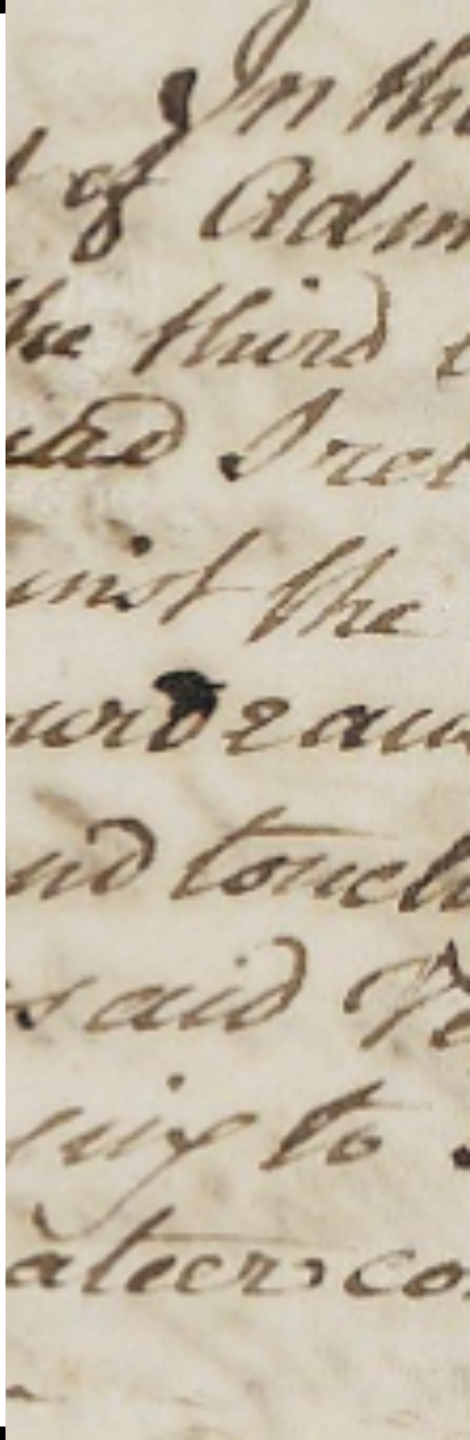
MONK IN PRACTICE:

Indexing Heterogenous Handwritten Collections

Anna CACERES, Andreas WEBER, Lambert SCHOMAKER

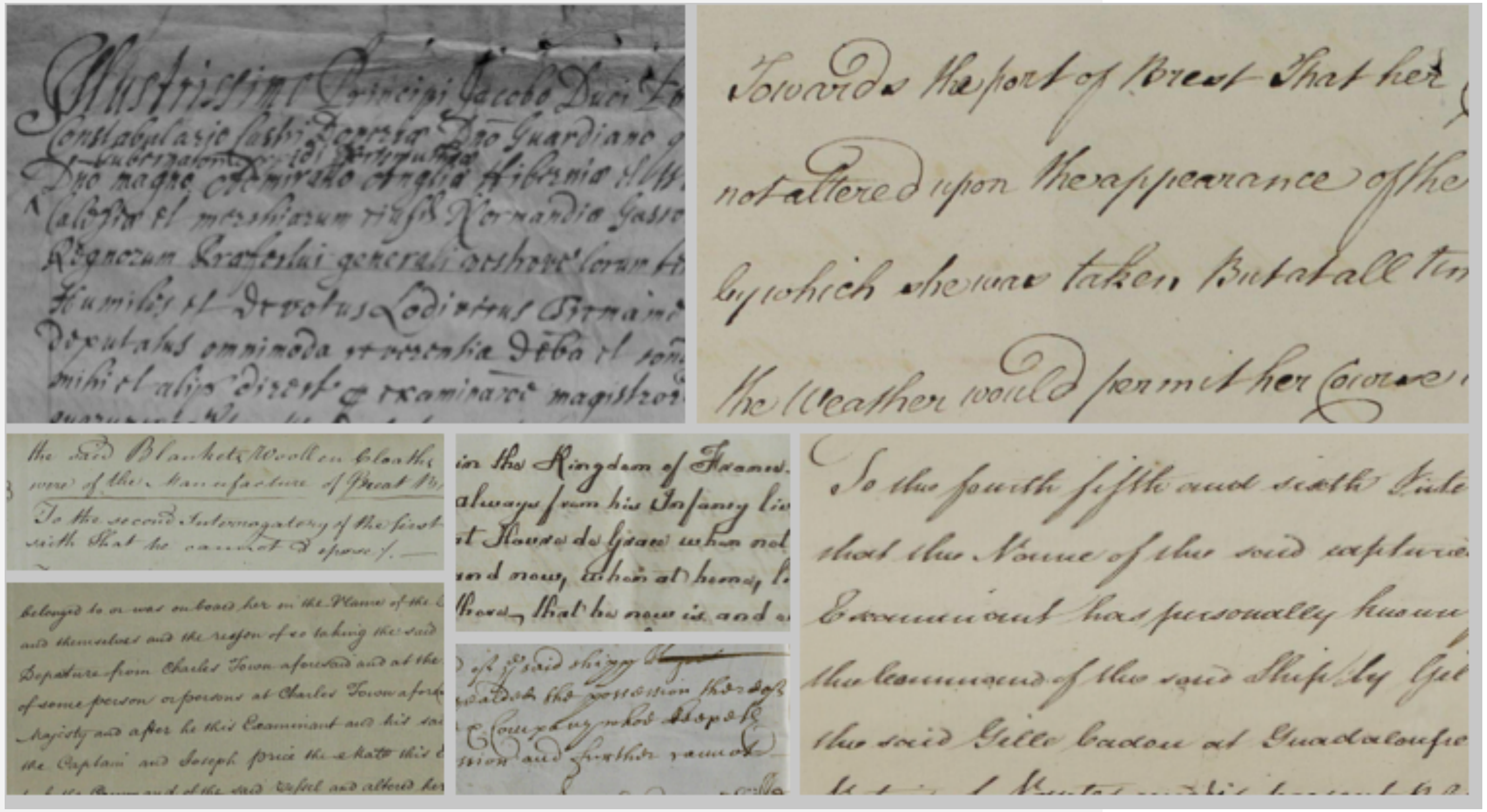
DOI: 10.5281/zenodo.3860811

JUNE 2020



INTRODUCING THE PRIZE PAPERS

Archives of British maritime juridical body the High Court of the Admiralty 1652-1815. Consist of two parts: interrogations and seized documents. Benefits of interrogations for HRT as follows:



Illustrissimo Principi Jacobo Duci Ebor
Constabulario Castri Dovoriæ Dño Guardiano
Dño magno Admirallo Angliæ Hiberniæ
Calisiæ et marchiarum earundem Normandiæ Gasc
Regnorum Præfecto generali
Humilis et devotus
deputatus omnimoda et
mihi et aliis direct examinare magistros

Towards the port of Brest That her
not altered upon the appearance of the
by which she was taken But at all tim
the Weather would permit her Course

the said Blankets Woollen Cloaths
were of the Manufacture of Great Br
To the second Interrogatory of the first
saith that he cannot depose./—

in the Kingdom of France.
always from his Infancy liv
at Havre de Grace when not
and now, when at home, l
there, that he now is and

To the fourth fifth and sixth Inte
that the Name of the said capture
Examinant has personally known
the command of the said Ship by Gil
the said Gille badou at Guadaloupe

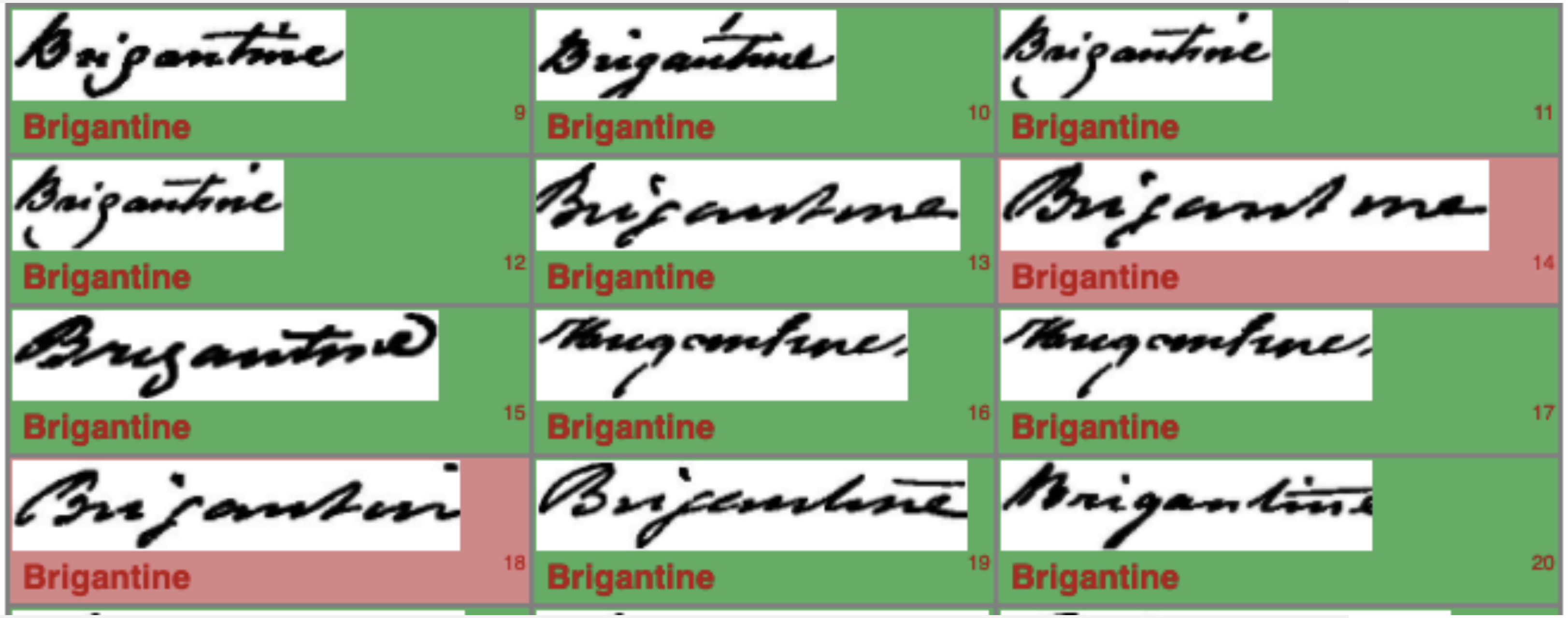
belonged to or was on board her in the Name of the
and themselves and the reason of so taking the said
Departure from Charles Town aforesaid and at the
of some person or persons at Charles Town aforesd
Majesty and after he this Examinant and his sai
the Captain and Joseph Price the Mate this
the Command of the said Vessel and altered her

of ye said shipp
the possession thereof
Company whoe
and further

- Highly standardized in content, all interrogations ask a set of roughly 33 questions
- Heterogeneous in script style, both due to use of different notaries and writer fatigue.
- Heterogeneous states of preservation.

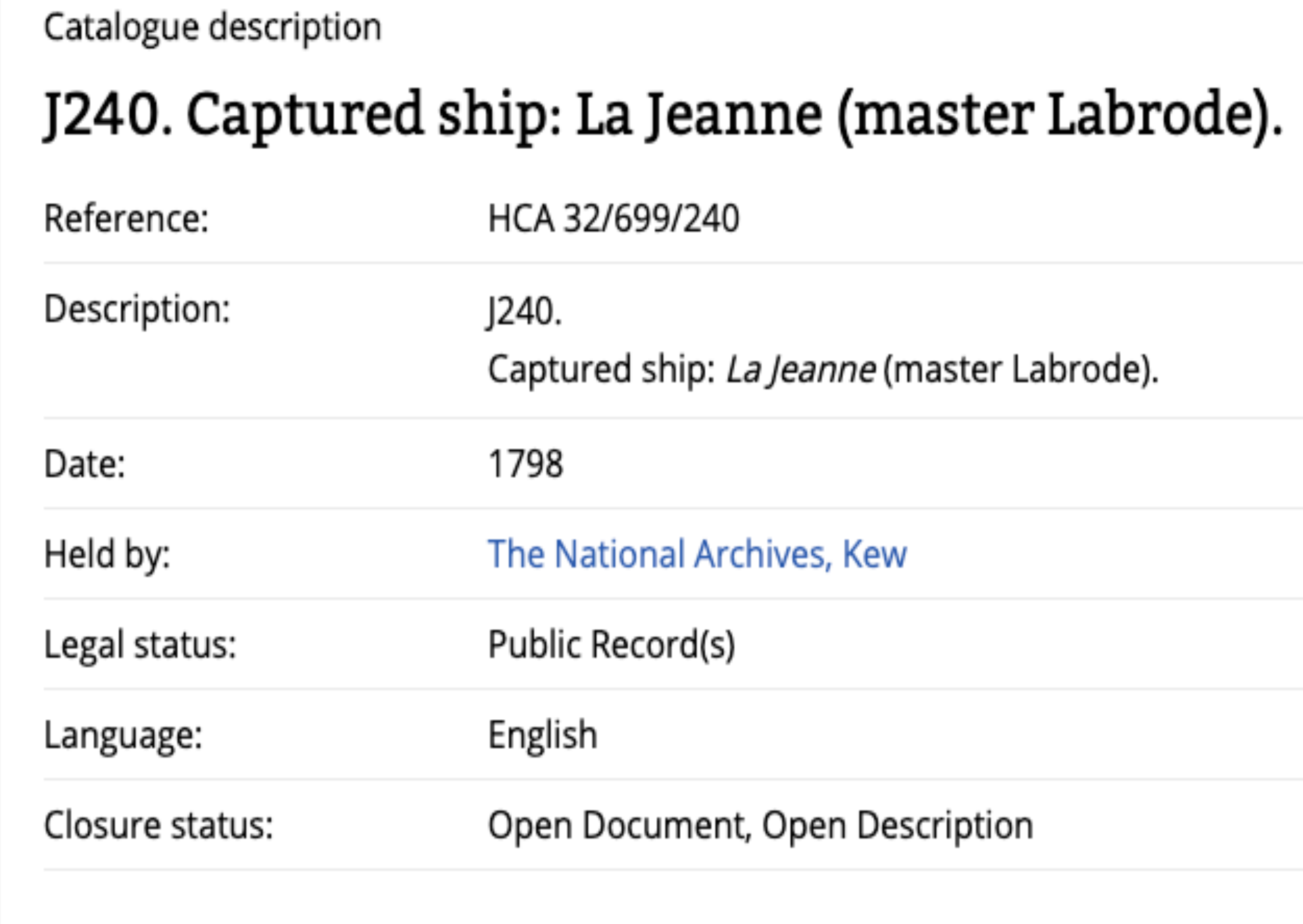
INTRODUCING MONK

Handwriting Recognition Technology (HRT) developed at the University of Groningen



- MONK is a human-in-the-loop, continuous, machine learning system.
- Multiple training functions allow alternation between breadth and depth of training.
- Elements of gamification in user interface.
- Subset of 2211 Prize Paper interrogation pages in MONK, total archive is roughly 20,000 pages.

BENEFITS OF INDEXING FOCUS IN HRT



Catalogue description

J240. Captured ship: La Jeanne (master Labrode).

Reference:	HCA 32/699/240
Description:	J240. Captured ship: *La Jeanne* (master Labrode).
Date:	1798
Held by:	The National Archives, Kew
Legal status:	Public Record(s)
Language:	English
Closure status:	Open Document, Open Description

- Allows for quick increases in use-ability and search-ability.
- Allows for targeting for words of low frequency, rather than generic but non-descriptive ‘function’ words.
- Effective digital indexing increases access to digital and physical collections.
- Indexing focus is most appropriate work around in light of limitations of current technologies.

PREPARING FOR TRAINING I: SECTIONING BY LEGIBILITY

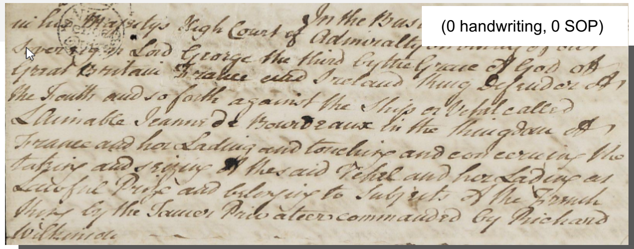
PURPOSE: to assign training samples of varying degrees of legibility, ensuring heterogeneity in training data.

METHODOLOGY: visually checking the scans, along two metrics:

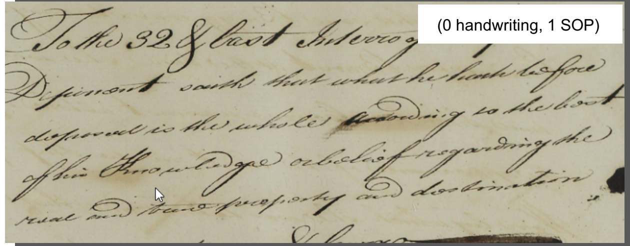
1) handwriting legibility
2) State of preservation (SOP)/ scan quality

Given binary assignation 1 = Good 0 = Bad. Results in 4 categories

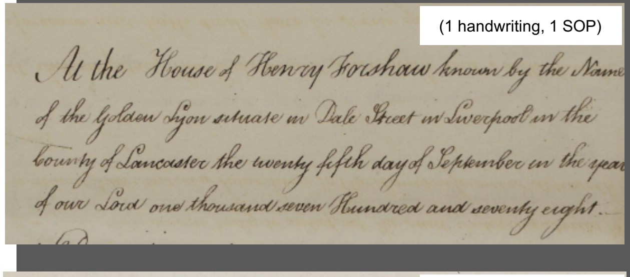
(0, 0) (0,1) (1,0) (1,1)


(0 handwriting, 0 SOP)

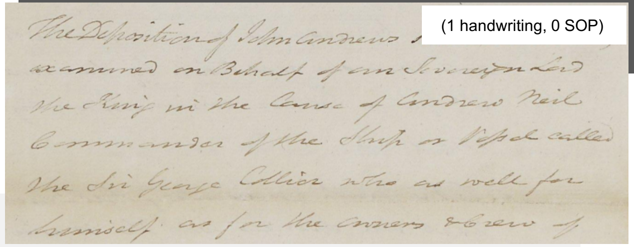
In the Busi
in his Majestys High Court of Admiralty on behalf of our
Sovereign Lord George the third by the Grace of God of
Great Britain France and Ireland King Defender of
the Faith and so forth against the Ship or Vessel called
L'Aimable Jeanne de Bourdeaux in the Kingdom of
France and her Lading and touching and concerning the
taking and seizing of the said Vessel and her Lading as
Lawful Prize and belonging to Subjects of the French
King by the Tamer Privateer commanded by Richard
Wilkinson


(0 handwriting, 1 SOP)

To the 32 & last Interrog
Deponent saith that what he hath before
deposed is the whole according to the best
of his Knowledge or belief regarding the
real and true property and destination

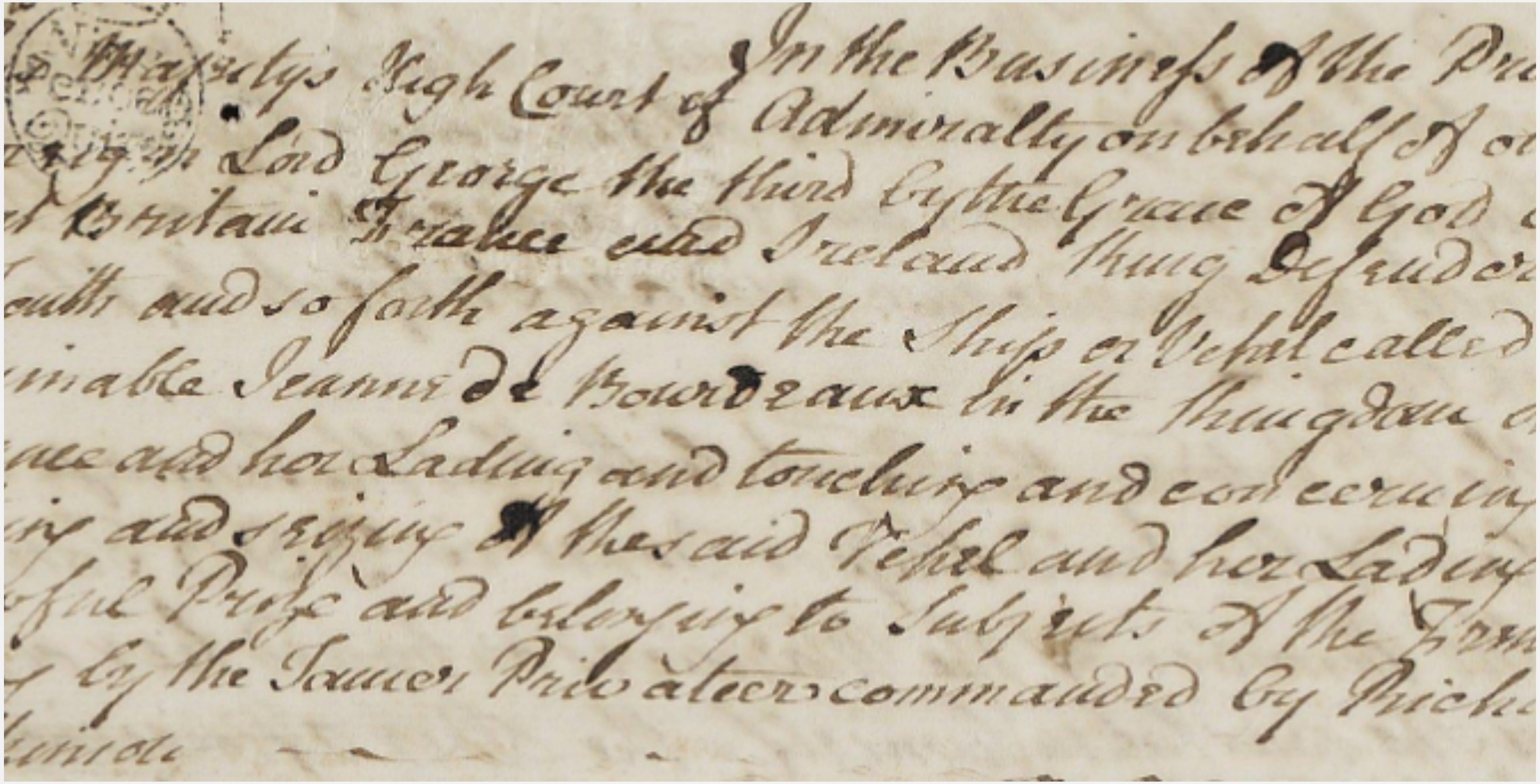

(1 handwriting, 1 SOP)

At the House of Henry Forshaw known by the Name
of the Golden Lyon situate in Dale Street in Liverpool in the
County of Lancaster the twenty fifth day of September in the year
of our Lord one thousand seven Hundred and seventy eight.


(1 handwriting, 0 SOP)

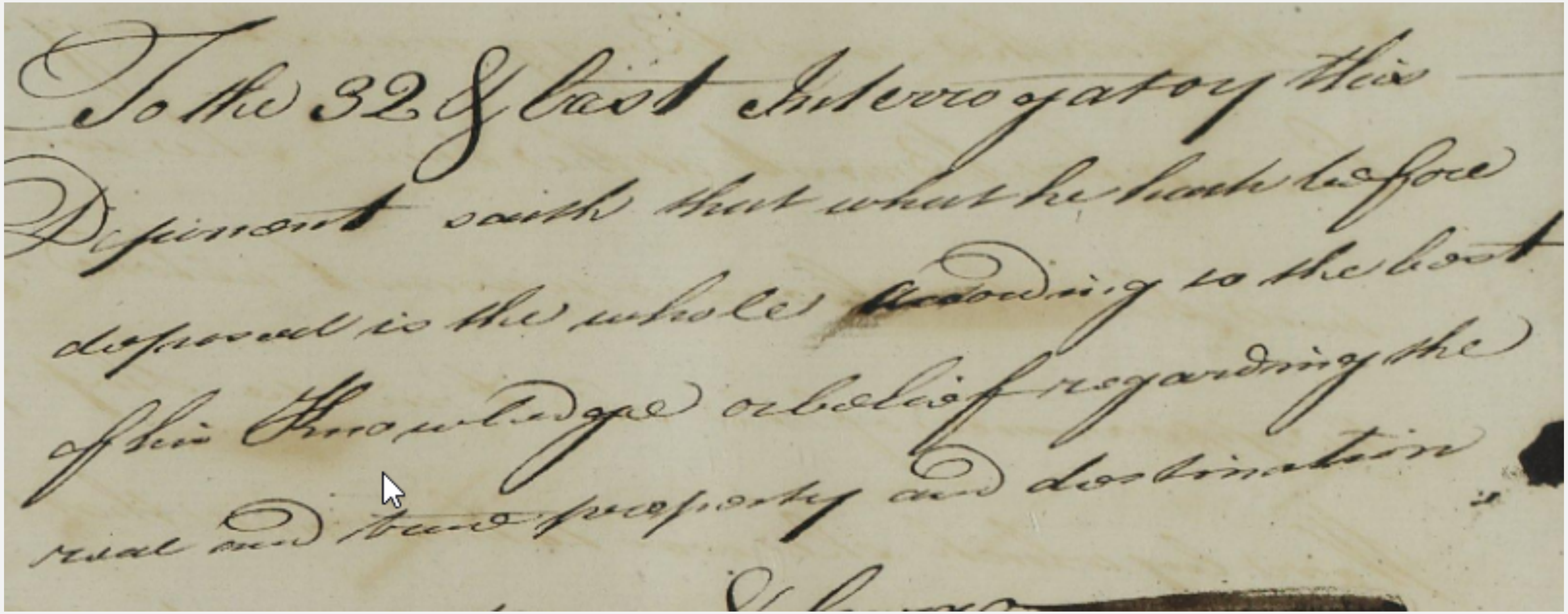
The Deposition of John Andrews
examined on Behalf of our Sovereign Lord
the King in the Cause of Andrew Neil
Commander of the Ship or Vessel called
the Sir George Collier who as well for
himself as for the Owners & Crew of

A. 0 HANDWRITING, 0 SOP



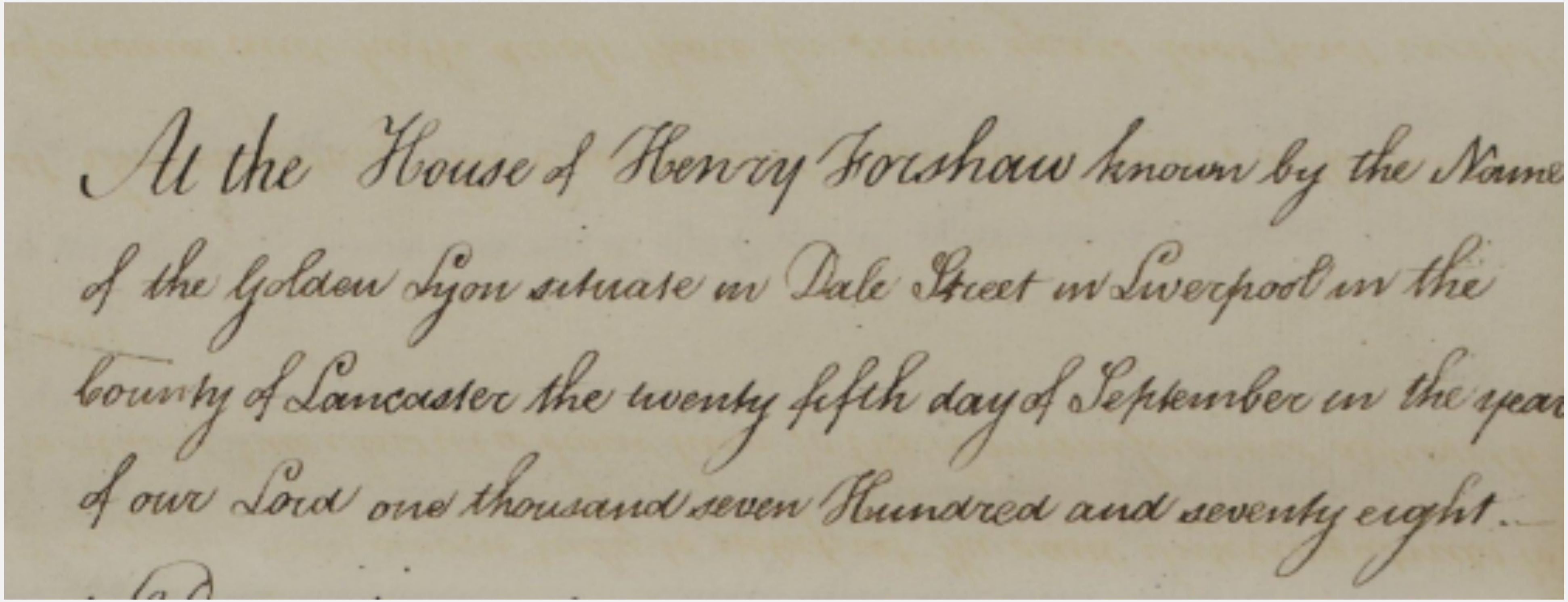
In the Business of the Pri
Majestys High Court of Admiralty on behalf of ou
reign Lord George the third by the Grace of God o
t Britain France and Ireland King Defender
Faith and so forth against the Ship or Vessel called
miable Jeanne de Bourdeaux in the kingdom o
nce and her Lading and touching and concerning
ing and seizing of the said Vessel and her Lading
ful Prize and belonging to Subjects of the Fren
by the Tamer Privateer commanded by Rich
imon

B. 0 HANDWRITING, 1 SOP



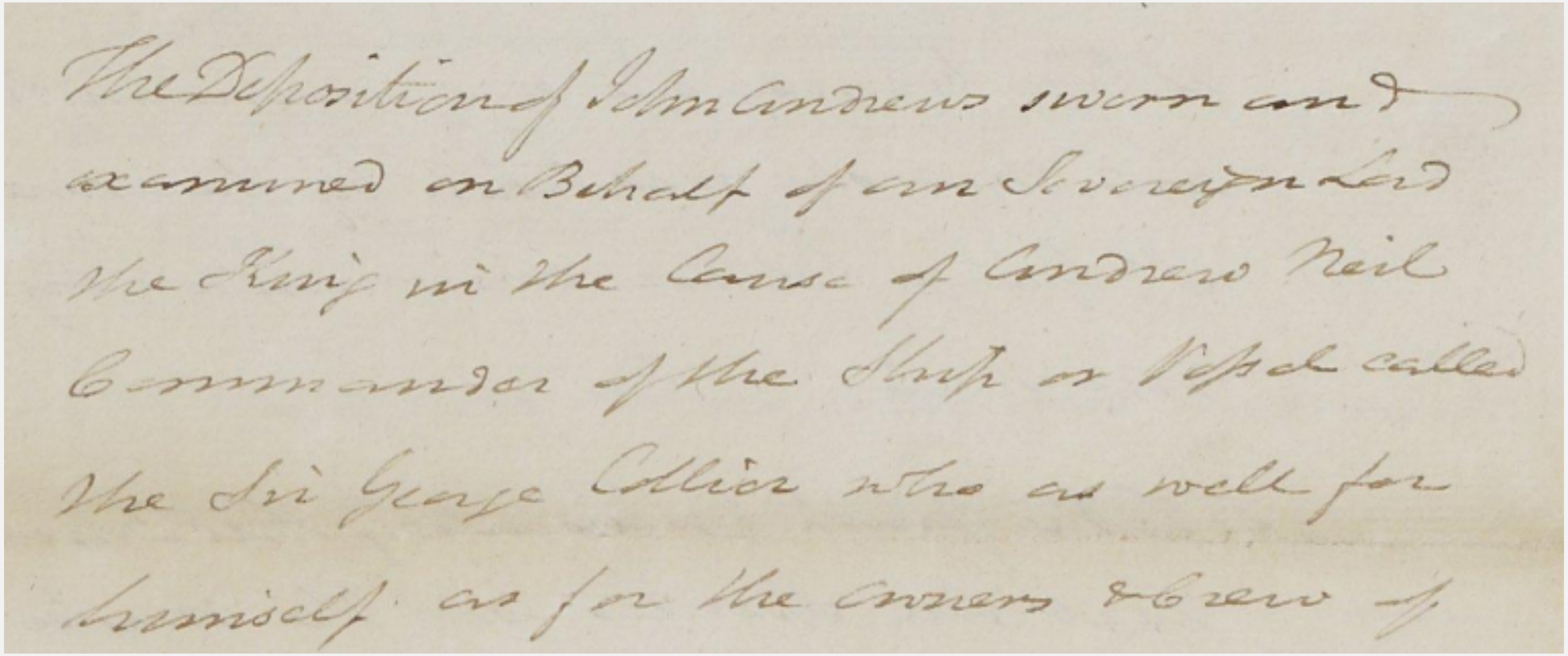
To the 32 & last Interrogatory this
Deponent saith that what he hath before
deposed is the whole according to the best
of his knowledge or belief regarding the
real and true property and destination

C. 1 HANDWRITING, 1 SOP



At the House of Henry Forshaw known by the Name
of the Golden Lyon situate in Dale Street in Liverpool in the
County of Lancaster the twenty fifth day of September in the year
of our Lord one thousand seven Hundred and seventy eight.

D. 1 HANDWRITING, 0 SOP



The Deposition of John Andrews sworn and
examined on Behalf of our Sovereign Lord
the King in the Cause of Andrew Neil
Commander of the Sloop or Vessel called
the Sir George Collier who as well for
himself as for the Owners & Crew of

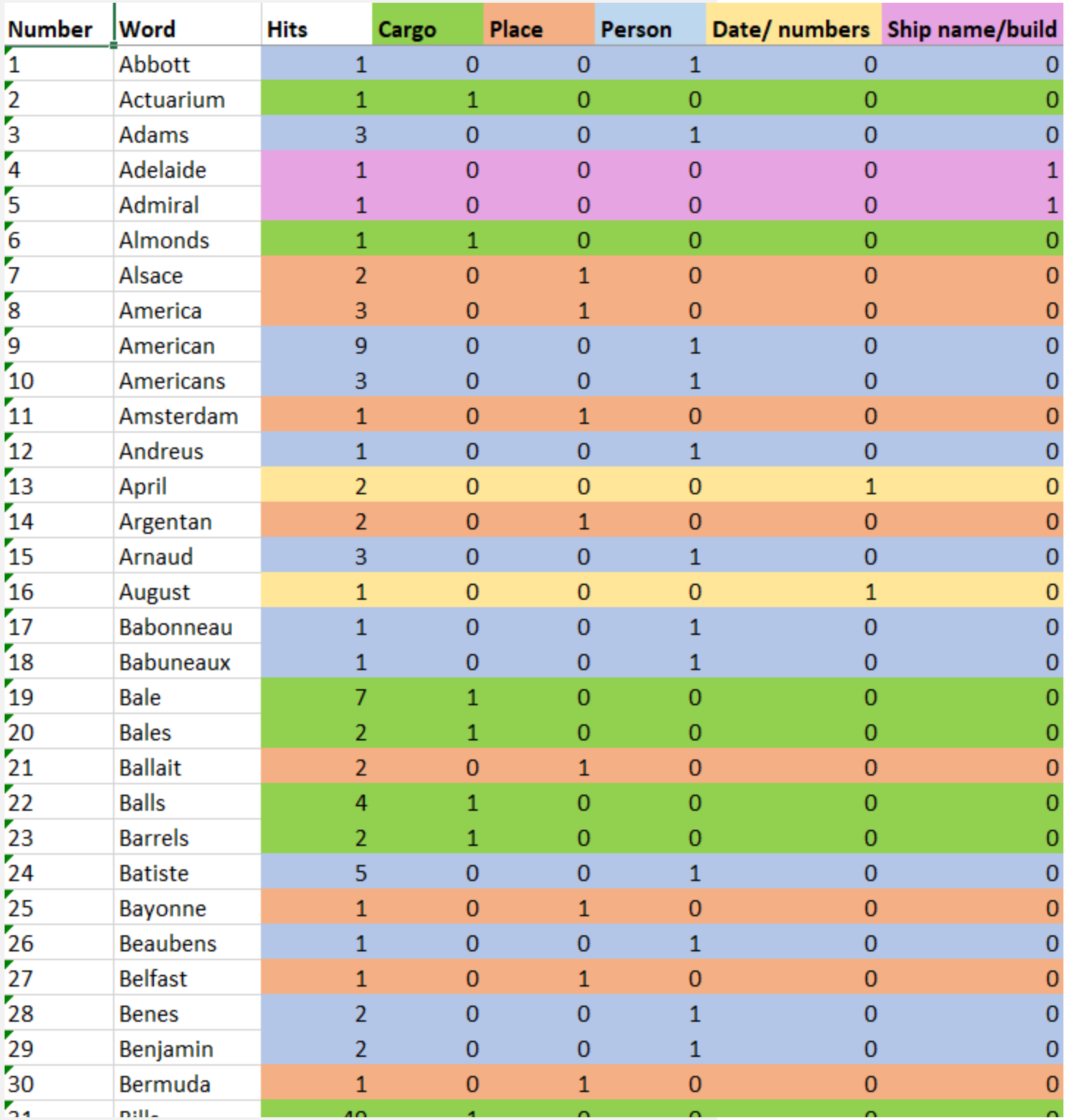
PREPARING FOR TRAINING II: EXTRACTING INDEXABLE WORDS

METHODOLOGY:
“Substantive” words correspond with commonly searched terms.

- Cargo,
- Place (geographical locations)
- Person (names, titles)
- Dates/ numbers
- Ship names/ build

List of substantive words extracted from both the metadata of the Prize Papers and from MONK’s Static Index.

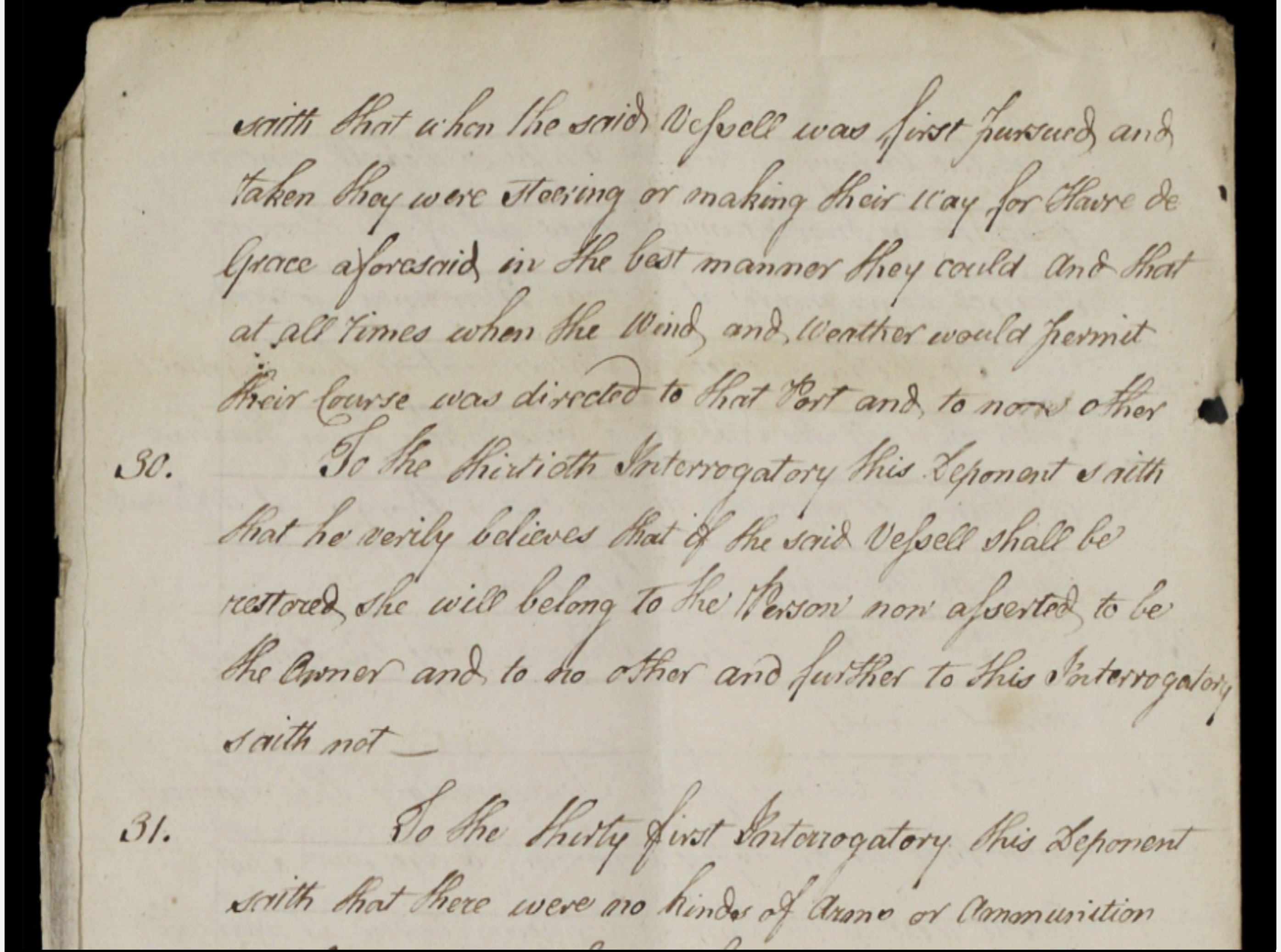
Possible to use the Static Index to identify number of ‘hits’ in the system prior to training, allowing for measures of progress.



Number	Word	Hits	Cargo	Place	Person	Date/ numbers	Ship name/build
1	Abbott	1	0	0	1	0	0
2	Actuarium	1	1	0	0	0	0
3	Adams	3	0	0	1	0	0
4	Adelaide	1	0	0	0	0	1
5	Admiral	1	0	0	0	0	1
6	Almonds	1	1	0	0	0	0
7	Alsace	2	0	1	0	0	0
8	America	3	0	1	0	0	0
9	American	9	0	0	1	0	0
10	Americans	3	0	0	1	0	0
11	Amsterdam	1	0	1	0	0	0
12	Andreus	1	0	0	1	0	0
13	April	2	0	0	0	1	0
14	Argentan	2	0	1	0	0	0
15	Arnaud	3	0	0	1	0	0
16	August	1	0	0	0	1	0
17	Babonneau	1	0	0	1	0	0
18	Babuneaux	1	0	0	1	0	0
19	Bale	7	1	0	0	0	0
20	Bales	2	1	0	0	0	0
21	Ballait	2	0	1	0	0	0
22	Balls	4	1	0	0	0	0
23	Barrels	2	1	0	0	0	0
24	Batiste	5	0	0	1	0	0
25	Bayonne	1	0	1	0	0	0
26	Beaubens	1	0	0	1	0	0
27	Belfast	1	0	1	0	0	0
28	Benes	2	0	0	1	0	0
29	Benjamin	2	0	0	1	0	0
30	Bermuda	1	0	1	0	0	0
31	Bills	40	1	0	0	0	0

TRAINING ON THE DAY

- Held at Brill publishers, Leiden, on October 10th, 2019.
- Recruited 14 expert volunteers with background in History or Archival Studies.
- Assigned varied tasks throughout the day seeking to exploit both **breadth** and **depth** training functions
- Each volunteer assigned own list of words and scans to avoid repetition in training data.
- Volunteer preference was for **breadth** over **depth** training functions
  - hunting for new words.
  - these offered more insight into narrative of the sources
  - also greater 'flow' in productivity.
- Total training time 5 hours.



saith that when the said Veſsell was first pursued and
taken they were steering or making their Way for Havre de
Grace aforesaid in the best manner they could And that
at all Times when the Wind and Weather would permit
their Course was directed to that Port and to none other

30. To the Thirtieth Interrogatory this Deponent saith
that he verily believes that if the said Veſsell shall be
restored she will belong to the Person now aſserted to be
the Owner and to no other and further to this Interrogatory
saith not —

31. To the Thirty first Interrogatory this Deponent
saith that there were no kinds of Arms or Ammunition

OUTCOMES

- Along depth axis overall increase in word accuracy of 3.87%
  - Corresponds with 761 newly labelled instances
- Main progress along breadth axis:
  - 113 new word classes labelled.
  - Doubling of total transcribed lines from 1143 to 2224.

REFLECTIONS

- Volunteer feedback suggests enthusiasm for workshop style crowdsourced events

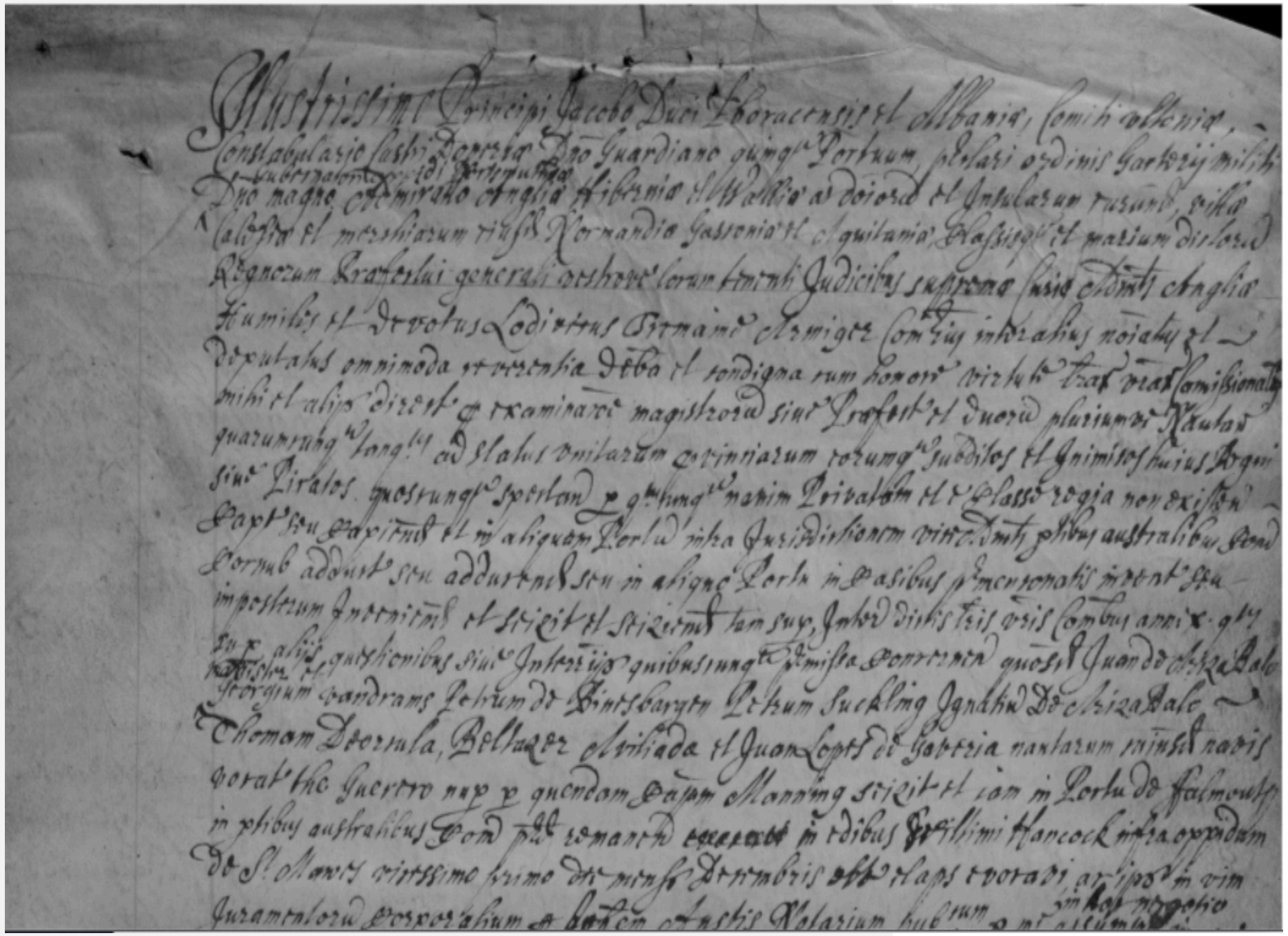
- Also suggests depth-based training activities likely to cause greater volunteer fatigue and require more breaks and/or gamification.

- Preliminary findings suggestive that:

  - targeted HRT training could aid the rapid expansion of archival indexing.

  - very little input from volunteers needed in terms of hours to achieve results..

- Larger-scale benchmarking study still needed.



Illustrissimo Principi Jacobo Duci Eboracensis et Albaniæ, Comiti Ultoniæ,
Constabulario Castri Dovoriæ Dno Guardiano quinque Portuum, praeclari ordinis Garterij militi
Gubernatori [illegible] Portsmouth
Dno magno Admirallo Angliæ Hiberniæ et Walliæ ac dominiorum et Insularum earundem, villæ
Calesiæ et Marchiarum eiusdem Normandiæ Gasconiæ et Aquitaniæ, Classiumque et marium dictorum
Regnorum Praefecto generali ac honorandis eorum tenenti Judicibus supremæ Curiae Admiralitatis Angliæ
Humilis et devotus [illegible] Armiger Comm[issari]us inter alios nominatus et
deputatus omnimoda reverentia debita et condigna cum honore. Virtute literarum vestrarum Commissionalium
mihi et aliis direct[arum] pro examinatione magistrorum sive Praefectorum et ducerum plurium Nautarum
quarumcumque longitudinis ad status unitarum provinciarum eorumque subditos et inimicos huius Regni
sive Piratos quoscunque spectantium per quamcumque navim Privatam et a Classe regia non existentem
capt[as] seu capiendas et in aliquem Portum intra Jurisdictionem vestram dictis partibus australibus Comitatus
Cornubiæ adduct[as] seu adducendas seu in aliquo Portu in partibus praementionatis moventes seu
imposterum inveniendas et seizit et seizendas tam sup[er] Interr[ogatoriis] [illegible] anni x. [illegible]
sive aliis quaestionibus sive Interrogatoriis quibuscumque praemissa concernentia quosdam Juan de Aizabalo
Georgium Vandrams Petrum de Vintsbargen Petrum Suckling Ignatium de Elizabale
Thomam Deornula, Beltazar Arriliada et Juan Lopes de Gabria nautarum cuiusdam navis
vocat' the Guertro nuper per quendam [illegible] Manning seizit et cap[tae] in Portu de Falmouth
in partibus australibus Corn[ubiæ] p[raedict]æ remanent [illegible] in aedibus Willelmi Hancock infra oppidum
de St Mawes vicessimo primo die mensis Decembris [illegible] elaps[i] evocavi ac ipsos in vim
Juramentorum Corporalium [illegible] Notarium [illegible]

CONTACT DETAILS

Anna Caceres,
Leiden University, Postbus 9500, 2300 RA Leiden, The Netherlands
annamcaceres@googlemail.com

Andreas Weber
University of Twente, BMS-STePS, 7500 AE Enschede, The Netherlands
a.weber@utwente.nl

Lambert Schomaker
University of Groningen, Bernoulli Institute, Nijenborgh 9, 9747 AG Groningen, The Netherlands
l.r.b.schomaker@rug.nl

MONK in Practice:
Indexing Heterogeneous Handwritten Collections

Anna Caceres[1], Andreas Weber[2], Lambert Schomaker[3]

[1] Leiden University, Postbus 9500, 2300 RA Leiden, The Netherlands
annamcaceres@googlemail.com
[2] University of Twente, BMS-STePS, 7500 AE Enschede, The Netherlands
a.weber@utwente.nl
[3] University of Groningen, Bernoulli Institute, Nijenborgh 9, 9747 AG Groningen, The Netherlands
l.r.b.schomaker@rug.nl

Abstract: This short paper describes how MONK, a machine-learning driven handwriting recognition system, can be used to rapidly index a heterogeneous handwritten collection with the help of volunteers. We discuss the setup and results of an event which saw volunteers come together to enrich a subset of the digitized Prize paper collection, a collection of historical handwritten documents of the High Court of Admiralty (1652-1815).

Keyword: *handwriting recognition, user study, heterogeneous archives, archives, active learning, Prize papers, machine learning*

Over the last decades archives, museums, research institutions and publishers have undertaken major efforts to index their digitized handwritten collections. This paper describes the setup and results of an event which saw 14 expert volunteers come together to enrich a digitized collection of a visually heterogeneous archive – the Prize Papers - (see figure 1) using MONK, a machine-learning driven handwriting recognition system developed at the University of Groningen. MONK does not require prior training. It starts from scratch and actively and continuously learns from the input of users (Schomaker, 2016 and 2019). The event took place in the offices of Brill publishers in Leiden in October 2019 and took less than one working day, with time for instruction.

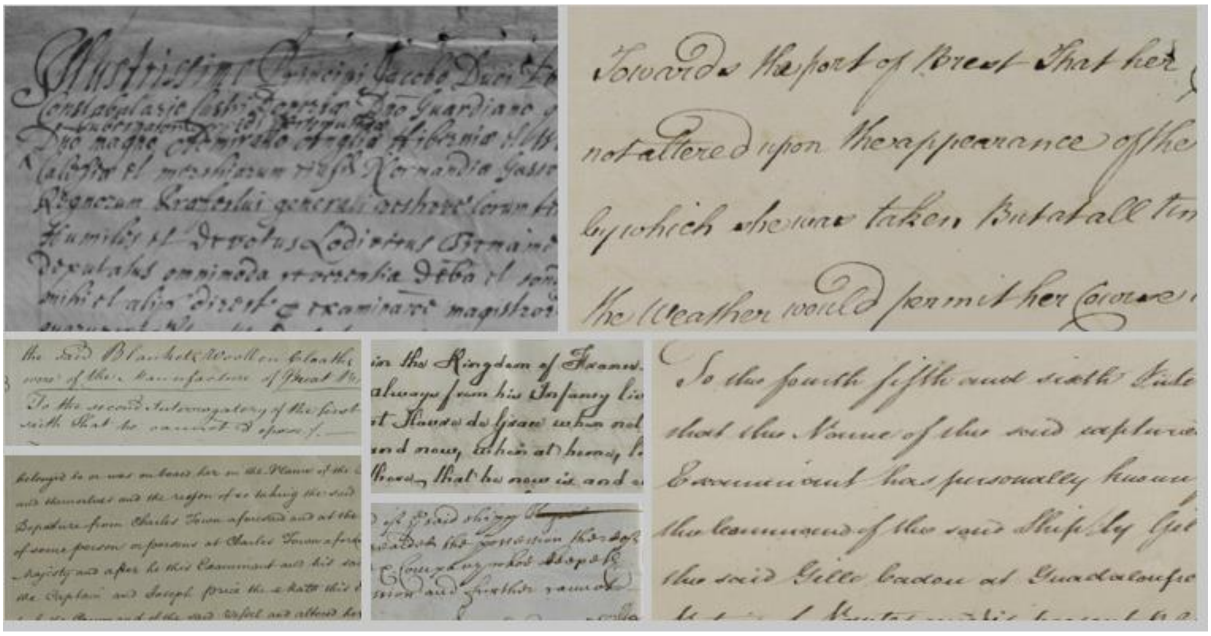


Figure 1: *Snippets from the Prize paper collection in the MONK system to show heterogeneity of script-styles.*

An indexing rather than line-by-line transcription focus, meant targeting labels on words known to be of indexable significance, or targeting areas of the document where such words were predicted to appear. Here, the format of the archive itself is of significance. The Prize Papers are the records of the High Court of the Admiralty, a British maritime legal body, and date from 1652 - 1815. (Van Lottum & Zanden, 2014). The archive consists of two parts: standardised interrogations of crew members on one hand, and miscellaneous seized documents from the ships on the other. For our use case we took a sample of 2111 pages from the interrogations, which were valuable because they presented a variety of script styles, paired with a highly standardised content, consistently asking a set of roughly 32 questions.

When identifying target zones for labelling then, we knew for example, that questions 7 and 8 enquire about the name, destination and origin of the ship so index-focused labelling should concentrate on these areas. Indexation is more valuable in the short and medium term, as it immediately increases the searchability, and thus useability, of historical documents (Zant et al., 2009; Colavizza, Ehrmann and Bortoluzzi, 2019). It further provides continuous learning systems with targeted training for word classes which typically appear less frequently, such as place names, people names and objects.

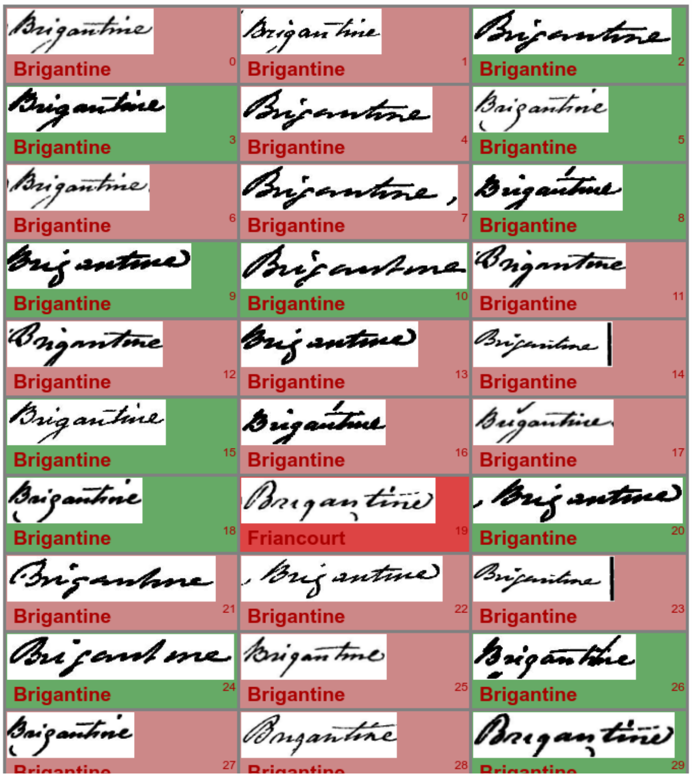


***Figure 2**: Example of a resulting hit list for word 'Brigantine' after labeling, using an LSTM recognizer in Monk (Ameryan & Schomaker, 2019). Green samples were used for training, Samples in light red correspond to the new harvest. A previously misrecognized sample 'Friancourt' (dark red) is now correctly recognized as Brigantine.*

Volunteers were assigned different labelling activities targeting both breadth (number of word classes recognised) and depth (accuracy of recognition) of knowledge in MONK. MONK generates suggestions both for word zones (beginning and ends of words) and word classes (alphabetic content) which users confirm or reject in various formats. Figure 2 shows a single-word hit list for "Brigantine", one example of machine-generated and human-corrected labelling. Training in different functions allowed both specific words and specific pages to be targeted.

The labelling efforts were primarily successful along the breadth axis with 113 new word classes labelled and a doubling of total transcribed lines - from 1143 to 2224. Along the depth axis there was an overall increase in word accuracy of 3.87% thanks to 761 newly labelled instances. This short paper details how MONK can facilitate the rapid indexation of heterogeneous archival material with a very limited involvement of volunteers. However, in order to make more general statements about the system's efficiency a much larger benchmarking study would be necessary.

References:

Ameryan, M. & Schomaker, L. (2019). A limited-size ensemble of homogeneous CNN/LSTMs for high-performance word classification, arXiv:1912.03223

Colavizza, G., Ehrmann, M. and Bortoluzzi,F., "Index-Driven Digitization and Indexation of Historical Archives," Frontiers in Digital Humanities, 6:4 (2019). https://doi.org/10.3389/fdigh.2019.00004

Lottum J. van & Zanden, J.L., "Labour Productivity and Human Capital in the European Maritime Sector of the Eighteenth Century," Explorations in Economic History, vol. 53, pp. 83-100, 2014.

van der Zant, T., Schomaker, L., Zinger, S., & van Schie, H. (2009). Where are the Search Engines for Handwritten Documents? Interdisciplinary Science Reviews, 34(2-3), 224-235. https://doi.org/10.1179/174327909X441126

Schomaker, L. (2019). Lifelong learning for text retrieval and recognition in historical handwritten document collections, arXiv:1912.05156 [chapter in book]

Schomaker, L. (2016). Design considerations for a large-scale image-based text search engine in historical manuscript collections. Information Technology, 58(2), 80-88. https://doi.org/10.1515/itit-2015-0049