is extremely buggy, incomplete, and obviously comes with no warranty of any kind, / The Segviewer is a very simple GUI for visualising the results of Ocropus's page segmentation components. It allows you to open a PNG file and compare the segmentation results as line and paragraph highlights. It was intended that component parameters would be accessible via the GUI (and not just the environment) but this has not yet been implemented. / TRANSCRIPT HELPER: This is a little GUI utility for generating and corrected ground-truth transcripts for use with Ocropus's "trainseg" mode. Originally written for testing Ocropus on handwritten material, it also allows you to manually segment a line of text into it's component characters and export a ".cseg.png" image.	0	
Pantera	Text Processing	NLP Tools	POS Tagger	The PANTERA is a Brill Tagger for morphologically rich languages, eg. Polish.	http://zil.ipipan.waw.pl/PANTERA	Bob	GPL	Polish			Yes		No	Unclear	No		0	
Polyglot 3000	Text Processing	NLP Tools	Language Identification	Polyglot 3000 is an automatic language identifier that quickly recognizes the language of any text, phrase or even single words. It is available for Windows 95/98/NT/ME/2000/XP/2003/Vista/2008/7/8.	http://www.polyglot3000.com/	Bob	unknown	More than 400	Win.	5m	Yes		No	No documentation online. Contact with sales possible.	No		0	
Rosette	Text Processing	NLP Tools	Language Identification	Automatically Detects the Language of Any Digital Text. Rosette® Language Identifier analyzes text, identifying the language and the character encoding scheme. Detecting the language of documents is a critical first step in any process that handles multilingual text. Our software recognizes 55 languages and 45 encodings and processes files extremely quickly and accurately.	http://www.basistech.com/language-identifier/	Bob	commercial	55			Yes	After request, no reply	Yes		Yes		4	
Rosette Base Linguistics	Text processing	NLP Tools	Lemmatization	Sophisticated morphological analysis, segmentation, and tagging of Arabic, Asian, and European language text	http://www.basistech.com/base-linguistics/	Bob	Commercial				Yes		Yes		yes		4	
Rosette Base Linguistics	Text processing	NLP Tools	POS tagger	Sophisticated morphological analysis, segmentation, and tagging of Arabic, Asian, and European language text	http://www.basistech.com/base-linguistics/	Bob	Commercial				Yes		Yes		yes		4	
Rosette Base Linguistics	Text processing	NLP Tools	Tokenizer	Sophisticated morphological analysis, segmentation, and tagging of Arabic, Asian, and European language text	http://www.basistech.com/base-linguistics/	Bob	Commercial	40			Yes		Yes		yes		4	
Rosette Entity Extractor (REX)	Text processing	NLP Tools	NER	Identify Names, Places, Organizations, and Other Entities in Your Text	http://www.basistech.com/entity-extractor/	Bob	Commercial	17			Yes		Yes		yes		4	
Rosette Linguistic Platform	Text processing	NLP Tools	Language Identification	Rosette® Language Identifier analyzes text, identifying the language and the character encoding scheme. Detecting the language of documents is a critical first step in any process that handles multilingual text. Our software recognizes 55 languages and 45 encodings and processes files extremely quickly and accurately.	http://www.basistech.com/language-identifier/	Bob	Commercial	55			Yes	Upon request	Yes		yes		4	
Rosette Linguistic Platform	Text processing	NLP Tools	NLP toolset and resources	Comprehensive linguistic analysis of unstructured text in Asian, European and Middle Eastern languages for enhancing information retrieval, text mining, and other applications.	http://www.basistech.com/products/	Bob	Commercial			1d	Yes	Upon request	Yes		yes		4	
Stanford coreNLP	Text Processing		NLP toolset and resources	Stanford CoreNLP provides a set of natural language analysis tools which can take raw English language text input and give the base forms of words, their parts of speech, whether they are names of companies, people, etc., normalize dates, times, and numeric quantities, and mark up the structure of sentences in terms of phrases and word dependencies, and indicate which noun phrases refer to the same entities.	http://nlp.stanford.edu/software/corenlp.shtml	Tomasz	GPL v2	English			Yes		Yes		Yes		0	
Stanford Log-linear Part-Of-Speech Tagger	Text processing	NLP Tools	POS tagger	A Part-Of-Speech Tagger (POS Tagger) is a piece of software that reads text in some language and assigns parts of speech to each word (and other token), such as noun, verb, adjective, etc., although generally computational applications use more fine-grained POS tags like 'noun-plural'. This software is a Java implementation of the log-linear part-of-speech taggers described in these papers (if citing just one paper, cite the 2003 one):	http://nlp.stanford.edu/software/tagger.shtml	Bob	GPL2	Any	java	1d	Yes		Yes		yes		4	

Succeed List of Tools

Name of the tool	Group	Type	Subtype	Description	Link to the tool/website	Entry author	Type of license	Language support	Tech. context	Time and effort for installation	1st criterion for exclusion: If there is no trial version available to test a tool within Succeed, the tool will be discarded: Trial version available (Yes/No)	Further Description	2nd criterion for exclusion: If there is no technical documentation or support available, the tool will be discarded: Documentation/S available? (Yes/No)	Further Description (e.g. available support, activity status/last update)	3rd criterion for exclusion: If there is no information about the tool being used in other projects, no information about existing benchmarks or no information from users about the tool, the tool will be discarded: Information assuring tool performance available? (Yes/No)	Further Description (Applicability to mass digitisation, quality and robustness)	Relevance for libraries: A rating from 0-5 indicates how relevant a tool is for libraries (0 = not relevant; 5 = very relevant). The description provides further insights on how the consortium came to this rating.: Rating (0-5)	Further Description
Stanford NER	Text processing	NLP Tools	NER	Stanford NER (also known as CRFClassifier) is a Java implementation of a Named Entity Recognizer. Named Entity Recognition (NER) labels sequences of words in a text which are the names of things, such as person and company names, or gene and protein names. The software provides a general (arbitrary order) implementation of linear chain Conditional Random Field (CRF) sequence models, coupled with well-engineered feature extractors for Named Entity Recognition. (CRF models were pioneered by Lafferty, McCallum, and Pereira (2001); see Sutton and McCallum (2006) for a better introduction.) Included with the download are good 3 class (PERSON, ORGANIZATION, LOCATION) named entity recognizers for English (in versions with and without additional distributional similarity features) and another pair of models trained on the CoNLL 2003 English training data. The distributional similarity features improve performance but the models require considerably more memory.	http://nlp.stanford.edu/software/CRF-NER.shtml	Bob	Free	Any	java	1d	Yes		Yes		Yes	Tested and compared as part of the Impact project	4	
Synapse	Text Processing	NLP Tools	NER	Competitive intelligence always concerns organizations, people, places, products, etc. This technology aims at tagging information in a text flow. The information automatically annotated is basically: Person's name, functions, organizations, dates, events, places, addresses, phone numbers, e-mail addresses and amounts.The technology is accurate for all types of texts, whatever the field. Whether legal or military posts, journalistic dispatches on terrorist acts or on economics news, it identifies the actors, their functions and relationships, as well as details of the events encountered. User can integrate its own dictionaries in the technology.	http://www.quaero.org/developpements-technologiques/	Bob	Commercial	1?	web service		No				No		0	
TexLexAn	Text Processing	NLP Tools	summerizer	TexLexAn is the project of an automatic text analyzer, classifier and summarizer. This software is at the frontier of the artificial intelligence and of the machine learning, and participates at its very modest level to the development of the softwares of the future. I take a lot of fun to develop it, I hope you will enjoy to try it.	http://texlexan.sourceforge.net/	Bob	Unclear	English, French, German, Italian, Spanish	C, Python		Yes		No	Stale project	No		0	
TexLexAn	Text Processing	NLP Tools	Text Classification	TexLexAn is the project of an automatic text analyzer, classifier and summarizer. This software is at the frontier of the artificial intelligence and of the machine learning, and participates at its very modest level to the development of the softwares of the future. I take a lot of fun to develop it, I hope you will enjoy to try it.	http://texlexan.sourceforge.net/	Bob	Unclear	English, French, German, Italian, Spanish	C, Python		Yes		No	Unclear	No		0	
TextCat	Text Processing	NLP Tools	Language Identification	TextCat is an implementation of the text categorization algorithm presented in Cavnar, W. B. and J. M. Trenkle, ``N-Gram-Based Text Categorization" In Proceedings of Third Annual Symposium on Document Analysis and Information Retrieval, Las Vegas, NV, UNLV Publications/Reprographics, pp. 161-175, 11-13 April 1994.	http://www.let.rug.nl/vannoord/TextCat/index.html	Bob	free	69	Perl	5m	Yes		No	No longer maintained	No		0	
TextCat	Text Processing	NLP Tools	Text Classification	TextCat is an implementation of the text categorization algorithm presented in Cavnar, W. B. and J. M. Trenkle, ``N-Gram-Based Text Categorization" In Proceedings of Third Annual Symposium on Document Analysis and Information Retrieval, Las Vegas, NV, UNLV Publications/Reprographics, pp. 161-175, 11-13 April 1994.	http://odur.let.rug.nl/~vannoord/TextCat/	Bob	Free		Perl script		Yes		No	not longer supported	No		0	
The Berkeley	Text Processing	NLP Tools	Parser	Parsing is the task of analyzing the grammatical structure of natural language. Given a sequence of words, a parser forms units like subject, verb, object and determines the relations between these units according to some grammar formalism. Our work has focused on learning probabilistic context-free grammars (PCFGs) which assign a sequence of words the most likely parse tree. The parser supports a variety of languages and achieves state-of-the-art performance on most of them. For additional information and related projects visit the Berkeley NLP website.	http://code.google.com/p/berkeleyparser/	Bob	GPL	3, others depending on availability of tree banks	Java		Yes		Yes	No active development	No		0	
The Oslo-Bergen Tagger	Text Processing	NLP Tools	Stemmer/Lemm	The Oslo-Bergen tagger is a robust morphological and syntactic tagger developed at the University of Oslo and at Uni Computing in Bergen over several years. The tagger consists of three main modules: a preprocessor with multitagger and compound analyser, a grammar module for morphological and syntactic disambiguation (Constraint Grammar) and a statistical module that removes the last of the remaining morphological ambiguity (only for Bokmål). The Constraint Grammar module uses a compiler developed at the University of Southern Denmark in Odense. The multitagger uses the lexicon Norsk ordbank.	http://tekstlab.uio.no/obt-ny/english/index.html	Bob	GPL	Bokmål and Nynorsk			Yes		No	Probably not	No		0	
The Stanford Parser	Text Processing	NLP Tools	Parser	This package is a Java implementation of probabilistic natural language parsers, both highly optimized PCFG and lexicalized dependency parsers, and a lexicalized PCFG parser. The original version of this parser was mainly written by Dan Klein, with support code and linguistic grammar development by Christopher Manning. Extensive additional work (internationalization and language-specific modeling, flexible input/output, grammar compaction, lattice parsing, k-best parsing, typed dependencies output, user support, etc.) has been done by Roger Levy, Christopher Manning, Teg Grenager, Galen Andrew, Marie-Catherine de Marneffe, Bill MacCartney, Anna Rafferty, Spence Green, Huihsin Tseng, Pi-Chuan Chang, Wolfgang Maier, and Jenny Finkel.	http://nlp.stanford.edu/software/lex-parser.shtml	Bob	GPL	3	Java		Yes		Yes	Active development	Yes		0	

Succeed List of Tools

Name of the tool	Group	Type	Subtype	Description	Link to the tool/website	Entry author	Type of license	Language support	Tech. context	Time and effort for installation	1st criterion for exclusion: If there is no trial version available to test a tool within Succeed, the tool will be discarded		2nd criterion for exclusion: If there is no technical documentation or support available, the tool will be discarded		3rd criterion for exclusion: If there is no information about the tool being used in other projects, no information about existing benchmarks or no information from users about the tool, the tool will be discarded		Relevance for libraries: A rating from 0-5 indicates how relevant a tool is for libraries (0 = not relevant; 5 = very relevant). The description provides further insights on how the consortium came to this rating.	
											Trial version available (Yes/No)	Further Description	Documentation/S available? (Yes/No)	Further Description (e.g. available support, activity status/last update)	Information assuring tool performance available? (Yes/No)	Further Description (Applicability to mass digitisation, quality and robustness)	Rating (0-5)	Further Description
TnT – Statistical Part-of-Speech Tagging	Text Processing	NLP Tools	POS Tagger	TnT, the short form of Trigrams'n'Tags, is a very efficient statistical part-of-speech tagger that is trainable on different languages and virtually any tagset. The component for parameter generation trains on tagged corpora. The system incorporates several methods of smoothing and of handling unknown words. TnT is not optimized for a particular language. Instead, it is optimized for training on a large variety of corpora. Adapting the tagger to a new language, new domain, or new tagset is very easy. Additionally, TnT is optimized for speed.The tagger is an implementation of the Viterbi algorithm for second order Markov models. The main paradigm used for smoothing is linear interpolation, the respective weights are determined by deleted interpolation. Unknown words are handled by a suffix trie and successive abstraction.	http://www.coli.uni-saarland.de/~thorsten/tnt/	Bob	Proprietary License				Yes	need to fill out form and fax/email it	No	Unclear	Yes		0	
transition-based dependency parser	Text Processing	NLP Tools	Parser	Bernd Bohnet and Joakim Nivre. 2012 A Transition-Based System for Joint Part-of-Speech Tagging and Labeled Non-Projective Dependency Parsing. EMNLP-CoNLL, pages 1455-1465 [pdf] bib]	http://code.google.com/p/mate-tools/	Bob	GPL	English, German, Chinese, other languages require training corpus	Java		Yes		Yes	Very concise	Yes		0	
TXM	Text Processing		text analysis tool	It offers a comprehensive range of analysis tools (concordances, collocate search, frequency lists, etc.) based on the powerfull CQP full text search engine (http://cwb.sourceforge.net) and a range of statistical functions (factorial analysis, classification, cooccurrency analysis, etc.) based on R packages (http://www.r-project.org).	http://sourceforge.net/projects/txm/	Tomasz	GNU General Public License version 3.0 (GPLv3)	English, French, Russian			Yes		No	Did not find	No	Not sure	0	
VARD 2	Text Processing		spelling variations	VARD 2 is an interactive piece of software produced in Java designed to assist users of historical corpora in dealing with spelling variation, particularly in EModE texts.	http://www.comp.lancs.ac.uk/~barona/vard2/	Tomasz	Creative Commons Attribution-Noncommerc Alike 2.0 UK: England & Wales License.	Early Modern Englis but can be extended via plugins			Yes		Yes		No	Not sure because last update was in August 2011	1	
WCRFT	Text Processing	NLP Tools	POS Tagger	WCRFT (Wrocław CRF Tagger) is a simple morpho-syntactic tagger for Polish producing state-of-the-art results.The tagger combines tiered tagging, conditional random fields (CRF) and features tailored for inflective languages written in WCCL. The algorithm and code are inspired by Wrocław Memory-Based Tagger.WCRFT uses CRF++ API as the underlying CRF implementation. Tiered tagging is assumed. Grammatical class is disambiguated first, then subsequent attributes (as defined in a config file) are taken care of. Each attribute is treated with a separate CRF and may be supplied a different set of feature templates.	http://nlp.pwr.wroc.pl/redmine/projects/wcrft/wiki	Bob	GPL	Polish	Python		Yes		Yes	Not much activity, no large community	No		0	
WCRFT (Wrocław CRF Tagger)	Text Processing		CRF tagger	WCRFT is a simple morpho-syntactic tagger for Polish producing state-of-the-art results. The tagger combines tiered tagging, conditional random fields (CRF) and features tailored for inflective languages written in WCCL. The algorithm and code are inspired by Wrocław Memory-Based Tagger. WCRFT uses CRF++ API as the underlying CRF implementation. Tiered tagging is assumed. Grammatical class is disambiguated first, then subsequent attributes (as defined in a config file) are taken care of. Each attribute is treated with a separate CRF and may be supplied a different set of feature templates.	http://nlp.pwr.wroc.pl/en/tools-and-resources/wcrft-tagger	Tomasz	unknown	Polish			Yes		Yes		Yes		0	
WMBT	Text Processing	NLP Tools	POS Tagger	WMBT (Wrocław Memory-Based Tagger) is a simple morpho-syntactic tagger for Polish producing state-of-the-art results. WMBT uses TiMBL API as the underlying Memory-Based Learning implementation. The features for classification are generated by using WCCL	http://nlp.pwr.wroc.pl/redmine/projects/wmbt/wiki	Bob	Unclear	Polish			Yes		No	Unclear	No		0	
WordFreak	Text Processing		annotation tool	WordFreak is a java-based linguistic annotation tool designed to support human, and automatic annotation of linguistic data as well as employ active-learning for human correction of automatically annotated data. Java based.	http://wordfreak.sourceforge.net/index.html	Tomasz	Mozilla Public License 1.1 (MPL 1.1)		Java		Yes		No		No	Old project	0	
Xerox	Text Processing	NLP Tools	Language Identification	This service will tell you the language your document is written in. Language identification is often the first, necessary step in a whole line of document processing.	http://open.xerox.com/Services/LanguageIdentifier	Bob	commercial	47	online service		Yes		Yes		Yes		4	
Abbyy FineReader Engine	Text Recognition	Core Text Recognition		State-of-the-art OCR engine	http://www.digitisation.eu/tools/ocr-engines/abbyy-finereader-engine/	Sebastian	commercial		Web Service/SDK		Yes	Through IMPACT Center of Competence	Yes	active (2013)	Yes		5	
ALTO-Edit	Text Recognition	Postcorrection		ALTO Editor for text and segmentation	https://github.com/impactcentre/alto-editor	Clemens	GPL	Not applicable	Javascript, Ruby	Considerable. Requires compilation of source code and configuration of Apache web server.	Yes	https://github.com/KBNLresearch/alto	Yes	community	Yes	Tested in small KB pilot, report available from IMPACT http://www.digitisation.eu/blog/view/article/impact-p	2	prototype
Asprise	Text Recognition	Core Text Recognition		Asprise OCR SDK library for Java enables you to equip your Java applications (Java applets, web applications, standard applications, J2EE enterprise applications) with optical character recognition (OCR) ability.	http://asprise.com/product/ocr/index.php?lang=java	Katrien&	Own license		Java		Yes		Yes	commercial	No	http://www.complaintsboard.com/complaints/asprise-ocr-c103169.html : "OCR product is a scam" is title review	0	
BIT-Alpha	Text Recognition	Core Text Recognition		Small French company that offered trainable OCR based on Neuronal Networks with support for Fraktur.	http://www.i-d-e.de/wordpress/wp-content/uploads/2009/08/tomasi.pdf	Clemens	Commercial	German, French	C	Available as Windows-based installer (MSI), thus installation effort rather minimal (might require Admin rights).	No		No	Last heard of in 2009. Website returns 404.	No	Quality of results *can* be extremely good given a huge training effort (neuronal networks). However, latest version tested appeared unstable. See reports from BSB, ZLB (available through IMPACT)(crashes, undocumented functions).	0	

Succeed List of Tools

Name of the tool	Group	Type	Subtype	Description	Link to the tool/website	Entry author	Type of license	Language support	Tech. context	Time and effort for installation	1st criterion for exclusion: If there is no trial version available to test a tool within Succeed, the tool will be discarded		2nd criterion for exclusion: If there is no technical documentation or support available, the tool will be discarded		3rd criterion for exclusion: If there is no information about the tool being used in other projects, no information about existing benchmarks or no information from users about the tool, the tool will be discarded		Relevance for libraries: A rating from 0-5 indicates how relevant a tool is for libraries (0 = not relevant; 5 = very relevant). The description provides further insights on how the consortium came to this rating.	
											Trial version available (Yes/No)	Further Description	Documentation/S available? (Yes/No)	Further Description (e.g. available support, activity status/last update)	Information assuring tool performance available? (Yes/No)	Further Description (Applicability to mass digitisation, quality and robustness)	Rating (0-5)	Further Description
Carleton OCR	Text Recognition	Core Text Recognition		Code repository for the Carleton OCR comps project 2010-2011	https://code.google.com/p/carletonocr/	Katrien&	MIT		Python		Yes		No				0	
ClaraOCR	Text Recognition	Core Text Recognition		Clara OCR is an Optical Character Recognition program. It features both a powerful GUI for the X Window System, and a Web interface. The Web interface is able to collect revision efforts from the Internet, using a simple revision model. It is intended to be used in the cooperative optical recognition of old books. It tries to facilitate fine- tuning, so an optical recognition project is enabled to invest resources in tuning the OCR, in order to achieve better recognition results for one specific book, and reduce the overall revision cost.	http://freecode.com/projects/claraocr	Katrien&	GPL		C		Yes		No	no more activity since 2001, version 0.9.9			0	
Collaborative Correction Platform (CONCERT)	Text Recognition	Postcorrection		A web-based platform, suitable for massive volunteer participation, which validates and corrects OCR results	http://www.digitisation.eu/tools/ocr-post-correction-and-enrichment/collaborative-correction-platform/	Sebastian	commercial	English, Dutch, German	Java web application, IBM stack (Webshere, DB2)	Major unless used as a web application hosted by IBM	Yes		Yes	IMPACT deliverable and User Guide	Yes	Crowd-correction tool tested at Hearst Archive (UCLA) and Japan Diet Library	0	
cs499ocr	Text Recognition	Core Text Recognition		Performs OCR with image processing and statistical pattern recognition.	https://code.google.com/p/cs499ocr/		GPL		Java		Yes		No				0	
Cuneiform	Text Recognition	Core Text Recognition		Cuneiform is an OCR system. In addition to text recognition it also does layout analysis and text format recognition. Cuneiform supports several languages.	http://en.openocr.org/	Katrien&	Own license		C		Yes		No	refer to openocr.org; last news of project 2009; last bug report 2010 unsolved			1	
Cutouts	Text Recognition	Postcorrection	Utilities for training and customization	Cutouts is a web application which allows to crowdsource preparation of training data for Tesseract OCR engine.	http://wlt.synat.pcss.pl/cutouts	Tomasz	free				Yes		Yes	PSNC support	Yes	in Poland	4	
Expervision OpenRTK	Text recognition	Core Text Recognition	OCR	OpenRTK 7.0 (Open Recognition Toolkit) is a C/C++ toolkit that provides an innovative solution to application developers, system integrators and OEM customers who need to integrate OCR capability into their applications with minimum engineering efforts.	http://www.expervision.com/ocr-sdk-toolkit/openrtk-ocr-toolkit-sdk	Tomasz	Commercial	English, French, German, Italian, Spanish, Portuguese, Danish, Dutch, Swedish, Norwegian, Hungarian, Polish, Finnish	C++		Yes		Yes	commercial	No		1	
EyeOCR	Text Recognition	Core Text Recognition		An OCR (Optical Character Recognition) application written in Java. Eye is easy and fun to use - no in-depth knowledge required. Eye is known to work on Linux, Windows and Mac OS X.	http://eyeocr.sourceforge.net/		Own license		Java		Yes		No				0	Very extensive toolset
FromThePage	Text Recognition	Postcorrection	Transcription	FromThePage is free software that allows volunteers to transcribe handwritten documents on-line.	http://beta.fromthepage.com/	Tomasz	GNU AGPL v3				Yes		Yes		Yes	Not many users	1	
Gamera OCR	Text Recognition	Core Text Recognition	Framework	OCR toolkit for Gamera: This is a Gamera toolkit for building standard text recognition applications. It is based on the Gamera framework and requires a working Gamera installation.	http://gamera.informatik.hsnr.de/addons/ocr4gamera/index.html	Katrien&	GPLv2		Python		Yes		Yes	community	Yes	projects using gamera: http://gamera.informatik.hsnr.de/links.html#usecases	4	Actively used in projects with apparently results which make it relevant in some fields (e.g. editions of classical greek and latin).
gOCR	Text Recognition	Core Text Recognition		GOCR is an OCR (Optical Character Recognition) program, developed under the GNU Public License. It converts scanned images of text back to text files.	http://jocr.sourceforge.net/	Katrien&	GPL		C		Yes		Yes	community (small)	No		0	
hOCR	Text Recognition	Core Text Recognition		HOCR is a Hebrew optical character recognition library.	http://hocr.berlios.de/	Katrien&	GPLv3		C		Yes		No	project has turned of bug tracker; no more development since 2008			0	
IBM Adaptive OCR Engine	Text Recognition	Core Text Recognition		IBM Adaptive OCR is a comprehensive software system which improves the recognition of historical texts significantly by applying adaptivity as one of the main features to the text recognition process. It integrates several other tools, such as the image enhancement toolkit, the ABBYY FineReader Engine, the post correction tool and the lexical resources developed during the IMPACT project.	http://www.digitisation.eu/tools/browse/ocr-engines/ibm-adaptive-ocr-engine/	Sebastian	commercial	English, Dutch, German	C, Java	Major unless used as a web application hosted by IBM	Yes		Yes	ICDAR2011 paper	Yes	Applicable to mass-digitisation with some constraints (manual interaction required).	0	
Inventory Extraction	Text Recognition			Allows for the extraction of a complete list of characters from a document, without reference to a specific language dictionary or a library of fonts.	http://www.digitisation.eu/tools/browse/experimental-prototypes/inventory-extraction	Sebastian	ASL 2.0	Not applicable	C++/C	Significant needs to be built from source (but ships with compiled binaries)	Yes	https://github.com/impactcentre/inven	Yes	not given	Yes	More of a niche tool to create font classifiers for currently unsupported alphabets/fonts	0	
JavaOCR	Text Recognition	Core Text Recognition		This OCR engine is implemented as a Java library, along with a demo application which shows the library in action. The core concept, at the character level, is image matching with automatic position and aspect ratio correction, using a least-square-error matching algorithm. It is a very simple yet reasonably effective implementation.	http://roncemer.com/software-development/java-ocr/	Katrien&	BSD		Java		Yes		No	It is on sourceforge, but hardly active; no replies to feature requests or bugs; http://sourceforge.net/projects/javaocr/			0	
Kognition	Text Recognition	Core Text Recognition		An omnifont OCR software for KDE. Due to the fact that each step of the OCR process can be visualized you can get a quick idea of how OCR works and where the problems lie. However the program may be of minor/no use for end users in its current state.	http://sourceforge.net/projects/kognition/	Katrien&	GPLv2		C++		Yes		No	no more activity since version 0.1.1, 2005-05-06			0	
Korrektor	Text Recognition	Postcorrection		GUI-based software for viewing and correcting document analysis results	http://www.iais.fraunhofer.de/diensteplattform-technologien.html	Sebastian	commercial		Java standalone GUI tool	web service	Yes	Free to use in Succeed	Yes	Support through Fraunhofer IAIS. Last update: 2012	Yes	Used in a large newspaper project with the National Library in Berlin	4	
Lios	Text Recognition	Core Text Recognition		Lios is a free and open source software for converting print into text using either scanner or a camera. It can also produce text out of scanned images from other sources such as pdfs, images or folders containing images.	http://code.google.com/p/linux-intelligent-ocr-solution/	Clemens	GPLv3	Bulgarian, Croatian, Czech, Danish, Dutch, English, Estonian, French, German, Hungarian, Italian, Latvian, Lithuanian, Polish, Portuguese, Romanian, Russian, Russian-English bilingual, Serbian, Slovene, Spanish, Swedish, Turkish, and Ukrainian.	Python	Requires Debian-based Linux (Ubuntu, Mint), then available from package repository, thus mininmal effort.	Yes		Yes	community support, last update: February 2013		Results equivalent to Tesseract 3 (thus rather good). Applicability to mass digitisation untested.	1	Appears to be an linux UI for either tesseract of CUNEIFORM. User interfaces for OCR are moderately relevant in a library context.
Longan	Text Recognition	Core Text Recognition		A flexible pure-Java OCR implementation. The aim of this project is to write a reasonably (competent, modular, understandable) OCR system.	https://github.com/Zarkonnen/Longan	Katrien&	ASL 2.0		Java		Yes		No				0	
NeuroOCR	Text Recognition	Core Text Recognition		Demo neural network OCR	http://www.codeproject.com/Articles/11285/Neural-Network-OCR		GPLv3		C#		Yes		No				0	

Succeed List of Tools

Name of the tool	Group	Type	Subtype	Description	Link to the tool/website	Entry author	Type of license	Language support	Tech. context	Time and effort for installation	1st criterion for exclusion: If there is no trial version available to test a tool within Succeed, the tool will be discarded: Trial version available (Yes/No)	Further Description	2nd criterion for exclusion: If there is no technical documentation or support available, the tool will be discarded: Documentation/S available? (Yes/No)	Further Description (e.g. available support, activity status/last update)	3rd criterion for exclusion: If there is no information about the tool being used in other projects, no information about existing benchmarks or no information from users about the tool, the tool will be discarded: Information assuring tool performance available? (Yes/No)	Further Description (Applicability to mass digitisation, quality and robustness)	Relevance for libraries: A rating from 0-5 indicates how relevant a tool is for libraries (0 = not relevant; 5 = very relevant). The description provides further insights on how the consortium came to this rating.: Rating (0-5)	Further Description
NewOCR	Text Recognition	Core Text Recognition		NewOCR.com is a free online OCR service based on Tesseract. It can analyze the text in any image file that you upload, and then convert the text from the image into text that you can easily edit on your computer	http://www.newocr.com/	Clemens	Own license	Same as Tesseract 3, see also website	Web Service	none (runs in Web browser)	Yes		No	active		Results equivalent to Tesseract 3 (thus rather good). Service responsive and enthusiastic comments from users on website. Applicability to mass digitisation not given. Requires manual page-by-page upload through web form. No API planned. But also no page limit!	1	Web interface to tesseract. This kind of interface is obviously useful for end users, but not directly relevant for libraries, who would be more interested in a data-only web service.
OCR gem	Text Recognition	Core Text Recognition		Recognize text and characters from image files using web services.	http://rubygems.org/gems/ocr	Katrien&	MIT	brazilian, byelorussian, bulgarian, catalan, croatian, czech, danish, dutch, english, estonian, finnish, french, german, greek, hungarian, indonesian, italian, latin, latvian, lithuanian, moldavian, polish, portuguese, romanian, russian, serbian, slovakian, slovenian, spanish, swedish, turkish, ukrainian	Ruby		Yes		No	current version 0.3.1, released april 2012, first release february 2012, no activity after this			0	Offers tools as webservice
ocrad	Text Recognition	Core Text Recognition		GNU Ocrad is an OCR (Optical Character Recognition) program based on a feature extraction method. It reads images in pbm (bitmap), pgm (greyscale) or ppm (color) formats and produces text in byte (8-bit) or UTF-8 formats. Also includes a layout analyser able to separate the columns or blocks of text normally found on printed pages. Ocrad can be used as a stand-alone console application, or as a backend to other programs.	http://www.gnu.org/software/ocrad/	Katrien&	GPL		C		Yes		Yes	research project	No	last version 0.21 (2011)	0	
OCRchie	Text Recognition	Core Text Recognition		The original OCR package could learn from a tif file and ascii translation, then recognize a document in the same font. This semester we added interactive learning, interactive segmentation of mathematics, page zoning (the ability to automatically or manually zone columns or regions of text, and interactive read-order specification.	http://www.cs.berkeley.edu/~fateman/kathey/ocrchie.html	Katrien&	unknown		C++		Yes		No	research project end of '90. Last visible update 2000			0	
ocre	Text Recognition	Core Text Recognition		Spanish OCR prototype	http://lem.eui.upm.es/ocre.html	Katrien&	unknown	English, Euskara/Basque, French, German, Polish, Português, Russian, Spanish	C		Yes		Yes	Seems project from 1 individual (last posting 2012)	No	last version: 0.042 (2012)	0	
OCRFeeder	Text Recognition	Recognition		optical character recognition system	https://live.gnome.org/OCRFeeder	Sebastian	GPL		GUI, Python		Yes	Open Source	Yes	documentation	No	use	0	
OCRopus	Text Recognition	Core Text Recognition		OCRopus is an OCR system focusing on the use of large scale machine learning for addressing problems in document analysis	https://code.google.com/p/ocropus/	Sebastian	Apache License v2		Python library		Yes	Open Source	Yes	Wiki and community (forum)	No		2?	waiting for results competitiens ICDAR 2013; new developments
OmniPage	Text Recognition	Core Text Recognition		State-of-the-art OCR engine	http://www.nuance.com/for-business/by-product/omnipage/index.htm	Tomasz	commercial	123 languages			Yes		Yes		Yes		2	no special support for libraries, eg. supporting library formats, fraktur.
Paradiit	Text Recognition	Core Text Recognition	Framework	The PaRADIIT (Pattern Redundancy Analysis for Document Image Indexing and Transcription) project is a research project conducted by the RFAI Team of the Computer Science Laboratory of Tours. The project focused on layout analysis, text/graphics separation, Optical Character Recognition (OCR) and text transcription processes dedicated to old books and historical documents. Additions: This is very much like the IBM concert tool also has ideas related to the inventory extraction! It consists of two processing steps: AGORA which extracts clusters of characters, and RETRO which presents something like IBM's carpets.	https://code.google.com/p/paradiit/	Katrien&	GPL		C#		Yes		Yes	project	Yes	Ongoing research project; several publications available, protoype available; worth mentioning, but not for implementation.	1	Since project results are still in the prototype phase, not ready for implementation.
Photoscore	Text Recognition	Core Text Recognition		Music OCR: music scanning & PDF to notation	http://www.neuratron.com/photoscore.htm	Tomasz	commercial				Yes		Yes		Yes		1	
Plasma OCR	Text Recognition	Core Text Recognition		An omnifont OCR engine. The long-term goal is recognition of formulas.	http://developer.berlios.de/projects/plasmaocr/	Katrien&	GPL		C, C++		Yes		No	version 0.1; no activity since then (2006)			0	All important tools bundled in a workflow
Post Correction Tool	Text Recognition	Postcorrection		Interactive post-correction of OCRed documents	http://www.digitisation.eu/tools/ocr-post-correction-and-enrichment/post-correction-tool/	Sebastian	unknown	German, Dutch	C, Java	Significant (compilation from source, includes Java UI and C backend)	Yes	Upon request from LMU - UI supposed to be open source?!	Yes	Supported through User Guide and technical documentation but developers have left the university	Yes	Supports batch correction of errors but tool itself not fully stable yet	1	

Succeed List of Tools

Name of the tool	Group	Type	Subtype	Description	Link to the tool/website	Entry author	Type of license	Language support	Tech. context	Time and effort for installation	1st criterion for exclusion: If there is no trial version available to test a tool within Succeed, the tool will be discarded: Trial version available (Yes/No)	Further Description	2nd criterion for exclusion: If there is no technical documentation or support available, the tool will be discarded: Documentation/S available? (Yes/No)	Further Description (e.g. available support, activity status/last update)	3rd criterion for exclusion: If there is no information about the tool being used in other projects, no information about existing benchmarks or no information from users about the tool, the tool will be discarded: Information assuring tool performance available? (Yes/No)	Further Description (Applicability to mass digitisation, quality and robustness)	Relevance for libraries: A rating from 0-5 indicates how relevant a tool is for libraries (0 = not relevant; 5 = very relevant). The description provides further insights on how the consortium came to this rating.: Rating (0-5)	Further Description
																http://chnm.gmu.edu/digitalhistory/digitizing/4 A study based on the Making of America project at Michigan found that about nine out of ten OCRed pages had 99 percent or higher character accuracy without any manual correction. A Harvard project that measured search accuracy instead of character accuracy concluded that uncorrected OCR resulted in successful searches 96.6 percent of the time with the rate for twentieth-century texts (96.9 percent) only slightly higher than that for nineteenth-century works (95.1 percent). To be sure, both of these projects used PrimeOCR, the most expensive OCR package on the market. Also http://www.lib.umich.edu/files/services/dlps/moa4 http://chnm.gmu.edu/digitalhistory/links/pdf/c BUT information seems hardly updated over many years. The phrase: "PrimeOCR is able to reduce OCR errors by an average of 65-82% over the best conventional OCR software products" appears with older versions dating back to 2006, and is still used with identical scores - which suggests it is just repeated without new statistics. This would put me (Jesse) off a bit. Other engines (at least abbyy) internally perform voting as well - should we believe they are not making any progress? And is it not strange that version 5.0 claims improved accuracy while mentioning identical percentages in the promo?		
PrimeOCR	Text recognition	Core Text Recognition	OCR	Prime Recognition's production OCR product, PrimeOCR is a Windows OCR engine that claims to reduce OCR error rates by up to 65-80% over conventional OCR by implementing "Voting" OCR technology.	http://www.primerecognition.com/prime_ocr.htm	Tomasz	Commercial	Danish, English, German, Norwegian, Spanish, Dutch, French, Italian, Portuguese, Swedish	?		No		Yes	commercial			1	Problematic PR (cf previous columns). Seems not really active after 2005; expensive.
Proofread page	Text Recognition	Postcorrection		Proofread Page is an extension for MediaWiki which allows you to edit transcriptions side by side with the page images. It is used on WikiSource for manuscript and early print transcription projects. Proofread Page supports workflow, but no markup.	http://dirt.projectbamboo.org/resources/proofread-page	Tomasz	GPL v2				Yes		Yes		No	Not sure	0	
ReadIris	Text recognition	Core Text Recognition	OCR	Readiris is a OCR solution designed for private users and small to large office users	http://www.irislink.com/c2-2115-189/Readiris-14--OCR-Software--Scan--Convert---Manage-your-Documents-.aspx	Tomasz	Commercial	140 languages	?		Yes		Yes	commercial		http://conference.ifla.org/past/ifla75/106-matusiak-en.pdf; http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2479(	2	less accurate than Finereader or Omnipage
Schnell OCR	Text Recognition	Core Text Recognition		A lightweight ocr module written in C	https://github.com/jagd/schnell-ocr		unknown		C		Yes		No				0	
Scripto	Text Recognition	Postcorrection	Transcription	A free, open source tool enabling community transcriptions of document and multimedia files	http://scripto.org/	Tomasz	GPLv3		PHP		Yes		Yes	community	Yes	Mailing list active	1	
SharpEye 2	Text Recognition	Core Text Recognition		Music OCR: You can use SharpEye to scan and convert printed sheet music into a music notation file or a MIDI file which can then be imported into a music notation program or MIDI sequencer	http://www.visiv.co.uk/	Tomasz	commercial $169				Yes	30 days	Yes		Yes	Not sure - the page is quite old I guess.	1	
SimpleOCR	Text Recognition	Core Text Recognition		SimpleOCR is the popular freeware OCR software with hundreds of thousands of users worldwide. SimpleOCR is also a royalty-free OCR SDK for developers to use in their custom applications.	http://www.simpleocr.com/	Katrien&	Own license	English, French			Yes		Yes		No	http://www.pcmag.com/article2/0,2817,1681515,00.asp; poor results: capitals already give problems	0	
SimpleOCRSDK	Text Recognition	Core Text Recognition		The SimpleOCR SDK is a fast, lightweight OCR engine designed to let developers add basic OCR functions to an application with minimal cost and none of the drawbacks of open source solutions.	http://www.simpleocr.com/Info.asp#SDK	Katrien	own license	English, French,			Yes		Yes		No	not for complex lay out	0	
SmartScore	Text Recognition	Core Text Recognition		Music OCR: Recognizes scores without any restriction on the number of parts. Process band arrangements, operas, hymns, musicals, instrumental and solo parts as well as full conductor's scores.	http://www.musitek.com/	Tomasz	commercial				Yes		Yes		Yes		1	
T-pen	Text Recognition			T-PEN is a web-based tool for working with images of manuscripts. Users attach transcription data (new or uploaded) to the actual lines of the original manuscript in a simple, flexible interface.	http://t-pen.org/TPEN/	Tomasz	ECL				Yes	online account	No	online service, no docs found	Yes	there are projects	1	
Tesseract	Text Recognition	Core Text Recognition		Tesseract is probably the most accurate open source OCR engine available	https://code.google.com/p/tesseract-ocr/	Sebastian	Apache License v2		CL tool (cross platform)		Yes	Open Source	Yes	very active community	Yes	Used and evaluated in various research and industry projects	5	

Succeed List of Tools

Name of the tool	Group	Type	Subtype	Description	Link to the tool/website	Entry author	Type of license	Language support	Tech. context	Time and effort for installation	1st criterion for exclusion: If there is no trial version available to test a tool within Succeed, the tool will be discarded		2nd criterion for exclusion: If there is no technical documentation or support available, the tool will be discarded		3rd criterion for exclusion: If there is no information about the tool being used in other projects, no information about existing benchmarks or no information from users about the tool, the tool will be discarded		Relevance for libraries: A rating from 0-5 indicates how relevant a tool is for libraries (0 = not relevant; 5 = very relevant). The description provides further insights on how the consortium came to this rating.	
											Trial version available (Yes/No)	Further Description	Documentation/S available? (Yes/No)	Further Description (e.g. available support, activity status/last update)	Information assuring tool performance available? (Yes/No)	Further Description (Applicability to mass digitisation, quality and robustness)	Rating (0-5)	Further Description
Text and Error Profiler	Text Recognition	Postcorrection		The Text and Error Profiler is software to analyse the OCR output from historical documents, using statistical modelling of document characteristics to improve OCR accuracy. It works by attuning itself to a particular document, rather than to common traits of printed documents from a certain era, resulting in a highly adaptive process. The tool uses its document-specific knowledge to allow the batch processing of erroneous words.	http://www.digitisation.eu/tools/ocr-post-correction-and-enrichment/text-and-error-profiler/	Jesse&	Licence pending. For further information, please contact the IMPACT Centre of Competence	Language-independent	C++		Yes		Yes		Yes		1	
Transcript	Text Recognition	Postcorrection		Transcript is a desktop-based manuscript transcription tool that supports word-processor style formatting.	http://www.jacobboerema.nl/en/Freeware.htm	Tomasz	free or 15 EUR				Yes		Yes		No		0	
Typereader	Text Recognition	Core Text Recognition		TypeReader®has been in the global market and received hundreds of appraisals from various industry technology magazines since 1991. The heart of this award winning OCR software product, ExperVision®'s OpenRTK®, is the only OCR Engine which won UNLV Test for consecutive years. Commercial (server/desktop)	http://www.expervision.com/ocr-software/desktop-ocr-typereader-7	Katrien	Own license		?		Yes		Yes	commmercial	Yes	Fast; only to consider if accurracy is not the most important (review PC Magazine 2008)	0	
Typewright	Text Recognition	Postcorrection		TypeWright1 is a tool for correcting the text-version of a document made up of page images.	http://www.18thconnect.org/typewright/documents	Clemens	ASL 2.0	English	Ruby on Rails	Probably significant as originally integrated with Collex collection management software. Currently being refactored.	Yes		Yes	commercial/communit (the tool is developed by Performant Software, a US company specialised on development of Digital Humanities Tools in funded projects); SVN (private) has updates every week.	Yes	Used by large US projects 18thConnect and NINES (collection of 180,000 books), further developed by Performant Software in eMOP project (KB, USAL partners)	1	Currently not yet available independently
Typewritten OCR	Text Recognition	Core Text Recognition		OCR Prototype for recognising typewritten documents incorporating background knowledge about the specific features of this type of documents.	http://www.digitisation.eu/tools/experimental-prototypes/typewritten-ocr/	Sebastian	unknown				Yes		Yes		No		1	prototype
Virtual Transcription Laboratory	Text Recognition	Postcorrection		Virtual Transcription Laboratory is Virtual Research Environment which works as a crowdsourcing platform for developing high quality textual representations of digital documents. It gives access to online OCR service and easy to use transcription editor. Images can be imported from various sources including direct import from digital libraries.	http://wlt.synat.pcss.pl	Tomasz	free				Yes		Yes	PSNC support	Yes	in Poland	4	
WeOCR	Text Recognition	Core Text Recognition	Web service	WeOCR is a platform for Web-enabled OCR (Optical Character Reader/Recognition) systems. It enables people to use character recognition over networks. A WeOCR server receives document images from users, recognizes text in the images, and returns recognition results to the users. WeOCR does not have its own character recognition engine. Instead, it is intended to accommodate various existing character recognition engines.	http://weocr.ocrgrid.org/	Katrien&	ASL 2.0		C		Yes		Yes	single researcher (Hideaki Goto) last release version 0.14 [june 2012]	No	(several publications) no recent activity; unknown how much it is used	0	
Word Spotting	Text Recognition			This tool provides an integrated GUI for indexing historical documents without an OCR engine. It works by segmenting documents into individual words and compiling a list of the most common words (keywords) in the text. Users are then asked to classify the keywords	http://www.digitisation.eu/tools/experimental-prototypes/word-spotting/	Sebastian	commercial	Not applicable	C++	Significant. Needs mySQL database.	Yes	Available through contacting NSCR	Yes	IMPACT deliverables	Yes	Not suitable for mass digitisation. Alternative to commercial "poor mans OCR".	0	
Wordsnap OCR	Text Recognition	Core Text Recognition		An app for OCR-based camera input on Android	https://code.google.com/p/wordsnap-ocr/	Katrien&	GPLv3		Java		Yes		No	all versions in 2009; version 0.3 last one; some activity, individual researcher's project			0	
XPLab	Text Recognition	Core Text Recognition		XPLAB tries to recognize patterns in a scanned document image by trained templates stored in a database. The main phases are training, recognition and maintenance. The user can switch easily between all phases in the same session. Some effort is made to simplify the training phase, which is the most time consuming part of interactivity.	http://www.pattern-lab.de/	Katrien&	GPL		C		Yes		No	version 0.4.6, last update dec 2012; most activity between 2004 and 2009 (change log)			0	Very extensive toolset