

The birth of a massive search engine for historical and multicultural handwritten collections

prof. dr. Lambert Schomaker

Lemy Fonry pastal alorentic Enrouse de



getahrever, rempl. des Moonrede. Il verlangt dat voortaan in het q Johreven energel des Proomede don Hill. de opening van de Staten Generaal. ruchen 10% het ande echer bladzyde worde vermel het woon Waarinede de Volgende bladzyde aanden

CERL Annual Seminar Oslo 28-10-2014



Continuous machine learning in Big Data

prof. dr. Lambert Schomaker

heming Fonry partat alorentic Enrouse de



Chempl. des Troonrede. H. M. verlangt dat voortaan in het 9 Achreven energel des Vroomede don Hell. de opening van de Staken Generaal te ge bruchen 10% het ande eener bladzyde nie worde vermel het woon waarinede de Volgende bladzyde aanden

CERL Annual Seminar Oslo 28-10-2014



The Monk System

prof. dr. Lambert Schomaker

demy Fonry partial alorentie forome de



Hetelnewey Trooverede. - H. M. verlangt dad voort aan in het ge Schreven enemge des Troomede dom H. M. bij No opening van de Staten Generaal te ge bruchen Top het oinde eene, bladzijde kiet worde vermelig het woond Waarmede de Nolgen de bladzijste aanvang

CERL Annual Seminar Oslo 28-10-2014





Monk e-Science web service addressing these **questions**:

- What? Word retrieval by 24/7 machine learning
 - Ambition: A European Google for handwriting
 - Server storage 1PB, filled 200 TB, with 37 collections
 - Qumran, Medieval, 17^{th} century, 20^{th} century, etc.
 - 400 M word images, 500 k labeled words, 30 k classes
- When? Medieval manuscript dating
 - On a server (MPS), uploading of charters from 1300-1550
- Who? Writer identification
 - On Monk server over internet
 - Using GIWIS Windows tool









Monk

Amsterdam.



Strafgeringenie







TTO foto



What?



(alast) Caop Trentterre



When? Medieval Paleographical Scale (MPS) with prof. Jan Burgers (paleographer, Huygens Inst.)





Medieval Paleographic Scale (under dev.)



Writer identification (1:N) and verification (1:1)

fully manual

1. Interactive, **2. Automatic**, **3**. OCR based

ROI based

and a set the for the prove that for finding the second by the the more shall - it as so that a the prove the terms and the the term the term the term and the term t men the fire a manufagting of that to at all allow a sole alles many the . at the to separate all a sharp of all life - and and a some sof maples any to be warfer the hard so of publick so plan a line and a so a top and an interest a sure gave from her in Who as allow higher on the ophilalline it and the ophil an a sindiate man to DE all all all all all all all and a stand and and a and to blive the first with any state all and and planter be all and a a futer and a ask par to starrate for a time rate man time it give but which we which we wanted allo all the art of the the set the all a to get aller the maked it were a shirt of my at a shirt allow at a sum Dr. S. B. and a alore - Douberman Vin rate beliefen Valer Sin polafe to him the set and i water pour pluster profess a product ou "From An dat dections Sec. Sec. porfe - The same an latin I'm for to all up as making it's set the finding the attack the popular an · some histor or to the link a little to Westingthe and rame Dry - Der De De informent S as app 2 and at a subjection and and a million of the start for mon taking Barpo Lines plainte I let mit multiful a star more it to the att mit to a so agen the liter all the right alpen with a summaries to a the right affer all her Do grantinger & Mr all timperen proder South Letter philes the at his me atte mothe matche pro the la alter from as a construction it is a mother at mother atilles to a me tiper ; and polar approximate tiller the power of polar or anter in such and the polar is no un and a matight and and the to the at a part when a should be at the at th



Region Of Interest





Researchers

- > Marius Bulacu, Axel Brink, Katrin Franke
- Ralph Niels, Louis Vuurpijl, Jean-Paul van Oosten,
 Sheng He, Jan Burgers, Petros Samara, Olarik Surinta
- > "The Nijmegen Handwriting Group 1984-1993"
- Netherlands Forensic Institute: Ton Broeders, Wil Fagel, Elisa van den Heuvel
- > Isabelle Guyon, Rejean Plamondon
- > Users (humanities researchers):
 - Jinna Smit, Mark Aussems, Masahiro Niitsuma, Mladen Popovic, Daniel Stoekl, Jetze Touber, Grace Fong, Elaine Treharne, ...





Overview

- > Digitisation of handwritten manuscripts
- Image processing challenges
- Text retrieval
 - Monk system: continuous learning
- > Difficult material & the role of Knowledge
- Conclusions





Digitisation

- Int. Conf. on Document Analysis & Recognition 2009, Barcelona
 - Representative of British Library: we have computed that we would need three petabytes (three thousand disks of a terabyte, i.e., 3 PB = 3 million gigabytes)
 - ... therefore we are obliged to use 'lossy' compression techniques for storing images ...
- › ??
- > We are using <u>10 PB</u> in Groningen just for some astronomers, manuscript archives and biomedical researchers !





Digitisation

- Current disk capacities allow for a full exploitation of al digital renderings of an object:
 - Gray scale
 - Color
 - Multi-spectral
 - 3D scanner (for Swedish haellristningar, Roman and Greek stone inscriptions, identification of the 'cisel')
- 1 PB for Monk ('in the noise', for astronomers)





Digitisation

- > Digitisation prevents wear and tear of the books
- > ... keep the books and documents in storage
- > Quality considerations
 - Color mode: RGB/gray/black-and-white
 - Dots per inch? Bits per pixel?
 - Mechanical safety of scanning system
- > Cost? (scanning, maintenance in digital life of object)
- > Speed?
- > Usage: How to make the digital access effective?





Case: Optical Character Recognition

- "The process of segmenting a text image into individual c h a r a c t e r images and classifying each as being a letter in an alphabet"
- > Impressive results on contemporary printed text in machine fonts: with some linguistic postprocessing results are close to 100%
- > Solved?





OCR ?

- > 'OCR' on historic documents does not work well
- > On *handwritten* manuscripts it doesn't work, at all!
- > Problems:
 - image quality
 - unknown character shapes
 - unknown statistical language models
- > However: pattern recognition and machine learning make enormous progress these days!
- > Which methods? How to apply them?





Current technology: neat text/known language

- > Why is 'OCR', i.e., letter by letter transcription on handwriting so difficult?
- Machine print: per character, per font, 8000 training examples are needed, typically
- > E. Barney-Smith: 200k instances of printed \mathbf{c} vs \mathbf{e}
- > Adress reading: reduced lexicon, zip codes etc., **help**
- In linguistic modeling: 20th century newspaper corpora do very little for 15th century acts
- Literary text, acts and charters each need their own knowledge models in order for OCR to work



Handwriting recognition: eat this!



- Many languages, scripts
- Over historical periods
- Contractions of letters
- 'Suggested' sloppy letter shapes
- Individual writer styles
- Image problems
- → Sliding window for character search usually problematic:







Monk -

Design considerations

- 1. Don't promise perfection
- 2. Don't promise 'transcription'
- 3. Don't promise exhaustive coverage (as in databases)
- 4. Make use of human trainers, volunteers
- > Word spotting:
 - "a Google for handwritten documents"





Monk -

Design considerations

- > Word spotting:
 - "a Google for handwritten documents"
- The word is a reliable chunk of information with many shape features: redundancy

mínímum

 > Big Data: With sufficient data, there is always a reasonable response on a query





Monk's world model:

- > Institutes
 - Collections
 - Books (i.e., documents)
 - Pages
 - Paragraphs
 - Lines



- Word zones and characters
 - Pixels



Example collection: Cabinet of the Queen



(KdK)

 > the Queen is the head of state
 > the head of state is a member of the government and signs all laws and many government decisions



KdK Archival system



- > all laws, decrees and correspondence are kept:
 - in a chronological order
 - and then in a numerical order





the Queen's Cabinet



- KdK Archive with a total extent of
 - (era 1798-1988):
 - 3,250 linear meter of shelves
- consisting of:
 - 28,000 boxes
 - average 1,000 pages per box
 - → 28,000,000 pages

•Of which Monk analyses the handwritten *Indices*, > 60k pages





Functions of Monk from user perspective

- Pages (read, annotate)
- > Lines (read, annotate, search)
- Words (cut ROI, labeling, inspect hit lists)

Human	Machine
Cut out region of interest with the mouse	Segmentation of image into objects
Label a ROI with a text label: region	Classify image object

 General search functions, 'google' over several collections

Peroolo van blang 10. Coursequentlas 9 11 1903 7 Kandongorechten. Mantongerechter. Maart 24 27 Rapps Hardlaart h 54 A. Kale, for beneening A Rappel & April 1 th & helt, the benaming tel Alfort by hat Manhengereald be Plietebet I in hed Ranton Harderwijk, I The Rosemal Syme & in het kanton Borgum, C.F. Trip Bashert fias - Mattheiten hiar Maach 31 19 Rappt / 27 Mac. A W. 128, Om a con l' de Levela 11 3 Rappt Ary April 1. 203 por aan Afflan Hay nel & la Sabloniere opign versal, con blad of rafe veryock , he verleenen een hours outslag and I lag the vorleaner intregne betrokking van han Tyne behekking van Rantoweedthe plaationen toprother plastourbanger in het kanton in hed kanton Halder, met daupheteren Backtergunag a suffer met dant betriging 11 Joslint fial __ Bethink filest April 1 2/ Vager With least whit I habit to entred duryon the for the bar of t 14 92 Tays 1 R 9 Spal 1 194 pm acm & Offerhand Betlind fins 1 Griffier by het Rantongerealt to Wearthat? It il Nortmann a 2 tot hor have anne tot 14 13 Rayar / P & Speel h loc our acon el " L' Sall Mantonceather plaatyvoringer in fet Kanton of ale boven was toublere as van het lipen Regeninger, Mr. J. J. Midda va Cappan bass Ministerie by de Randougereakten in - Meshint fias het anontripenent illeddollherg, von de kan Spril 6 11 Kappt for Spoul to by A lat. For benaoming long Meddelburg a Goes, haf Hand place Middelburg. do Kantongenetten en had apport formend Rotten dam, voor de kantone Schielden, Brielle an 111019 Kappt Ry April & Sq & lake ter henoeming Tot hap terstalle glanden of Helk. a such kanten Hoerden for Helk. I in het kanten Holen for den Bootheryn Sommelidigh toi Hantplagte Schiedan M" Litt Roalers Day Lermap - Borlin fiad April y 51 Rappt the April 1.148, en aan H. A. Aleyn vande foll op sign cogace to verleanen ed an Kapp / 17 April 17 70 , kakt to beneering Colondslagent signe beheckking von Ranton Yeather plaats we wanger in helplanton Harden Meyer Mirtura Wigh on her ho most dankhotinging - Deshitfiel - 12 alfit pier Mervolg op blad 12 Corvolger black 11



rall pm her valle m march m

KdK 1903 Scan Company A

Teroolgon Mar 173 19 ? Portification and little get 1997 " Portification en Militaire Colon 1997 " Porfification en Militaire Coloren Call to so frage it is to for the the son of the two of two of the two of two Out to say fractional to late to the late to be a state on one of the say that a grant day to the state of a state to be for the the termulation of the state being to only for infinition as working and state being to only the tool so Roy It' Collar Chadalan a to be them. Helen to be configured to the the second sec . For he ten himgen tappetievely & de Morten and Karelenet & loot a van de verten a Hallemet Out so 45 Kapt May bet 1 37 Of ader van 2 19. Soom an te altrait, om abbreve oan befor by het Henryar and they fall to 23 bearing Cath 30 45 Tayst Mary Pett 4 57, 69 active van I 84 Trol 4 se Frank March and the se It admer so It themas Sigh to far day on all here so a de server Sert 4 20 Rapp flipp latter in the admirant to terred and in Tamp and in a determinant for a series and a series and in Roman stiller Predict for - Reel + Had a see Tay the little in the second and a second and a second and a second and a second a Your y are Rapy Ild at beth to 19, It had onleaner war Yord 4 22 Rapp Ild 28 Dett to 19, For hid contemen an Augunance to be the house and have have been and the set on the Say the fill Hard with the hard the Hard man the the set on the Dark house the beale of a got that and the state north of Martin and the set of the set o All the definition of the and Thatten _ Bestud field _ Bushat feel Nort 11 34 Koppt Minter 13 30. 17 adure on Ghelden Julie Proster to the Seguring the held seasoft governet and some herder formal to an bola the Need it is Koppeld to alere to 3 30 to advise on globaldes Join a Prostantin on the Symming to had availed to a work a some dashed formalle and before the populantor 1 30 Made genow to berting haarden Buchuit he genow he besting haarden ; Steel 25 4 Fordell of the Aby the Adams on on Second States and the second seco Terd 25 44 Tappelling terd to by the descharmen and generate balance way be had an added at generate balance water interest former at been to be desting on the Welds Starket fail 23 W Rayd Bafton Party tothy conte so de Helling om de Helder Becht Nerorly op black 845 Mororly op Alach 845 Decorly op Had 84 -----Lin Twoolgon Had 178 Teroolgoan Mar 1/3. 314 1197 " Fortification en Militanel Chouse 147 " Portification en Militaine Colon 147 " Portification en Militaire Com Out to so front de to for the to the new form Coll is in fact the in fight of in the association of the second of the Cold 50 30 Karri dill'17, laft 17, lipotia la con dil 19 Mary agritta da la contrata da contrata da contrata da 19 Mary agritta da la contrata da la contrata da la contrata da contrata d ten kungen taspectieval & to de Werken on to Karrienalkilloot a. van de verling Hollowed Un hingen rapertievel & an de Marten an de Karelenald Stort a. va. de verleng Hallamet this _ Baslinta to this Out so 45 Kapt Milly lett 4 59, to advise a fifth Out so ys Kaper Miry both h so, by aday van 2 8.4 Out so 45 Kapt May bett 4 59 . Of adies von I S.A Boxman to Utreahl, bu abloven ou les Most 4 20 Page fil & lotto a. It adwood to the Sond 4 to Farge filling tother a Mantur on the terms and information on all terms on the content for the Manuscheller - Parlant first Vind 4 20 Rapp fli 2 bill h 12. Materica. It Reasons with the Jacker, on allow con the levels was to Ramanshellowt Allert Restint find _ Bul +! Stort 4 22 Tappet All 28 latter 19 He had and and and and a first on the second state of the second state How 4 22 Rapp Mar bet h 19. At had container van How 4 22 Rapp Mars beth no 19, tothed colene can I compose a la constance de la Augument of the factor of the constraint of the factor of _ Bushet fees __ Bestert feet tatting ____ Besturt feel Hert it is Roget Marter 1 3. pader on Glether Nort it at Rapp Mentert & 30. Spader con Gledke Nort is so Ropp Mentert & 30. Opedier on Gleddes for he Buston on very uning tot had consiste 12 te Buston, on organing to hed overall generan de Merting haarden Berhint fins, genoa se Verting Readen Berlint fich genera te Verteng Kaarden Barlins Stort 23 W Loger Dag torta for the forman on on Jones of the soft of the soft of the soft of the Jones of the soft of the soft of the soft of the Son to State of the Store of the Store of the Son to State of the Store of the Store of the How 23 44 Rappell 19 10 tor biby the hol contenes con any Fort 23 44 Rappella ig tort hby tot holo colena con con Jon and the second of the for the second of Jenne Caten wege to light a trans and and Arabitation had weller integenes tomme at them to be shelling on de Peter Statut first Aurorly of Mark 845 Aroolg op black 845 Nerorly of Med. 945

KdK 1897 Scan Company B

01 201 Par. ne herfen van s el fer Us res en Vi

KdK 1897 Scan Company B

of Revenue Sun Courses & Confort por porte as Course Very & and General Office and Course of Radia Called a softem for the pair of the softem south (General Radia Called and Softem for the softem for the softem for and the course manual Radia Called a course of the softem for a softem for and the course manual Radia Called a course of the softem for a softem for the new start of the Contro allor and current

Ething Save gaption for private grantite gritan deviallanderer 3" a new to the fine of lower of grann fire for dear the public state gaptie de vandere for same of the same case for dear of the and best my magne of magne gabter of anyone for fider ude a Okian prove handling per grad plan had been at the state of the learning of the state of the party handling the state of the standard of the stand prove the first party had the the state of the state of the state of the party and share to be the state of the state of the party party house the least and plant the state of the party party house the state of the party house the state of t

man friende por formo los Con - of andrew entran for the court - for the court Bath as former big of 9- 7 of for he ng app Pan Alin any home to make

Je Chille Fred and for the second of the transford the transford the second of the sec

4) Eline Ener galera for Firenz grunde given de mellenerary products of a new station of a second product of the second product of the second product of the second sec

Sam

SP de

an mitting an

of inter eit

Abud gran Jumy Sydrig

Bo Werkdy

At fifting the open of the start of the fifting the start of the start

from prinford - motorm and from the for the Course Der Byr of andrer entrenn Some finder work - A de some for so for first for for function over the off of some for some for the form the source of the so my firmit a file and one pro the mile

Jer Kunnen Sun Bruch & Reaforf & Jerpen 35 Course Bring & Bar General (Jack we Change of father for your 35 Course Bring & Bar See and the second second second for the second for the second second second second second for the second for the second for the second second second second second second second for the second second second second second second second second for the second sec

A place can applied in place grante provide a statistic cover place to be a new of place in place for place for place the place and the statistic cover place of the statistic coversite of the state of the state of the state of the place control of the state control of the state is not a state of the is not state of the is not state of the state is not state of the st

on printer - motorin an furthe for forward Con

-friend from George

for OGT For

Sold the of anon entry

Donn Brond muchans fo

Inophantik onver ~ f

boll as former 620

Senart during of Pi

Strang of the program

for and and and

I There and barries a first of provide Connect line proved and the second secon

the the address of the

the the say at you

I chan and in firmating and and the firm part of the firm

It for a set of the provide state of the set of the set

The property of the property o

Jer Runne Bur Bring & Conforth of parent & Commen Program (1986) General Ofta and Cannon for Andre & Califordia and for for the part and the month of a General Andre of the one of the for the part of the control Confort of the control of the one of the part of the part and Confort of T

It for a set and the former par for given a set of the former of the former of the former of the set of the se

I trom principar - interna - from a for an and a sure and a sure of a sure o

1

- State

SAL 1421 Company C

flus y

219

100117 LG. 28 220 albho fru? 64**8**° / Dropoly G 6 Sho Yor'

SAL 1421 Company C

1921 Hang Ritra godfi 19 Jimbregto my me film for coma of Anthe to a pring inf" litrate Con momende or 1457 He forwork fremens with fare fremens zoon boilen porte brunde for forwork fremens with gettind performing with surfamper 4- Gebbon Gan Japobpe Se Drugte artist gullen politis to wing funder effect, with origin fremen on afflage Gan mor orforman Gunna wyld walo 1559 fy are why to How War frage palacions to to when pur bourg prove my figur & four van wylny jand Washcauset on med? fy Ir from Dos voup wy con Minog yo a res brifter such and ser var Minga

on eadmodner pe pumende ze hypan noldon bæc hi bonne serped an hir milite rpa mucele vyolicon rpa micclum rpa hi nu heopa rpupan to hyp sebylde nellas zebizan. Dar popo rynd ze cpedene be pam pidep copenum achentilias papopo pe paze cope nanzernermad. Se hælend cpæd. bonne bar punopa on sinnas ahebbad fonne coppe hearda Jbehealdap ron han de copen aligredner zenealed Spylce he prucelice by se copenan manode. Jonne mid_ dan eapser pita ze lom læcad bonne oza bær mýcelan somer bid ar coped. ahebbad ponne coppe hearda fir Hadrad on copput mode rop pi ponne re middan capid bid fe endod pe ze nu luxedon ponn bid ze hende reo alizedner be se rohon. On halsu se pricu bid se lome hearod se rec ron hær manner mode ron dan he f hearod se purpad pam odpui lumum spa spa & mod ze dihr baze pohrar pea hebbad une hearda ponne pe une mod apæpad to ze rem ber heorenlican adeler bade sod lupias hi pind se ma_





Scanning and image preprocessing

- Quality of scanning companies varies considerably even if 'strict' standards (NL: Metamorfoze) are applied
- OCR has **very different demands** from web access and human reading requirements
- 'Simple operations' such as ink/paper separation are not at all solved fundamentally in image-processing science due to a pervasive 'chicken and egg' problem:

"In order to know the text, a system needs a good segmentation, but

a perfect segmentation into lines and words is only possible knowing the text "





Coarse overview current contents Monk, 71 documents, 53000 page scans

- KdK Dutch administrative 1893-1906
 20 books
- > Dutch Admirality 1760-1823 5 books
- Printed, Elzevirium, 1616
- Qumran scrolls (2463 scans)
- Middelduits (example document)
- Accounts, 1425 Gelria
- Schepenbank Louvain, 1421-1559, 3 books
- Colonial diary (1932)
- Municipal year report 1855
- > 20k illuminated initials
- Russian handwritten newspaper 1672
- Scholarly correspondence 1674-1682
- Chronicon Boemorum 1201
- Homiliarum Opatovicense 1150
- Resoluties Staten Generaal 1627
- Medieval charters 1300-1550 in 25yr periods
- ➢ Witch trial 1605

- ➢ 'Beowulf' and related, 4 books
- Ming Qing poetry , 7 books
- Arabic document, 291 scans
- Charlotte Perkins-Stetson diaries 1883- (682p.)
- Wittenbergsches (fraktur machine print)

Cooperations with:

- Huygens institute, Sorbonne,
- Harvard,
- o Stanford,
- Czech National Library,
- Dutch National Library, McGill Univ.,
- Utrecht Univ.,
- 0 Univ. Uppsala
- City archive Louvain,



Pattern recognition and machine learning

- Several years of experimentation
 - started in 2005
 - Monk was switched on, to autonomous mode, in 2009
- For developing and optimizing two functions:
- **Retrieval**: return images for a given keyword
- **Recognition**: return the most likely word given an image






Boosting performance

- Old mine-shaft
 elevator principle: *Fahrkunst*
- After having trained with method A to its 'max'
- An orthogonal method B can reach a higher performance
- > Then method A again, etc.
- Until the real asymptote is reached



Method A – Method B



THE SOLUTION:

24/7 learning over internet

> RuG HPC cluster

Labeling by humans

"Interactive Supercomputing"



Hit list computation







Lessons learned during Monk development

- A shape feature which is powerful for Retrieval may not be strong in Recognition!
- Feature B: hit list should provide nice, intuitive ranking in a satisfying 'hit list'
- Feature A: target word class should survive competition with the other word classes (emerging needle from the heterogeneous hay stack)



Examples of feature patterns for two words



van der Zant. Schomaker, Haak (2008). IEEE PAMI

Amsterdam

Groningen





Distance Measures for nearest-neighbour matching

• Hamming:

$$d = \sum_{i} |p_{i} - q_{i}|$$

Euclid

$$d = \sqrt{\sum_{i} (p_i - q_i)^2}$$

• Minkowski $d = \left(\sum_{i} |p_{i} - q_{i}|^{n}\right)^{\frac{1}{n}} \qquad bhattachary$ $d = \sqrt{1}$

Hausdorff

$$d = \max_{i} \left(\left| p_{i} - q_{i} \right| \right)$$

• Chi Squared (
$$\chi^2$$
)
$$d = \sum_{i} \frac{(p_i - q_i)^2}{p_i + q_i}$$

• Bhattacharyya $d = \sqrt{1 - \sum_{i} p_{i} \cdot q_{i}}$

Schomaker, Bulacu & Franke (2004) 9th IWFHR

21/33







Tilburg:

Orvolg van blady 433. 454 , Fraficken, Heringen an Bedrijven appt MD of April 11: 158 Tot verdaging der hesterling 330 In Lake het door / Magner to Filbudg ingesteld here Legen con betlint from Burg. a Holle dias demedite Waarbig hem vergunning is queigert tot oprichter Van done inrichting ter veretaarbigking van Heukwarken.





Does it work?

- > Do you need to label continuously?
 - Then transcription would have been better/easier
- > How good is the pattern recognition?
- > Do you need a supercomputer?
- > How many people are working 'back office'?
- > Can I send my collection to Monk?
- > What is the benefit, what is the output of the process?





#Labels, total as a function of 'days on-line', KdK 1903,1893,1897







#Labels, total as a function of 'days on-line', KdK 1903,1893,1897







18-09-2012 | 49

Monk

Jumps in the number of harvested word labels (y-axis) coincide with reaching a critical training dataset size (time axis)









Performance

 Retrieval precision increases monotonously with number of word examples

#training examples	What to expect in Top 50
1	Correct instance is good luck
5	A few correct hits
20	Many useful hits

- > Lexicon of 11k words: > 81% **recognition rate**, KdK
- > Lexwords with 50+ examples: > 95% correct
- > Idem, Cliwoc captain's logs: 93%





Good results: *bramzeijls* (top gallant's) hit list after training on just 2 examples

aramzeijlsbramzeijlsbramzeijlsbramzeijlsbramzeijlsbramzeijlsbramzeijlsGrowtługeGrowtługeGrowtługeGrowtługeGrowtługeGrowtługeGrowtługeBramzeijlsGrowtługeBramzeijlsGrowtługeGrowtługeGrowtługeBramzeijlsGrowtługeGrowtługeGrowtługeGrowtługeGrowtługeBramzeijlsGrowtługeGrowtługeGrowtługeGrowtługeGrowtługeBramzeijlsGrowtługeBramzeijlsGrowtługeGrowtługeGrowtługeBramzeijlsGrowtługeGrowtługeGrowtługeGrowtługeGrowtługeBramzeijlsGrowtługeGrowtługeBramzeijlsBrowtługeGrowtługeBramzeijlsGrowtługeBramzeijlsGrowtługeBrowtługeBramzeijlsBramzeijlsGrowtługeGrowtługeBramzeijlsBrowtługeBramzeijlsBramzeijlsGrowtługeBramzeijlsBrowtługeBramzeijlsBrowtługeBramzeijlsGrowtługeBramzeijlsBramzeijlsBrowtługeBramzeijlsBramzeijlsJoantugeBramzeijlsBramzeijlsBramzeijlsBramzeijlsBramzeijlsJoantugeBramzeijlsBramzeijlsBramzeijlsBramzeijlsBramzeijlsJoantugeBramzeijlsBramzeijlsBramzeijlsBramzeijlsBramzeijlsBramzeijlsBramzeijlsBramzeijlsBramzeijlsBramzeijlsBramzeijlsBramzeijlsBramzeijlsBramzeijls <td< th=""><th>alyestmon</th><th>bromtzeyls</th><th>branteyes</th><th>bromtheyes</th><th>bromteyes</th><th>bromteges</th><th>brantzeyes</th></td<>	alyestmon	bromtzeyls	branteyes	bromtheyes	bromteyes	bromteges	brantzeyes
GrownHungesGrownHungesGrownHungesGrownHungesGrownHungesGrownHungesGrownHungesbranzeijlsGrownHungesGrownHungesGrownHungesGrownHungesGrownHungesBranzeijlsBranzeijlsBrownHungesbranzeijlsGrownHungesGrownHungesGrownHungesGrownHungesGrownHungesBrownHungesBrownHungesbranzeijlsGrownHungesGrownHungesGrownHungesGrownHungesBrownHungesBrownHungesbranzeijlsGrownHungesGrownHungesGrownHungesBrownHungesBrownHungesbranzeijlsGrownHungesGrownHungesBrownHungesBrownHungesbranzeijlsGrownHungesGrownHungesBrownHungesBrownHungesbranzeijlsFrownHungesGrownHungesBrownHungesBrownHungesbranzeijlsFrownHungesBrownHungesBrownHungesBrownHungesbranzeijlsFrownHungesBrownHungesBrownHungesBrownHungesbranzeijlsBrownHungesGrownHungesBrownHungesBrownHungesbranzeijlsBrownHungesBrownHungesBrownHungesBrownHungesbranzeijlsBrownHungesBrownHungesBrownHungesBrownHungesbranzeijlsBrownHungesBrownHungesBrownHungesBrownHungesbranzeijlsBrownHungesBrownHungesBrownHungesBrownHungesbranzeijlsBrownHungesBrownHungesBrownHungesBrownHungesbranzeijlsBrownHungesBrownHungesBrow	bramzeijls	bramzeijls	bramzeijls	bramzeijls	bramzeijls	bramzeijls	bramzeijls
bramzeijls	bromtinges	bramteyls	bromteyes	brandeyes	bromteyld.	bramteyes	Squettmond.
GrownHaugh bramzeijlsGrownHaugh bramzeijl	bramzeijls	bramzeijls	bramzeijls	bramzeijls	bramzeijls	bramzeijls	bramzeijls
bramzeijls	bramtleyls	bromteyls	bromtyges	bramthayld	. bromthayls	bramteyls	bramtleyts
bramzeijls	bramzeijls	bramzeijls	bramzeijls	bramzeijls	bramzeijls	bramzeijls	bramzeijls
bramzeijls	bromteyes	bramteyld	Algert Imord.	Sycetmard.	branteyls.	Eromteyls.	bromteyes
bramzeijls	bramzeijls	bramzeijls	bramzeijls	bramzeijls	bramzeijls	bramzeijls	bramzeijls
bramzeijls	agustmord	bromteyes	bramteyls.	You Tayl	bramteyld	marzeyld.	bromteyes
bramzeijls bramzeijls bramzeijls bramzeijls bramzeijls bramzeijls	bramzeijls	bramzeijls	bramzeijls	bramzeijls	bramzeijls	bramzeijls	bramzeijls
bramzeijls bramzeijls bramzeijls bramzeijls bramzeijls bramzeijls	Jul tryl	Jul tryl	Spectmond.	mourtanged	Inordeyls	Equettmord.	Headmark
	bramzeijls	bramzeijls	bramzeijls	bramzeijls	bramzeijls	bramzeijls	bramzeijls

	Boek: KdK 1893	▼]		KdK 1893 n=47 Baron	KdK 1903 n=3 Baron
Hoofdstuktitel:					ni mi
Woordherkenner:	ToDo	•		Bur Barrow	Millio Alana
Woord:	Baron	Selectie uit hit list:	Niet (HUMAN)		1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
Aantal regels in hit list:	10	Eerste itemnr. in hit list:	0		ani -
Weergave:	Postzegels 🔻	Vernieuw:	Onderstaand overzicht	()) ana-	17thesa 🕥

ToDo / Baron / (no comment)

Items in hit list of Baron numbered: 0-2999. (now showing: 0-48)

Begin Prev Reload Next End

Baron	Baron	Baron	Baron	Baron	Baron	Baron
Baron	Baron	Baron	Baron	Baron	Baron	Baron
Baron	Baron	Baron	Baron	Baron	Baron	Baron
Baron	Baron	Baron	Baron	Baron	Baron	Baron
Baron	B oven	Baron	Baron	Baron	Boven.	Baron
Baron	Baron	Baron	Baron .	Baron.	Boven	Boven
Baron	Baron	Boven	Baron	Boren.	Baron	Baron Curon

	Boek: KdK 1893			Kd n=	IK 1893 47 Baron	KdK 1903 n=3 Baron
Hoofdstuktitel:						ni ni .
Woordherkenner:	ToDo				Ster O'S taken	Mar (1.9843-2
Woord:	Baron	Selectie uit hit list:	Alles	aller.		
Aantal regels in hit list:	49	Eerste itemnr. in hit list:	0	0.		Athe
Weergave:	Postzegels v	Vernieuw:	Onderstaand overzicht	-# RECOMPUTE	C Contraction	Maran D

ToDo / * / (no comment)

Items in hit list of Baron numbered: 0-2999. (now showing: 0-48)

Begin Prev Reload Next End

Baron	Baron	Baron	Baron	Baron	Baron	Baron
Baron	Baron	Baron -	Baron .	Baron	Baron	B too.
B woo	Baron	Baron	Baron	Baron	Baron	Baron
Baron	Baron	Baron	Baron	Baron	Baron	B Ween
Bron	Boven Baron	Baron	Baron	Boven	Boven.	Baron
Baron	Baron	Baron	Baron .	Born.	Boven	Boven
Baron	Baron	Boven	Baron	Booen.	Booen.	Baron

THE WORK SYSTEM / Lambert Benomaker

	Boek: KdK 1893	•			KdK 1893 n=11 Exploitatie	KdK 1903 n=1 Exploitatie
Hoofdstuktitel:	•					ALL O CLA
Woordherkenner:	ТоDo	•			Alaida Catilonialas	Mimphie Matorianie
Woord:	Exploitatie •	Selectie uit hit list:	Niet (HUMAN)	ste.		
Aantal regels in hit list:	10	Eerste itemnr. in hit list:	0	505	0.00.4	W. total
Weergave:	Postzegels •	Vernieuw:	Onderstaand overzicht	-# RECOMPUTE	Supplication	Unitoritatie S

en

Exploitatie

Exploitatie

ToDo / Exploitatie / (no comment)

Enpl

Exploit Lap

Exploita

apl

Exploit

Capl

laplo

en

Exploitatie

Begin Prev Reload Next End Items in hit list of Exploitatie numbered: 0-1025. (now showing: 0-48)

Exploitati

Inj Exploitatie

ritabre	Inploitate	laplostated	l'aploitatie	enploytate	enploytate	Enploitable)	
	Exploitatie	Exploitatie	Exploitatie	Exploitatie	Exploitatie	Exploitatie	
ortated	Enploitates	l'aploitatie	enfortatio	enfortatio	Inplantation	Inplastation	
	Exploitatie	Exploitatie	Exploitatie	Exploitatie	Exploitatie	Exploitatie	
interen	aploster and	haplostatie	enployate	enployate	Duplotitie	Corplostatie	
	Exploitatie	Exploitatie	Exploitatie	Exploitatie	Exploitatie	Exploitatie	
interror	Explortates	aplostera	Poploitatie	la filostation	Cofelostations	Explortation	
	Exploitatie	Exploitatie	Exploitatie	Exploitatie	Exploitatie	Exploitatie	
telle	nplostatio	Saplantate	, aplitera	Exploitate	enploitate	enploitate	
	Exploitatie	Exploitatie	Exploitatie	Exploitatie	Exploitatie	Exploitatie	
lotate	Infortate	Corplostates)	Inplantatic	Infloitates	Inplortates	onplotaties	

Exploitatie

Exploitatie

en

The Monk system / Lambert Schomaker

Exploitatie

Explortatio

terrow

Exploitatie

Exploitatie

en

Exploitatie

Qa,

Exploitatie

At rank 343 in the list: pink newlings





Items in hit list of ende numbered: 0-3999. (now showing: 0-48)

Begin Prev Reload Next End





Monk usage example: May 2013 Qumran scrolls: Daniel Stoekl (Sorbonne), Mladen Popovics (Groningen)







Qumran scrolls: 2400 photographs

- > Using **Monk** for character labeling
- > With Daniel Stoekl and Mladen Popovics
- > Using its 24/7 machine-learning cycle:
 - Label → Train
 - \rightarrow Label some More \rightarrow Train
 - → Easily label Many
- Thousands of characters 'mined' out of the Qumran collection of photographs in just two weeks, with very little effort in human labeling



א וראוריש לוריר ופוא ושרא או	אנה עטרונגאות שנורה	יזאה אריבי איז ארוביב with ybas; C
?noLineLabel navis-Qlrug-Qu	mran_extr09_2181-line-002-y1=326	6-y2=515.txt?
	Recognizer hypotheses:	re-Recognize Word
	Alef	0.515
K X	Gimel	0.555
	Paleo_Tsadi	0.582
Alef	Greek Alpha	0.597
Save label	@PAPYRUS_EDGE	0.607
	Yod	0.614
<u></u>	@Top_Of_Waw-triangular-close	ed 0.618
	Paleo_Shin	0.657
Next in [Sordex/Alef]: 'Alef'	@FRAGMENT_EDGE	0.662
	Delta	0.675



Back to ov	erview of trained words Book: Qumran: Qumran scrolls		Qumran scrolls n=2311 Shin	Qumran scrolls n=2311 Shin Separability
Word recognizer: Hit list, word:	Sordex Shin	Select from hit All	set recompute Within'	
Zoom:	1.0x •	First itemnr. in hit list:	*	
View:	Thumbnails (orig)	Refresh: Hit list below	-# READY	

Items in hit list of Shin numbered: 0-3999. (now showing: 0-99)

Begin Prev Reload Next End











Allograph harvest in Qumran collection

244 NY 字 27 X6 XK 匀 101 👕 X6 3W yy v 🐖 NW 🏲 WY LK 🗤 2 yper 11/ 11 - - 11 - N II 20 - が い コ ツ る - が 65 NAN YA MA 101 -24 T IV M T Y \sim -104 100 MR 72 121 Terr 25 1 52 2-19





Allograph harvest 33k labels in NWO/MPS project with Huygens institute

Back to ove	rview of trained words		MPS 1300 ["] MPS 1300 ["] n=151 @a_charter_1300 n=151 Separability @a_charter_1300	
	2 Book: 1300refset: MPS 1	300 [*] 🔻		
Word recognizer:	YearLetters	•		
Hit list, word:	@a_charter_1300 •	Select from hit list:	set recompute 'within'	
Zoom:	1.0x •	First itemnr. in hit list:		
View:	Thumbnails (orig)	Refresh: Hit list below		
tems in hit	list of <u>@a_charter_1300</u> numbe	ered: 0-3999. (now showing $a_{3}^{2} = a_{10}^{2} = a_{11}^{2} = a_{12}^{2} = a_{13}^{2} = a_{14}^{2} = a_{$	$\begin{array}{c} \text{Ig: } 0-99 \end{array}) \begin{array}{c} \text{Is } 15 \\ \text{If } 16 \\ 16 \end{array} \\ \begin{array}{c} 16 \\ 17 \end{array} \\ \begin{array}{c} 16 \\ 18 \end{array} \\ \begin{array}{c} 19 \\ 18 \end{array} \\ \begin{array}{c} 19 \\ 19 \end{array} \\ \begin{array}{c} 20 \\ 20 \end{array} \\ \begin{array}{c} 21 \\ 21 \end{array} \\ \begin{array}{c} 22 \\ 22 \end{array} \\ \begin{array}{c} 22 \\ 23 \end{array} \\ \begin{array}{c} 23 \\ 24 \end{array} \\ \begin{array}{c} 24 \\ 25 \end{array} \\ \begin{array}{c} 25 \\ 26 \end{array} \\ \begin{array}{c} 25 \\ 26 \end{array} \\ \begin{array}{c} 25 \\ 26 \end{array} \\ \begin{array}{c} 25 \\ 25 \end{array} \\ \begin{array}{c} 25 \\ 26 \end{array} \\ \begin{array}{c} 25 \\ 25 \end{array} \\ \end{array} \\ \begin{array}{c} 25 \\ 25 \end{array} \\ \end{array} \\ \end{array} \\ \begin{array}{c} 25 \\ 25 \end{array} \\ \end{array} \\ \end{array} \\ \end{array} \\ \begin{array}{c} 25 \\ 25 \end{array} \\ \end{array} \\ \end{array} \\ \end{array} \\ \end{array} $ \\ \begin{array}{c} 25 \\ 25 \end{array} \\ \end{array} \\ \end{array} \\ \end{array} \\ \end{array} \\ \end{array} \\ \end{array} \\ \begin{array}{c} 25 \\ 25 \end{array} \\ \end{array} \\ \end{array} \\ \end{array} \\ \end{array} \\ \end{array} \\ \end{array} \\ \end{array} \\ \end{array} \\ \end{array}	I 32
A 33 A 34 A 35			A A	65
			1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2	98
99		har	vested during March 2014	

Yes, also Chinese characters (handwritten & woodblock, Grace Fong & Harvard Yenchin collection)





Necessary metaphor for the Monk process

- Not: a stream-lined mechanical factory
- Yes: crystal-like growth, arborization,
- Randomly over a large surface
- Energy comes from:
 - human labels
 - machine re-compute actions



















dBid	Year	Short name	Pages \$	Scans	BookID (url base)	.reindex .files	.IDs+txt	/Index	/Sordex	Words	Days
1	1893	KaK 1803 /*1 biz. 1-1316	1-1384	1384	navis-NL-HaNA 2.02.04 3950/	1 days 1 month	s 1 months	1 davs	1 davs	12984 (=)	205.3 (+1.0)
2	1897	KdK 1807 blz. 781-1020	781-1632	852	navis-NL-HaNA 2.02.04 3955/	5 days 1 month	s 1 months	1 weeks	1 months	10514 (=)	264.4 (+1.0)
3	1898	KdK 1898 blz. 1-842	1-842	842	navis-NL-HaNA 2.02.14 7813/	5 days 1 month	s 1 months	1 weeks	1 weeks	8939 (=)	225.8 (+1.0)
4	1899	KdK 1800 blz. 1-022	1-922	922	navis-NL-HaNA 2.02.14 7815/	1 days 1 month	s 1 months	1 months	1 months	9210 ()	205.0 (+1.0)
5	1899	KdK 1899 blz. 802-1782	862-1794	933	navis-NL-HaNA 2.02.14 7816/	5 days 1 month	s 1 months	1 weeks	1 weeks	9136 ()	183.7 (+1.0)
6	1901	KoK 1901 biz. 1-1040	1-1099	1099	navis-NL-HaNA 2.02.14 7819/	5 days 1 month	s 1 months	1 weeks	2 davs	9999 (=)	195.2 (+1.0)
7	1901	KdK 1901 blz. 1041-2024	1041-2038	998	navis-NL-HaNA 2.02.14 7820/	5 days 2 week	s 2 weeks	1 weeks	2 weeks	9286 ()	181.0 (+1.0)
8	1902	KdK 1902 blz. 974-2005	982-2073	1092	navis-NL-HaNA 2.02.14 7822/	5 days 1 month	s 1 months	2 months	1 weeks	6485 (=)	197.3 (+1.0)
9	1903	KdK 1903 blz. 1-1094	1-1094	1094	navis-H2 7823 0001-1094/	4 days 1 month	s 1 months	4 days	4 days	9013 ()	504.3 (+0.9)
10	1903	KdK 1903 blz. 1200-1346	1299-1346	48	navis/	1 years 9 month	s 9 months	1 weeks	1 weeks	1411 (-)	116.3 (+1.0)
11	1903	KdK 1903 blz. 1348-1401	1348-1401	54	navis2/	1 years 9 month	s 9 months	3 months	3 weeks	1861 ()	116.3 (+1.0)
12	1904	KdK 1904 blz. 1-998	1-1064	1064	navis-NL-HaNA 2.02.14 7825/	4 days 1 month	s 1 months	4 days	4 days	7624 (=)	171.5 (+1.0)
13	1904	KitK 1004 biz 000-2008	999-2110	1112	navis-NL-HaNA 2.02.14 7826/	5 days 1 month	s 1 months	1 weeks	4 days	8662 ()	168.7 (+1.0)
14	1905	KdK 1005 blz, 1-1120	1-1190	1190	navis-NL-HaNA 2.02.14 7827/	5 days 1 month	s 1 months	1 months	1 weeks	7808 (=)	174.0 (+1.0)
15	1905	KdK 1005 blz, 1121-2247	1121-2257	1137	navis-NL-HaNA 2.02.14 7828/	9 hours 1 month	s 1 months	1 weeks	1 weeks	5895 (=)	145.7 (+1.0)
16	1906	KdK 1005 blz, 1161-2330	1161-2334	1174	navis-NL-HaNA 2.02.14 7830/	1 days 1 month	s 1 months	2 weeks	4 days	14891 (=)	260.2 (+0.8)
17	1809	KdK 1800, scannr, 130-730	525-912	388	navis-ni-hana h26506 0556 0001-0739/	1 years 2 week	s 2 weeks	1 vears	1 vears		
18	1810	KdK 1810 scannr, 100-822	1-624	624	navis-ni-hana h26506 0557 0001-0822/	1 years 2 week	s 2 weeks	2 months	1 months	687 (=)	77.1 (+0.9)
19	1903	KoK 1003 reg this 1-1004 XMI -layout	1-1094	1094	navis-H2 7823 0001-1094b/	NonEx NonE	x NonEx	NonEx	NonEx		(
20	1779	Adm 1770 - 177 1177 /*1	1-538	538	clwoc-Adm 177 1177/	9 hours 2 week	s 2 weeks	2 weeks	1 months	1192 (=)	179.6 (+1.0)
21	1779	Adm 1770 - 177 1180	1-205	205	clivoc-Adm 177 1189/	1 years 9 month	s 9 months	1 weeks	1 weeks	295 (=)	97.2 (+1.0)
22	1783	Adm/Fam 1130 38 Jaar 1783	1-90	90	clwoc-Fam 1135 38/	8 months 8 month	s 8 months	2 months	5 days	1325 (=)	854 (+1.0)
23	1760	Adm 177 1157 Jaar 1760	1-558	558	clwoc-Adm 177 1157/	9 months 8 month	s 8 months	3 months	1 days	1055 (=)	1017 (+1.0)
24	1823	Mar. D23005 4063 Jaar 1823	1-416	416	olivoo-Mar D23905 4053/	1 months 4 month	a A months	1 months	5 days	670 (=)	881 (+1.0)
25	1020	(Ibbo Emplus REH	1-70	70	ubrua-libbo Emplus REH 1616 Omslag p72/	2 years 2 week	s <u>+ monore</u> s - 9 weeks	NonEy	NonEx	0/0(-)	00.1 (+1.0)
25	1947	Outran scralls	1-2/63	2463	Olicio-Oumran extribi	5 days 2 month	s 2 months	5 days	5 days	(23.(-)	72.9 (41.0)
20	1400	Minishute	1-4	2400	navis_marilaraLtart_li arxiani	7 months 4 month	s <u>A months</u>	1 months	R weeks	423 (-) 811 (=)	133.0 (±0.0)
28	1425	Celderse rekening 1/25	1-20	20	Caldaroh, rekeningen, 1725/	A months A month	s Amonths	2 weeks	1 weeks	1645 ()	122 / (-3.2)
20	1420	Schenenhank Lewen 1421 P1	1-20	794	PAL7316/	1 days 1 month	s 1 months	1 months	3 weeks	4044 ()	227.6 (41.0)
30	1457	Schepenbark Leuven 1457	1-960	860	SAL7751/	7 months 6 month	s 6 months	2 months	2 weeks	1/33 ()	107.5 (±1.0)
34	1559	Schepenbark Leuven 1457	1-1373	1373	CALTAST	7 months 6 month	s <u>6 months</u>	2 months	2 weeko 3 dave	1450 ()	107.5 (±1.0) 80.5 (±1.0)
30	1779	Adm 1770 - (bst)	100,070	1070	oliupa Adm 177 1177b/	1 months 1 months	s <u>timontes</u>	2 months	11 months	274 ()	777.7 (41.0)
32	1912	P // Astro - 1032 (*1	1.3	113	D/Letto-1932	1 days 1 month	s 1 months	1 weeks	A months	275 ()	120.5 (±1.0)
24	1002	CV Applaged m - 1955 (*)	1-020	220	CV-Applaceterr \$140	E dave if month	s 1 months	2 months	2 weeks	1907 ()	125.0 (±1.0) 205.2 (±1.0)
25	1000	Socialitata	1 20527	205	Claria National	<u>o uaya</u> i munun 1 yaara - 2 waak	o <u>i munuto</u> O weeks	<u>2 monute</u> MonEx	AleeEv	1007 ()	250.2 (+1.0)
20	1620		1-20027	20027	Macanual Edge data da	I years 2 week	o <u>2 weeko</u>	2 months	INUTEX	210 ()	67 A (14 M)
30	1672	Cuper Brown	1-202	202	Cuper Brauel	<u>4 montris</u> 4 montri 2 weeks 2 montri	s <u>4 monuns</u>	2 montris	E dave	319 (=) 1405 (=)	101.2 (11.0)
30 30	1074	Cuper_braun	1-91	000	Cuper Diably	<u>o weeks</u> 2 month		2 weeks	<u>o uaya</u>	1400 ()	101.3 (±1.0)
30	1002	Cuper_Altri	1-922	922	Cuper Alling	d daus - 2 baus	s <u>i monuis</u>	1 monute	1 weeks	1900 (=)	02.2 (+1.0)
39	1201	Chronicon Boemorum 1-151	1-151	151	RNMP-VIII P 69 2020/	<u>1 days</u> 3 hour	s <u>o nours</u>	1 days	1 days	1179 (=)	20.1 (-0.0)
40	1627	HSG 1027 [] scan 5101.1.	1-90	90	RSG 102//	3 weeks 4 week	s <u>4 weeks</u>	4 weeks	<u>1 days</u>	1/54 (=)	10.7 (+1.0)
41	1300	MPS 1300 []	1-90	90	MPS1300	10 hours 1 day	s <u>10ays</u>	9 hours	10 minutes	007 (+3)	9.5 (+0.7)
42	1325	MPS 1325	1-141	141	MPS1325/	19 nours 1 day	s <u>1 days</u>	9 nours	36 minutes	900 (=)	11.4 (+1.0)
43	1350	MPS 1350	1-69	89	MPS1350/	25 seconds 13 hour	s <u>13 nours</u>	9 nours	<u>1 days</u>	1091 (+18)	4.7 (+0.9)
44	1375	MPS 13/5	1-64	84	MPS13/5/	13 nours 1 day	s <u>1 days</u>	9 Hours	4 minutes	/10 (+1)	9.9 (+1.0)
45	1400	MPS 1400	1-196	196	MPS1400/	7 nours 8 hour	s <u>8 hours</u>	a nours	2 hours	990 (+16)	9.3 (+0.9)
40	1425	MPS 1425	1-198	198	MPS1425/	24 seconds 18 hour	s <u>18 hours</u>	9 Hours	10 hours	991 (+10)	8.3 (+0.9)
47	1450	MPS 1450	1-245	245	MPS1450/	4 nours 1 day	s <u>1 days</u>	9 nours	2 seconds	973 (+2)	9.9 (+0.9)
48	1475	MPS 1475	1-205	205	MPS1475/	1 days 15 hour	s <u>15 hours</u>	9 hours	<u>1 days</u>	1015 (+1)	9.8 (+0.9)
49	1500	MPS 1500	1-141	141	MPS1500/	O seconds 5 hour	s <u>5 hours</u>	5 hours	16 hours	996 (=)	9.9 (+1.0)
50	1525	MPS 1525	1-115	115	MPS1525/	O seconds 2 hour	s <u>2 hours</u>	2 hours	<u>3 minutes</u>	985 (+13)	10.6 (+0.9)
51	1550	MPS 1550	1-113	113	MPS1550/	24 seconds 1 day	s <u>1 days</u>	9 hours	17 minutes	987 (+23)	9.8 (+0.7)
52	1605	HulsBergh_0214_7268	1-76	76	HulsBergh 0214 7268	1 weeks 1 week	s <u>1 weeks</u>	1 weeks	1 weeks	558 (=)	4.8 (+1.0)

collection on 02-04-14 52 'books' and growing fast

Monk

On 28-10-14 72 'books'

Static index example: Chronicon Boemorum (1201 AD, Cosmas)

Monk search engine for handwritten words - KNMP collection -

KNMP/CB - Chronicon Boemorum, Cosmas, p. 1-151 boek KNMP-VIII_F_69____2C2O(1201).

L. Schomaker, ALICE, Rijksuniversiteit Groningen/ Tomas Klimek, National Library of the Czech Republic, Monasterium (Last update of these static tables: Thu Jan 30 17:41:00 CET 2014)

<u>A B C D E F G H I J K L M N O P Q R S T U V W X Y Z @ abcdefghijklmnopqrstuvwxyz</u>

Ø	<u>Piura</u>	<u>aspicias</u>	<u>dedisse</u>	<u>familia</u>	<u>huc</u>	<u>memorat</u>	<u>patro</u>	<u>romano</u>	<u>uolo</u>
6	<u>Placet</u>	<u>aspicio</u>	<u>dedita</u>	<u>famula</u>	<u>hui</u>	<u>memorati</u>	<u>patroni</u>	<u>rota</u>	<u>urbe</u>
@ampersand	<u>Plura</u>	<u>assistebant</u>	<u>defensio</u>	<u>fatis</u>	<u>huic</u>	<u>memoria</u>	<u>patronis</u>	<u>rubentes</u>	<u>urbes</u>
	<u>Podium</u>	<u>auctoritas</u>	<u>defraudare</u>	<u>fecerit</u>	<u>huma</u>	<u>memorie</u>	<u>patu</u>		<u>urbio</u>
٨	<u>Polonias</u>	<u>auctoritate</u>	<u>delphi</u>	<u>fecunt</u>	<u>humani</u>	<u>menia</u>	<u>pauca</u>	S	<u>urbis</u>
A	<u>Polonie</u>	<u>auderem</u>	<u>denario</u>	<u>felicit</u>	<u>humanitatis</u>	<u>meo</u>	paucis	-	<u>usus</u>
	<u>Poloniensi</u>	audientes	<u>denegare</u>	<u>felix</u>	<u>hunc</u>	<u>minor</u>	<u>pauco</u>	sacra	<u>utero</u>
Adhec	<u>Polonios</u>	<u>aura</u>	<u>depilet</u>	<u>femina</u>	<u>hystoriaca</u>	<u>mittere</u>	<u>pax</u>	sacris	<u>utro</u>
Adhuc	<u>Pomora</u>	aure	<u>deposuit</u>	<u>ferata</u>		<u>modis</u>	pectore	salutat	
<u>Adque</u>	<u>Post</u>	<u>auro</u>	descendere	<u>ferre</u>	i	<u>molestia</u>	<u>pectu</u>	salute	v
Anno	<u>Postera</u>	<u>auru</u>	<u>deunde</u>	<u>ferro</u>	•	monens	<u>pecunia</u>	sanati	•
Annoncanta	<u>Praga</u>	<u>ausis</u>	<u>diaboli</u>	<u>feruus</u>	iacet	<u>monete</u>	<u>pede</u>	saniens	vocabit
Aut	<u>Pridie</u>	<u>aut</u>	<u>diadema</u>	<u>festina</u>	ibi	mons	<u>pena</u>	sciencia	vundat
		<u>auta</u>	<u>dicens</u>	<u>festis</u>	ille	mora	<u>penas</u>	scient	variaat
В	Q		<u>dicere</u>	<u>fiat</u>	illi	<u>mordax</u>	<u>peticione</u>	scientes	
	-	b	<u>dictare</u>	<u>fide</u>	illicitis	<u>mortalis</u>	<u>pia</u>	scriptura	
Babenberk	Qua		<u>dicto</u>	<u>fidele</u>	illos	<u>morte</u>	<u>pla</u>	scripture	
Bawarios	Quare	barba	<u>dicunt</u>	fideliores	illuc	<u>mortis</u>	placeat	seccessit	
Benedict	Que	bawaria	<u>didit</u>	<u>fidelitate</u>	illustri	<u>mortuos</u>	placent	secunda	
<u>Boemie</u>	Qui	bello	<u>die</u>	<u>fides</u>	impetrare	<u>mortuus</u>	<u>placita</u>	sed	
Boemii	Quid	bene	<u>dignitas</u>	<u>filia</u>	impetu	<u>mox</u>	<u>placuit</u>	sede	
<u>Boemo</u>	Quinto	benediccione	<u>dignitate</u>	<u>filio</u>	infacto	<u>moxq</u>	<u>plaustris</u>	sedis	
Line transcriptions, actual (literal and text-rendered @Monk-codes)



Note: some @Monk-codes for word shapes have not been codified here, the user can click on them to do that

Line transcriptions, decontracted, literal and text-rendered @Monk-codes



Line transcriptions, word-by-word translation from Latin to Dutch

onk menu:	▼ Go	Monk - () Stadsarchief Leuven, Scher	penbank 1421 [*]	Ischomaker Log out
st=[Jair (MMonk,version:),]			
(Transcription of text lines, scan= <u>7</u>	➡,	
an: 7 Page: [recto] 4				
[bovenrand-bo [marge-boven] •ltem <i>Renerus</i> Hendrik en (aa	ek] deveri Lemmens de∣v n)Johannes deveri Le	ran Stenhuffel zoon van Peter voorheen Lemmens he emmens zijn(bez.) <i>fratribus</i> Wim de∣van Ophem zoon	eft erkend zich,dat hij,dat zij moeten heer van Johannes voorheen	
de van Ophem	<i>militer</i> en Hendrik Le	ammens zoon <i>Reveri quandam</i> Lemmens <i>fratris predi</i>	<i>icti?</i> van Peter	
100 ponden gr	ote van Tours <i>veterui</i>	<i>n</i> goede goederen <i>legales</i> bij tot naar op waarschuw	ıing en preterea famulum conductum Al	bus Graven
juni 27 •Item Johannes genoemd Jaco <i>fratribus filiis v</i> ponden <i>grosso</i> •Item Johannes <i>ac</i> huis en hof en goed goederer bewaren en <i>re</i> Johannes Sam <i>securitatem</i> ge	s genoemd Vos zoon I ps de van Steenhuffle van Peter voorheen Le orum van Tours oude s Vos bakker in aanwe tuin met zijn(bez.) aar eren van Johannes de n van dezelfde Willem levare de van <i>illo mo</i> pain <i>presbyter quo</i> bi enoemd Quintinus de	Nycolai genoemd Vos de van Steenhuffle heeft erken <i>presbytero Renero</i> Hendrik en Johannes genoemd(p mmens en Hendrik de van Horenbeke de van Steenh goede goederen en <i>legales</i> bij tot naar op waarschur zigheid heeft beloofd Quintinus de van Valkenborch horigheden van dezelfde Quintinus gelegen in Dorps van Troembeke <i>exeuntes</i> met een klein kleine hof tu Nage en goed goederen aan meester Hubertus voor <i>dio siliginis mensuris</i> Leuvense <i>annue pensionis qua</i> j tot naar op leven <i>suam</i> heeft bij tot naar op goed gov van <i>premissa facundia prefatus</i> Johannes Vos 3 goud	d zich,dat hij,dat zij moeten aan de heer pl.) Lemmens de van Steenhuffle huffle 100 wing en <i>preterea famulum conductum</i> h bakker straat tussen goed goederen Willem Na uin in Bakeleynstraat tussen heen de van <i>aquis presbyteri indempn</i> <i>am</i> heer oederen voornoemde en bij tot naar op den	r (aan)Johannes noofdelijk aan dezelfden gel <i>es penitus</i> <i>maiorem</i>

Word-shape codification: attributes of a literal or @Monk-code



provisional transcriptions. The text shown comes either from human line transcription or (human) confirmed recognized words (see ...)

```
囲
Cuper Braun 0014
6 [R] IOANNI BRAVNIO (continued) ut patet ex Liv. ...
 ... et ... 3 ...
tintinnabula illa summi sacerdotis ex ore fuisse, quae tü fuerunt ex auro
 ... B ... quibus ... u ...
vocant et quia oris in saevis Bracchii, in quibus et nutrices @GR chalko-
 ... quia ... qui ... a ...
@GR drutai sacro sine dubio noie vocantur; plurimus usus, strepiturque
... g ... ipse doces ... p ... fuisse ...
undem Deü coli= et ipse doces p 96. Solenne Ethicus fuisse Bacchi
sacra celebrare, oneis crepidis et cymbalis ejusmode calciamenta
habuisse Empedoclem, teste Tertulliano c.q. de pallio; i Andque probas
... @Epsilon ... s ...
auctoritate Plutarchi; apud quem tü l. 6.Sym. quë laudas, nihil tale
auctoritata Plutarchi, apud quem tü 1.6.Sym, que laudas, nihil tale
repperi. Sane tintinnabulorü in vestibus, ut apud Judaeos ...., usum
apud veteres fuisse nullibi reperio. Lego quidem apud M. Paulum
venetum 1.2.c.23 cursores Tartarü Regium habere, cinetoria
 ... nihil ... ad ...
sonoris tintinnabulis plena gestantes, sed hoc nihil ad rem
 ... n ... a ...
unde o abs re est, quod putem comparatione H m
... @o - ... a ... quod ... ut ... u ... m ...
strepituum, quos .... tintinnabulis, Bacchantes oreis cymbalis edunt
 ... quo ... s ... n ...
 ... b ...
instituere, nec illü dixisse Sacerdotë Bacchi et tintinnabula habuisse,
illa autem @GR ho deiknumenos en tois enantiois tou meteooron thursos
 ... a ... bene ... n ...
```





Conclusions

- > Great progress is made in handwritten text retrieval in recent years
- The concept of iterative refinement by human labeling & computing really works
 - → interactive machine learning over Internet
- Image preprocession and layout analysis are important, still crude and not fundamentally solved
- > Image-quality standards (NL norm: *Metamorfoze*) do not guarantee high-performance handwritten text recognition
- Approach works for several historical periods (1200, 1400, 1600, 1800): but each needed some minor tweaking of the image preprocessing
- > Think big!









How to improve on human and machine transcriptions?

- The legal transactions in the acts of the Leuven Alderman's rolls, 1421: Towards a formal semantic model of acts (Marianne Ritsema van Eck, august 2011). Presented at Digital Humanities 2012
 - → experts fight over the transcription
 - \rightarrow students need to know more than paleography
 - \rightarrow the machine needs a model for top-down reasoning





Semantic modeling and layout of acts: example: 'IOU'



Fig. 4.9 (MS page V64.1 / scan 128) The relevant regions of interest, that correspond to the structural elements the act is composed of, are indicated on the MS page with bright colours.

Monk & the Schepenbank – Lambert Schomaker



Spatial modeling: Layout vs Semantics in the acts of the Schepenbank





Digging into Data / Global currents -- US/Canada/NL project with Mohamed Cheriet of Ecole de technologie supérieure(Montréal) Andrew Piper (McGill University) (principal investigator) Elaine Treharne (Stanford University)

ira, cum dimidiato fulget orbe, non est globo fimi nam proc fe fert. Nam how funt ab una parte p forma circulari, al hunc modum ford eandem figuram bullas habuille infantium; wilk ank puches ex circulo annexo alleri angulorum Avel B. et i lavorium in ostiis , fit et multorum hodie fimi hilfimum est. BVLLE fuerint nominatoz. Ino et veteres hanc ob caufam DVILAS Lunulas vel pe



Digging into Data / Global currents -- US/Canada/NL project with Mohamed Cheriet of Ecole de technologie supérieure(Montréal) Andrew Piper (McGill University) (principal investigator) Elaine Treharne (Stanford University)

ira, cum dimidiato fulget orbe, non est globo fimi nam proche fert. Nam how funt ab una parte p forma circulari, ad hunc modum forc eandem figuram bullas habuille infant Wijk ande peches ex arculo annexo allen angubrum Avel B. et il lavorum in ostris , ut et multorum hodue Armi hijfimum est BVLLE fuerint nominator. Ino et veletes han ob caufam DVLLAS Lunulas vel p





Conclusions

- > Great progress is made in handwritten text retrieval in recent years
- The concept of iterative refinement by human labeling & computing really works
 - → interactive machine learning over Internet
- Image preprocession and layout analysis are important, still crude and not fundamentally solved
- > Image-quality standards (NL norm: *Metamorfoze*) do not guarantee high-performance handwritten text recognition
- Approach works for four historical periods (1200, 1400, 1600, 1800): but each needed some minor tweaking of the image preprocessing
- > Think big!