Digital Humanities Across Borders

Class 3: Digitizing and digitized text

Hat tip to Ryan Cordell

Hat tip to Ryan Cordell

Written by Ryan Cordell
on January 10, 2019

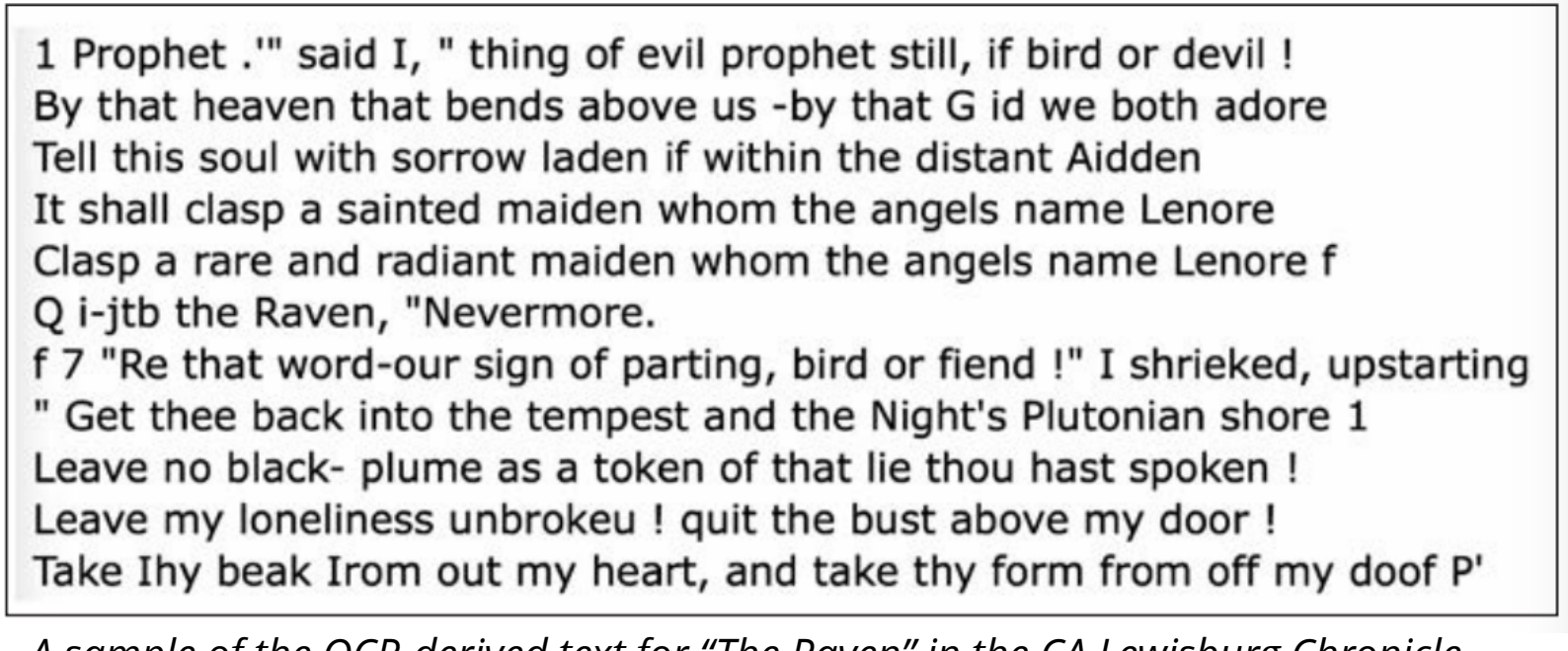
Why You (A Humanist) Should Care About Optical Character Recognition

Yesterday David Smith and I announced the release of "A Research Agenda for Historical and Multilingual Optical Character Recognition," a report funded by the Andrew W. Mellon Foundation and conducted in consultation with the NEH's Office of Digital Humanities and the Library of Congress. These groups realized that many of the digital humanities projects they support struggle with similar issues related to the quality of their source text data. They asked us to survey the current state of OCR for historical and multilingual documents, and to recommend

New OCR report

Why should you care about OCR?

Why should you care about OCR?



1 Prophet .'" said I, " thing of evil prophet still, if bird or devil !
By that heaven that bends above us -by that G id we both adore
Tell this soul with sorrow laden if within the distant Aidden
It shall clasp a sainted maiden whom the angels name Lenore
Clasp a rare and radiant maiden whom the angels name Lenore f
Q i-jtb the Raven, "Nevermore.
f 7 "Re that word-our sign of parting, bird or fiend !" I shrieked, upstarting
" Get thee back into the tempest and the Night's Plutonian shore 1
Leave no black- plume as a token of that lie thou hast spoken !
Leave my loneliness unbrokeu ! quit the bust above my door !
Take Ihy beak Irom out my heart, and take thy form from off my doof P'

A sample of the OCR-derived text for "The Raven" in the CA Lewisburg Chronicle.

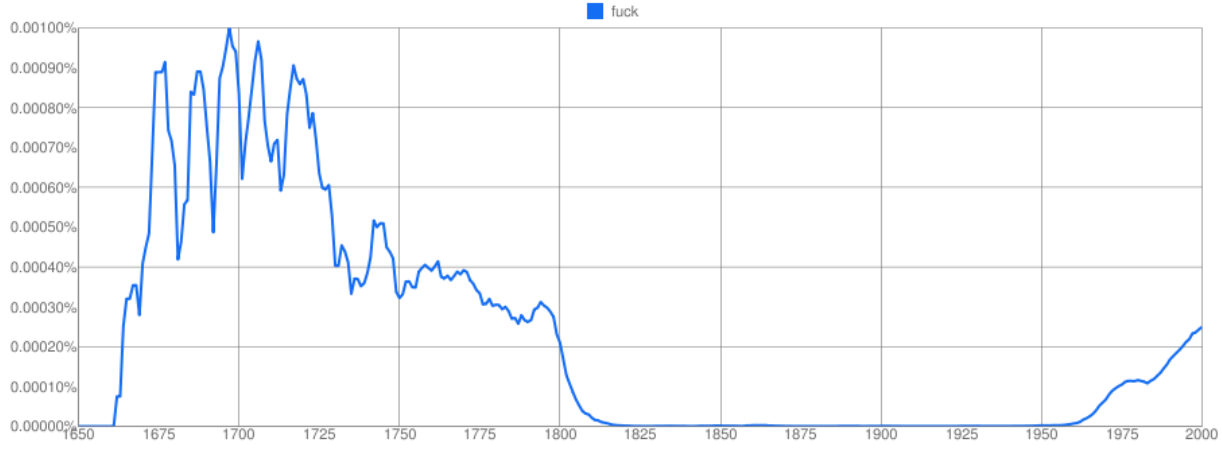
From Ryan Cordell's article"Q i-jtb the Raven".

OCR errors matter

OCR errors matter

fuck

Graph these **case-sensitive** comma-separated phrases: fuck
between 1650 and 2000 from the corpus English with smoothing of 3 .
Search lots of books



Search in Google Books:

1650 - 1676	1677 - 1716	1717 - 1725	1726 - 1783	1784 - 2000	fuck

OCR errors matter

fuck

Graph these **case-sensitive** comma-separated phrases: fuck
between 1650 and 2000 from the corpus English with smoothing c
Search lots of books

Search in Google Books:

1650 - 1676	1677 - 1716	1717 - 1725	1726 - 1

"fuck"

About 430 results (0.25 seconds)

Jan 1, 1650–Dec 31, 1676 › **Search English pages**

The reformed common-wealth of bees: presented in severall letters ... - Page 28

Samuel Hartlib - 1655 - 102 pages - Full view

... communicated to the dew, so that to **fuck** such clusters of Floures ina morning is almost as pleasant as to **fuck** a Honey.combe for taste- ... sowing Anise at several times , for it is the floure oue. ly of it, which the Bees **fuck** on. ...

books.google.com - More editions - Add to My Library ▼

Four books on the eleventh of Matthew: viz. I, Christ inviting ... - Page 741

1659 - Full view

... bring us to th% Heavenly Canaan* and the Saints c hey may **fuck** Mi lk and Honey continually, ... wicked have nothing but Swill and Dogs meat, to **fuck**, and to feed upon. And here is the difference ...

books.google.com - Add to My Library ▼

A supplement to the morning-exercise at Cripple-gate: or, Several ... - Page 341

Samuel Annesley - 1674 - 887 pages - Full view

... or greater charity > namely when the impediment to her giving **fuck** is natural ... and yet gave **fuck** to a Puppy, that her milk might be more artificially dried up. ...

books.google.com - More editions - Add to My Library ▼

The morning-exercise at Cripple-gate: or, Several cases of ... - Page 590

No cover image

1671 - 648 pages - Full view

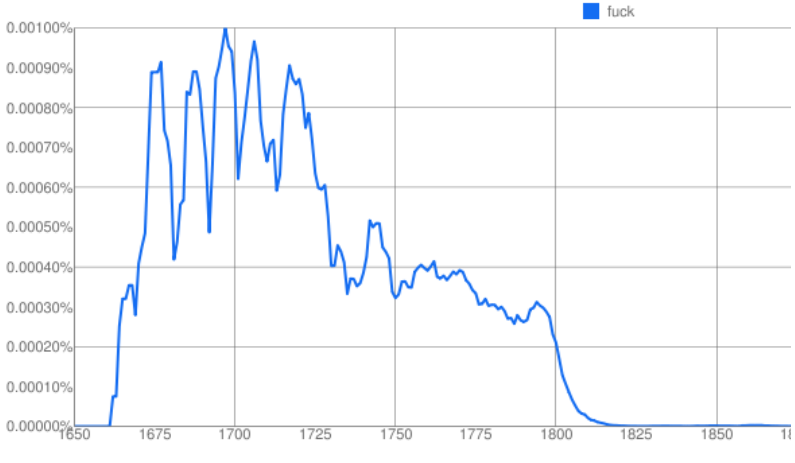
... which themselves do cot partake of, doth soon improve into malice also a-gainst that natural humane life which themselves are also partakers- of ; their desires to **fuck** the blood, as I may so fey, of good mens souls and graces, ...

books.google.com - More editions - Add to My Library ▼

OCR errors matter

fuck

Graph these **case-sensitive** comma-separated phrases: fuck
between 1650 and 2000 from the corpus English with smoothing c
Search lots of books



Search in Google Books:

1650 - 1676	1677 - 1716	1717 - 1725	1726 - 1

"fuck"

About 430 results (0.25 seconds)

Jan 1, 1650–Dec 31, 1676 › **Search English pages**

The reformed common-wealth of bees: presented in severall letters ... - Page 28

Samuel Hartlib - 1655 - 102 pages - Full view

... communicated to the dew, so that to **fuck** such clusters of Floures ina morning is almost as pleasant as to **fuck** a Honey.combe for taste- ... sowing Anise at several times , for it is the floure oue. ly of it, which the Bees **fuck** on. ...

books.google.com - More editions - Add to My Library ▼

Four books on the eleventh of Matthew: viz. I, Christ inviting ... - Page 741

1659 - Full view

... bring us to th% Heavenly Canaan* and the Saints c hey may **fuck** Mi lk and Honey continually, ... wicked have nothing but Swill and Dogs meat, to **fuck**, and to feed upon. And here is the difference ...

books.google.com - Add to My Library ▼

A supplement to the morning-exercise at Cripple-gate: or, Several ... - Page 341

Samuel Annesley - 1674 - 887 pages - Full view

... or greater charity > namely when the impediment to her giving **fuck** is natural ... and yet gave **fuck** to a Puppy, that her milk might be more artificially dried up. ...

books.google.com - More editions - Add to My Library ▼

The morning-exercise at Cripple-gate: or, Several cases of ... - Page 590

No cover image

1671 - 648 pages - Full view

... which themselves do cot partake of, doth soon improve into malice also a-gainst that natural humane life which themselves are also partakers- of ; their desires to **fuck** the blood, as I may so fey, of good mens souls and graces, ...

books.google.com - More editions - Add to My Library ▼

of; their desires to fuck the blood, as I may so say, of good mens souls and graces, makes them delight to fuck the blood of their bodies; witness *Cain*, the first that learnt this bloody trade, by killing his brother

OCR errors matter

- Letter б (b) for the digit 6 (six)
- Letter o for the digit 0 (zero)
- Letter з (z) for the digit 3 (three)
- Frequently confused в (v) and б (b).
- Very frequently confused и, н, and п. Some of these could be corrected through global search and replace operations (e.g., of those three letters, only и can stand alone as a word, only н occurs before a in a two-letter word) or global search with verification before replacement (e.g., in word-initial position before o, п is overwhelmingly the most likely, н is possible, and и is impossible).
- Mistook м for iл. This was easily corrected with a global search and replace operation.
- Inconsistently recognized ы, often making it ьi.

(David Birnbaum, OCR report for *Bdinski Sbornik)*

New OCR report

New OCR report

- 1. Improve statistical analyses (tools for training and adapting post-OCR correction models; look at impact on error rates on common text mining methods)

New OCR report

- 1. Improve statistical analyses (tools for training and adapting post-OCR correction models; look at impact on error rates on common text mining methods)
- 2. Formulate standards for annotation & evaluation of document layout

New OCR report

- 1. Improve statistical analyses (tools for training and adapting post-OCR correction models; look at impact on error rates on common text mining methods)
- 2. Formulate standards for annotation & evaluation of document layout
- 3. Use existing digital editions for training & test data

New OCR report

…

New OCR report

...

- **6. Train & test OCR on linguistically diverse texts.**

New OCR report

...

- **6. Train & test OCR on linguistically diverse texts.**
- **7. Convene OCR institutes in critical research areas.**

Word segmentation

2 Chinese Word Segmentation

The state-of-the-art Chinese Word Segmentation (CWS) algorithms are bi-directional LTSMs [28]. CWS is an important yet challenging pre-processing step for Chinese NLP; this is because both characters and words carry semantic meaning, which are sometimes related, and sometimes unrelated. For example, related characters and words include

Char	游	to swim (verb)
Char	泳	swimming (noun)
Word	游泳	to swim (noun/verb)

Char	微	micro
Char	波	wave
Char	爐	oven
Word	微波爐	microwave oven

while unrelated characters and words include

Char	香	fragrant
Char	港	harbor
Word	香港	Hong Kong

and

Char	幽	dark or quiet
Char	默	silently or secretly
Word	幽默	humor(ous)

In addition, the same sentence can sometimes be segmented differently and still remain grammatically correct, making CWS dependent on context. For example:

結婚/的/和尚/没/結婚/的/學生。	A married monk and an unmarried student.
結婚/的/和/尚没/結婚/的/學生。	Married and unmarried students.

Let's give it a try!