

Maximising the utility of Jeremy Bentham's manuscripts

READ



Dr Louise Seaward
Bentham Project, University
College London

@TranscriBentham

Summary

- The Bentham Project
- Transcribe Bentham
- HTR models
- Keyword Spotting
- The future?

Jeremy Bentham (1748-1832)



*'It is the greatest happiness of the greatest number
that is the measure of right and wrong'*

The Bentham Project

- Scholarly edition of Bentham's *Collected Works*
- 75,000 folios of Bentham's writings – at UCL and The British Library
- 33 volumes completed out of a projected 80 volumes
- Not yet halfway to completion – after nearly 60 years!



Transcribe Bentham



- Launched in 2010, initially as short-term experiment
- One of first humanities crowdsourcing projects
- Volunteers transcribe and mark-up pages of Bentham's manuscripts
- 20,000+ pages transcribed at high level of accuracy – thank you to all volunteers!

Benefits of crowdsourcing

1. Preservation

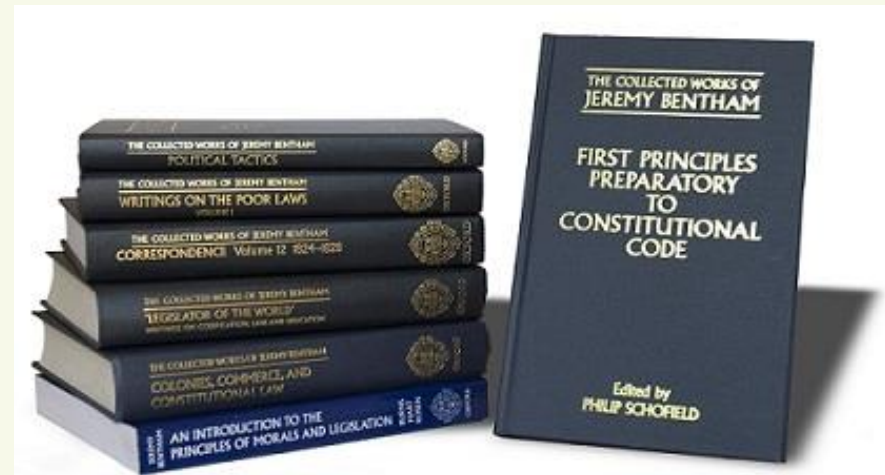
Bentham's writings digitised and transcribed – 95,000 images

2. Scholarship

Transcripts used to produce Bentham's *Collected Works* and can be reused in other research

3. Public engagement

Involving the public in research and Bentham studies



JEREMY BENTHAM'S PRISON COOKING

A COLLECTION OF UTILITARIAN RECIPES



WITH SPECIAL CONTRIBUTIONS BY Chef Fergus Henderson,
FOOD HISTORIAN Dr. Annie Gray, AND THE CO-ORDINATOR
OF TRANSCRIBE BENTHAM, Dr. Tim Causer

Transcribe Bentham

Welcome to the Transcription Desk



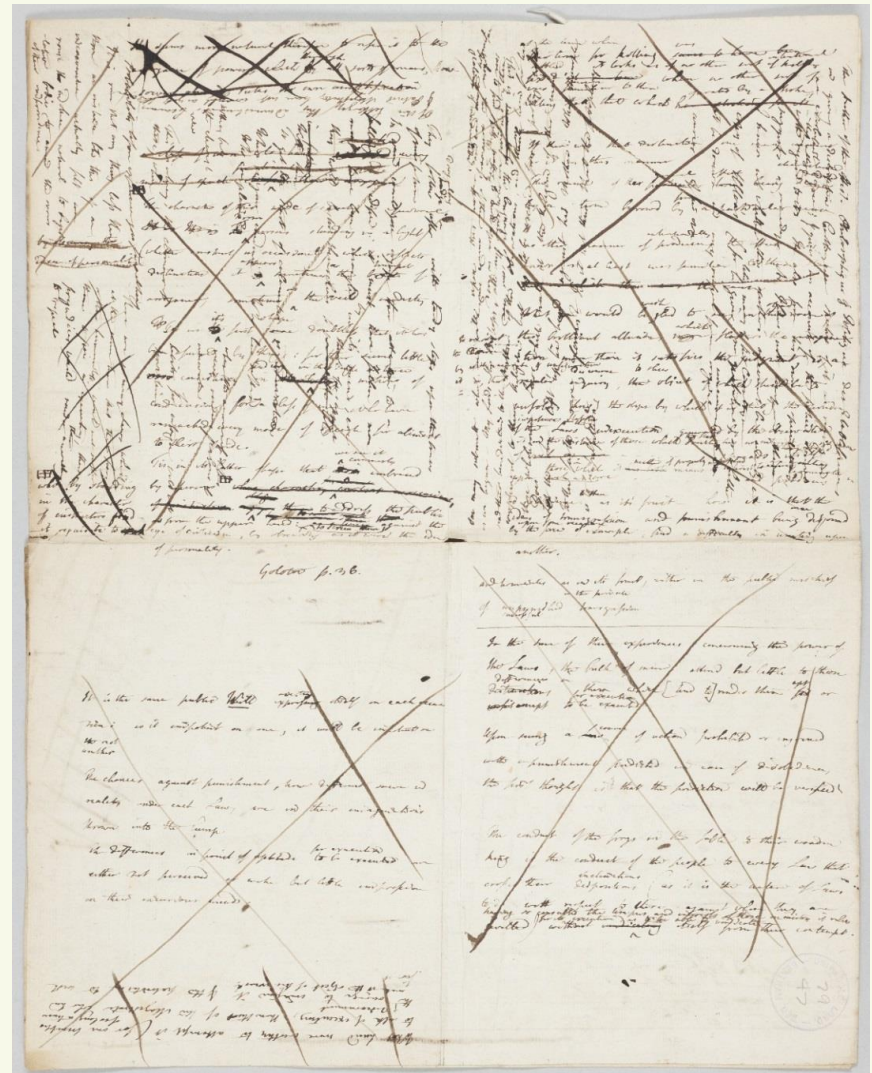
The Transcription Desk is the heart of a major online initiative to transcribe the manuscripts of the English philosopher Jeremy Bentham. It is managed by the [Bentham Project](#) at University College London.

You are invited to assist us by using the Transcription Desk to type up the text of Bentham's manuscripts.

















These transcripts will make it easier for anyone to access and read Bentham's papers and will be used by scholars at the Bentham Project in the production of the edition of [The Collected Works of Jeremy Bentham](#).

At the last count, volunteers have transcribed more than 20,000 pages of Bentham's writings! *Why not join us in our mission?*

- [Check out the project website and blog](#)
- [Sign up to our newsletter](#)
- [Follow us on Twitter](#) and [like us on Facebook](#)



TEI toolbar

Button																
Function	Line Break	Page Break	Heading	Paragraph	Addition	Deletion	Questionable Reading	Illegible Text	Marginal Note	Underline	Superscript	Unusual Spelling	Foreign Language	Ampersand	Long Dash	User Comment
Rendering	-	-	text	-	text	text	text[?]	[...]	text	text	te ^{xt}	text	text	&	—	-



Maximise Minimise

Wikitext Preview Changes

'''[{{fullurl:JB/116/172/001|action=edit}} Click Here To Edit]'''
<!-- ENTER TRANSCRIPTION BELOW THIS LINE -->

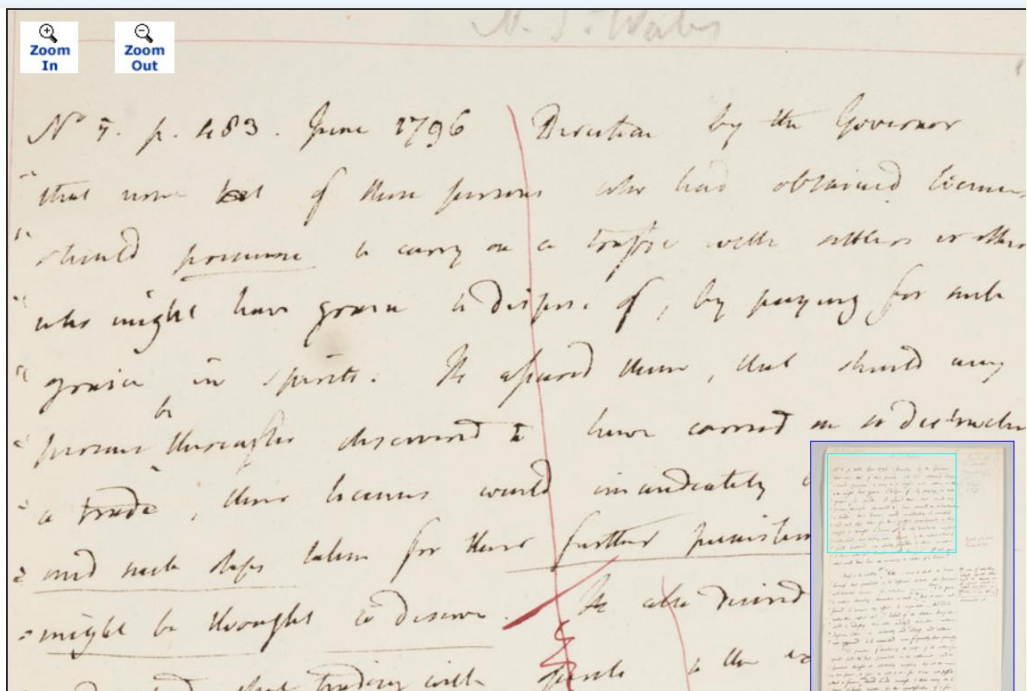
<head>13 July 1802 7 <lb/>N. S. Wales<lb/> 1 </head>

<p>N<hi rend='superscript'></hi> 7. p. 483. June 1796 Direction
by the Governor
<lb/>

"that none but of these persons who had obtained
licenses
<lb/>

"should <hi rend='underline'>presume</hi> to carry on a traffic
with settlers or others
<lb/>

"who might have grain to dispose of, by paying for such
<lb/>"grain in spirits. He assured them, that should any
"persons <add>be</add> thereafter discovered to have carried on so
destructive<lb/> "a trade, their licenses would immediately be
recalled, <lb/>"<hi rend='underline'>and such steps</hi> taken for
their <hi rend='underline'>further punishment as they <lb/> "might
be thought to deserve.</hi> He also desired as might be
<lb/>"understood, that trading with spirits to the extent which he
<lb/>"found practiced was strictly forbidden to others, as well as
<lb/>"to those who had licensed public houses." <note><gap/> of



User activity

- Dependent on a small group of **‘super transcribers’**
- 660 users have transcribed something at least once
- 31 super transcribers have worked on 95% of the 20,000 transcribed pages
- 11 super transcribers have transcribed more than 500 pages
- 15 super transcribers have contributed in the past year
- Around 3-5 users participating each week

We need to motivate our existing
super transcribers **AND**
encourage new people to take
part –

*Handwritten Text Recognition
technology could help!*

Let's go back to the start...

- Part of tranScriptorium project (2013-2015)
- Collaboration with Pattern Recognition and Human Language Technology (PRHLT) research centre at the Universitat Politècnica de València
- Using easier Bentham material as training data – writing by secretaries
- Around 900 pages of ground truth processed using Hidden Markov Models
- **Model with 18% CER**

tranScriptorium

Next step: neural networks

- 900 pages of simple ground truth reprocessed in Transkribus using neural networks from Computational Intelligence Lab (CITLab), University of Rostock
- **Model with 3.66% CER on test set**
- Model struggles to recognise more difficult handwriting
- Usually between 5-20% CER on a random page from the collection
- Model and dictionary are freely available in Transkribus – ‘English Writing M1’
- ‘English Writing M1’ is a good base model for training and recognising other collections

If then between two pleasures, the one produced
the possession of
by money, the other not, a man had as his those
enjoy the one as the other, such pleasures are to

If then between two pleasures, the one produced ↵
the possession of ↵
by money, the other not a man had as his those ↵
enjoy the one as the other, such pleasures are to ↵

8.9% CER on this page

Bentham's handwriting

- Advances in Layout Analysis sped up ground truth production – yay!
- Created new ground truth in Transkribus based on Bentham's *worst* handwriting
- First model – **57,000 words → 26.53% CER on test set**
- Experimented with Text2Img matching but too many errors
- Next model – **81,000 words → 17.75% CER on test set**
- This CER is too high for reliable transcription now – but with Transkribus, the future is bright!

made of issuing the proposed Annuity, now, and
in the payment of the interest or dividend,
as they become due, the provisions will be amended
to apt to appear in third trifling details
the which course of procedure being in every section

made of issuing the proposed Annuity hopes, and
the in the payment of the interest or dividend
in
in they become due, the provisions will be amended
to apt to appear in third trifling details
the which course of procedure being in every section

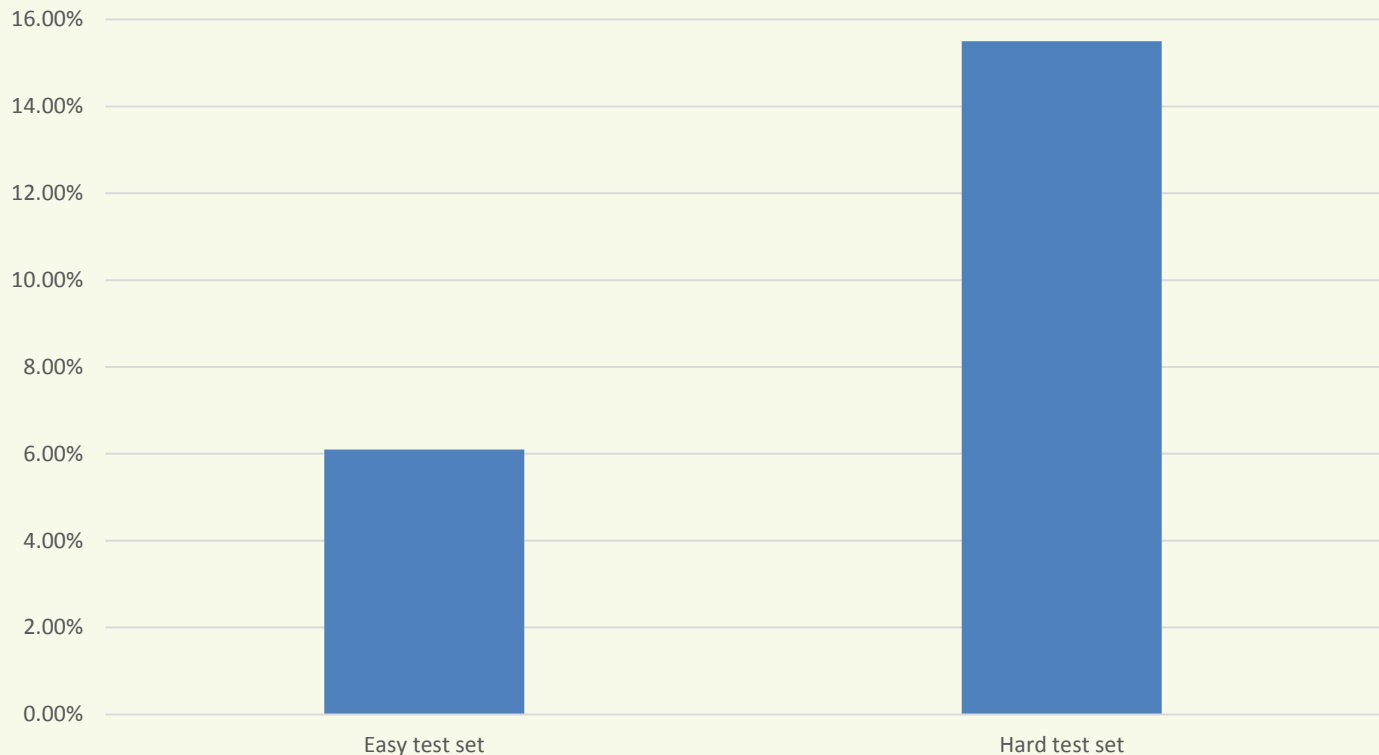
34.4% CER on this page

Keyword Spotting

- KWS works well even when HTR models have a relatively high error rate
- Collaboration with Pattern Recognition and Human Language Technology (PRHLT) research centre at the Universitat Politècnica de València
- We shared 95,000 images, 1200 pages of ground truth and metadata records
- Data cleaned and 95,000 images segmented in Transkribus in batch mode
- Valencia processed ground truth with Laia toolkit – neural network HTR and probabilistic word indexing

Keyword Spotting

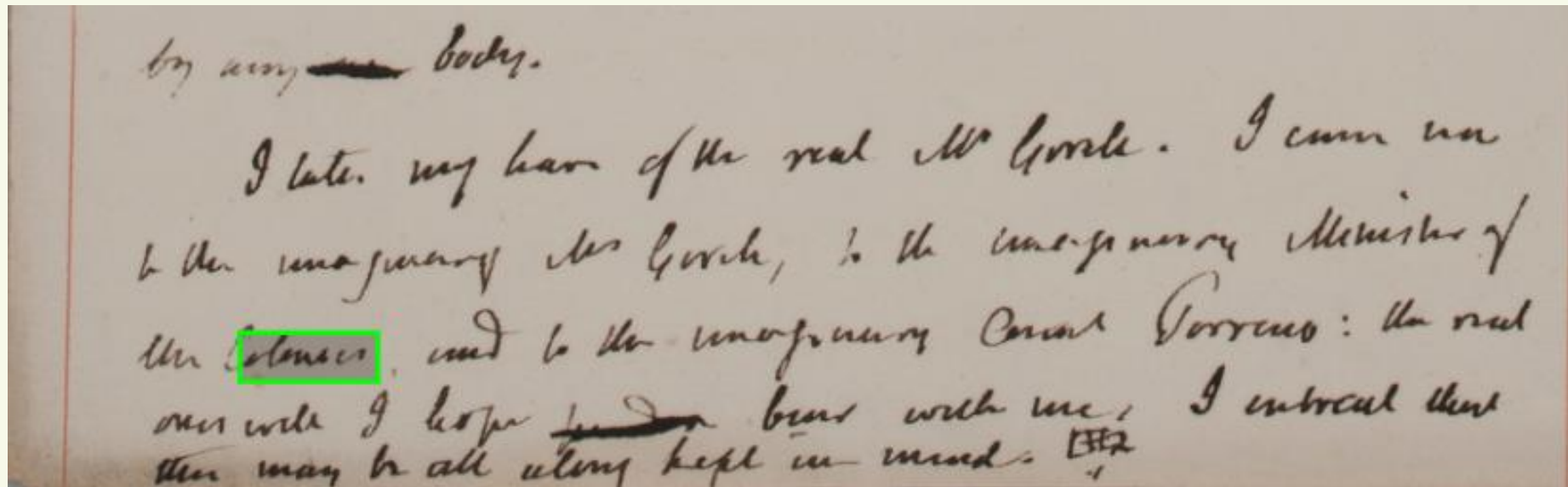
CER of model when tested on different data sets from UCL Bentham collection



Keyword Spotting interface

Search 90,000 images of Bentham's writings:

<http://prhlt-carabela.prhlt.upv.es/bentham/>



Search for 'democracy'

against the democracy; I am not against democracy; I reverence
the democracy, and I reverence the King: but because that beautiful
structure which consists of a due mixture of its component parts
would be injured, if not destroyed, by making so essential an alter-
ation in the Commons House of Parliament. The British Consti-

Search for 'democracy'

1. Under every form of government but the
 democracy the sole object which enters as a
 the nature of man can ever be really in
 happiness of the nation or rulers: and to
 founded as the virtue of the people if the ruler
 made and made to be a part of the happy

Keyword Spotting in the wild

- Since 15 October 2018 - 114 unique users from 25 countries have made searches
- A user in Italy searched for: 'Naples'
- A user in USA searched for: 'Jesus'
- A user in Austria searched for 'anarchy'



Keyword Spotting in the wild

Transcribe Bentham volunteers and other users have been recording their searches on a Google sheet:

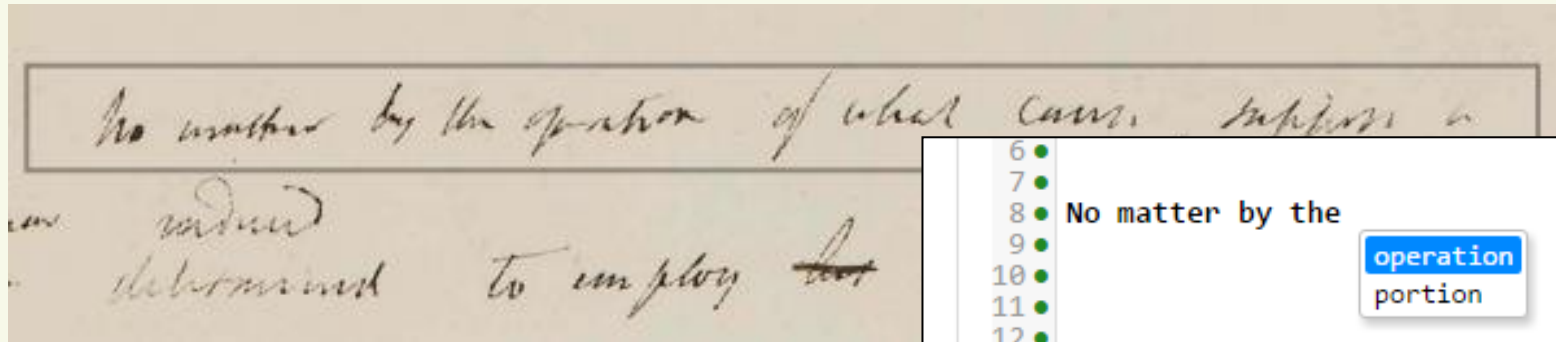
- A search for 'legislature' at 80% confidence returned 1137 matches
- A search for 'muzzy' at 80% confidence returned 0 matches
- Searches for 'Manchester' and 'massacre' show that Bentham wrote about Peterloo massacre of 1819



Use cases for KWS

- Allows anyone to search all of Bentham's writings
- Useful for researchers interested in Bentham, philosophy, law, history and more
- Will allow Bentham Project researchers to find previously unknown text
- Will help Transcribe Bentham volunteers to find interesting material to work on

The future...



- Include Valencia KWS technology in Transkribus GUI and Web
- Connect KWS site to existing digital Bentham resources – catalogue etc.
- Improve HTR models with PHRLT and CITLab – more specific training data for different hands and languages
- Integrate HTR into Transcribe Bentham
- Volunteers could check and correct automated transcripts or ask for computer-generated word suggestions
- Potential to attract new volunteers who are daunted by Bentham's handwriting

‘Many hands make light work’

A portrait of James Oglethorpe, a man with long, wavy hair, wearing a dark jacket over a light-colored shirt. A speech bubble is positioned above his head.

Thank you!

My thanks go to:

- Staff at the Bentham Project
- Transcribe Bentham volunteers
- PRHLT team at UPV
- CITlab team at Rostock
- Transkribus team at Innsbruck
- Our other READ colleagues

Thanks for listening!

<https://www.ucl.ac.uk/bentham-project/>

<http://transcribe-bentham.ucl.ac.uk/>

transcribe.bentham@ucl.ac.uk

@TranscriBentham



READ

