



Word Spotting: Indexing Handwritten Manuscripts

R. Manmatha

Multimedia Indexing and Retrieval Group
Center for Intelligent Information Retrieval
University of Massachusetts, Amherst

George Washington's manuscripts

58

Mr. Bantidges answer to Mr. Linnens enclosed, is left open for your perusal, and may be delivered to him, or not, and at any time you may find it convenient. —

As the post hour is at hand and I have many letters to close before the mail — I hardly know what I have written or whether you will be able fully to comprehend my meaning. My love in which Mrs. Washington's letter is presented to Henry the Children, and

I am dear Sir
Your aff. serv^t
G. Washington

Mr. Bantidge's answer to Mr. Linnens, Philad^a 21th March 1776 —

Dear Sir,

Presuming that you have received my last sent thro' the hands of Mr. Keith, with such aid as he was able to afford you respecting my administration of the Public Estate, and supposing as the Chancery term is closed or about to close, that the enclosed letter from that Gentleman, would not get to your hands in time to influence your measure, I resolved at first not to send it. — But upon second thoughts have changed my mind, leaving it to chance & your own judgment to give it the best effect of which the information is susceptible. — My best wishes attend you Mrs. Washington and with sincere regard & friendship

I am — Your aff. serv^t
G. Washington

165

Letters in 1758.

d. and to prevent this advantageous Commerce from suffering in its infancy by the sinister views of designing, selfish men, of the different Provinces — I humbly conceive it absolutely necessary, that Commissioners from each of the Colonies be appointed, to regulate the course of that Trade, and fix it on such a basis that, all the attempts of one Colony to diminish, ^{and} thereby weakening and diminishing the general system, might be frustrated. To effect which the General would (I fancy) cheerfully give his aid. —

Altho' none can entertain a higher sense of the great importance of maintaining a Post upon the Ohio than myself, yet under the unhappy circumstances that my Regiment is, I could by no means have agreed to leave any part of it there, had not the Governor an express order from the Lords would, to shew that the King's Troops ought to garrison it; but he told me, as he had no instructions from the Ministry relative thereto, he could not obey it —

and our men that are left there, are in such a miserable situation, having hardly rags to cover their nakedness — exposed to the inclemency of the weather in this rigorous season, that unless provision is made by the Country for supplying them immediately, they must inevitably perish! and, if the First V. Regiment

George Washington's manuscripts

165

Letters in 1758.

it. and to prevent this advantageous
 Commerce from suffering in its infancy
 by the sinister views of designing, selfish
 men, of the different Provinces. I hum-
 bly conceive it absolutely necessary, that
 Commissioners from each of the Colonies
 be appointed, to regulate the trade of
 that Trade, and fix it on such a basis
 that, all the attempts of one Colony en-
 davoring, ^{subtly} and thereby weakening and
 diminishing the general System, might
 be prevented.



Collection

- Have roughly 6,400 scanned pages of George Washington's manuscripts from the Library of Congress.
- Scanned in 8 bit graylevel at 300dpi.
- Scanned from microfilm
 - Quality not as good as scanning from original.
 - » For example, boundary artifacts, noise etc.
 - Probably done for reasons of cost, fragility of manuscripts and security.



Word spotting: Indexing Handwritten Documents

- Index historical documents written by a single author
 - Make an index like one at the back of a printed book.
- Examples
 - Presidential papers at the Library of Congress.
 - W.E.B. Dubois collection at Umass.
 - Margaret Sanger's correspondence at Smith and NYU.
- Variation in a single author's writing is small.
- R. Manmatha and W. B. Croft, *Wordspotting: Indexing Handwritten Manuscripts*, in *Intelligent Multimedia Information Retrieval*, ed. Mark Maybury, AAAI/MIT Press, 1997.



One Possible Approach

- Recognize words (e.g. optical character recognition – OCR).
- Use text indexing and retrieval.
- Handwriting recognition is a hard unsolved problem.



Word Spotting: Approach

- Match words as images.
 - Avoid Recognition as much as possible.
- Algorithm
 - Segment page into words.
 - Create classes (lists) containing all instances of similar ‘looking’ words in the document.
 - Words in a class contain links to the original pages.
 - User identifies single template word from each class and provides ASCII equivalents.
- Difficult parts
 - Segmentation of a page into words.
 - Matching words.

Senior Document

Now what tipped for number ten? by Walter Terry
 With one mighty spurt Mr Selwyn Lloyd has dashed from his
 rut and is now in the race for real power within the Conservative
 party. In so unobscured a contest the most difficult task is to judge
 one's times properly. Mr Lloyd has done this superbly with his budget
 line he was a non-starter today he is running well along the track
 towards number ten Downing Street. But wait a minute - Selwyn Lloyd,
 the little Liverpool lawyer, as he was contemptuously described a few years
 back, as prime minister? Laughable they used to say the man could
 hardly make a decent speech, puffing and pondering over a dreary brief
 Dominant. But Mr Lloyd as Prime Minister is ridiculous no more. The
 very thought, I am sure, has struck Mr R. A. Butler, home secretary and
 apparently the heir to Downing Street. For Mr Lloyd old nerves gone
 and seemingly dominant for the first time in his political career has made
 a tremendous impact on the Tories of Westminster with his budget.
 Maybe they don't see some of its details, especially the general tax. But
 the key significance of it is that for the first time in ten years of

Now what tipped for number ten? by Walter Terry
 With one mighty spurt Mr Selwyn Lloyd has dashed from his
 rut and is now in the race for real power within the Conservative
 party. In so unobscured a contest the most difficult task is to judge
 one's times properly. Mr Lloyd has done this superbly with his budget
 line he was a non-starter today he is running well along the track
 towards number ten Downing Street. But wait a minute - Selwyn Lloyd,
 the little Liverpool lawyer, as he was contemptuously described a few years
 back, as prime minister? Laughable they used to say the man could
 hardly make a decent speech, puffing and pondering over a dreary brief
 Dominant. But Mr Lloyd as Prime Minister is ridiculous no more. The
 very thought, I am sure, has struck Mr R. A. Butler, home secretary and
 apparently the heir to Downing Street. For Mr Lloyd old nerves gone
 and seemingly dominant for the first time in his political career has made
 a tremendous impact on the Tories of Westminster with his budget.
 Maybe they don't see some of its details, especially the general tax. But
 the key significance of it is that for the first time in ten years of



Classes

Class	Class	Class	Class
the	Lloyd	Minister	Significance
the	Lloyd	minister	
the	Lloyd		
the	Lloyd		
the	Lloyd		
the			



SLH Rankings for “Lloyd”



First image template.

Others matches in rank order under an “affine” transformation.



People Working

- Previously
 - Nitin Srimal – Grad Student.
- Since grant started
 - Joshua Sarro, Eric Mulvihill and Liz Yon – Undergraduate Students.
 - Shaun Kane, Andrew Lehman, Elizabeth Partridge – REU's (site REU award)
 - Fangfang Feng – staff programmer



Why is Word Spotting Difficult?



Handwriting style

Cursive Discrete

Words are non uniformly
spaced in this sentence

Words not uniformly spaced



scale scale

Words at different scales

Image
Picture

Ascenders and descenders connected together



words and words of

Orientation of words



Noise and degradation



Page Segmentation

- A novel scale space technique to segment a handwritten manuscript page into words.
 - Can handle variations in spacing, size of words, speckle noise.
- Intuition
 - First segment pages into lines.
 - Smooth (“smear”) line images to create blobs.
 - At an appropriate scale, the blobs will correspond to words.
 - Automatic determination of scale is important.



Assumptions

- Documents are gray level.
- Words are mostly horizontal and are written in lines.
- The spacing between words is more than the spacing between characters.



Segmentation Algorithm

- Pre-processing
 - Clean-up image.
- Line Segmentation
 - Projection profile.
- Blob Analysis
 - Finding appropriate scale for extracting words.
- Word Extraction and Post-Processing
 - Extract words.
 - Use connected components to link ascenders and descenders
- *R. Manmatha & N. Srimal, "Scale space technique for word segmentation in handwritten manuscripts," Proc. 2nd Intl. Conf. on Scale-Space Theories in Computer Vision (Scale Space 99), pp. 22-33, Springer-Verlag, 1999*

Results of Pre-Processing

165

Letters in 1758.

d. and to prevent this advantageous commerce from suffering in its infancy by the sinister views of designing, selfish men; of the different Provinces - I humbly conceive it absolutely necessary, that Commissioners from each of the Colonies be appointed, to regulate the mode of that Trade, and fix it on such a basis that, all the attempts of one Colony ^{and the} diminishing, and thereby weakening and diminishing the general system, might be frustrated. To effect which the General would (I fancy) cheerfully give his aid -

Altho' none can entertain a higher sense of the great importance of maintaining a Post upon the Ohio than myself, yet under the unhappy circumstances that my Regiment is, I would by no means have agreed to leave any part of it there, had not the Governor an express order for it. I intended to shew that the Kings Troops ought to garrison it; but he told me, as he had no instructions from the Ministry relative thereto, he could not do it -

and our men that are left there, are in such a miserable situation, having hardly rags to cover their nakedness - exposed to the inclemency of the weather in this rigorous season, that unless provision is made by the Country for supplying them immediately, they must inevitably perish! and, if the First V. Regiment

165

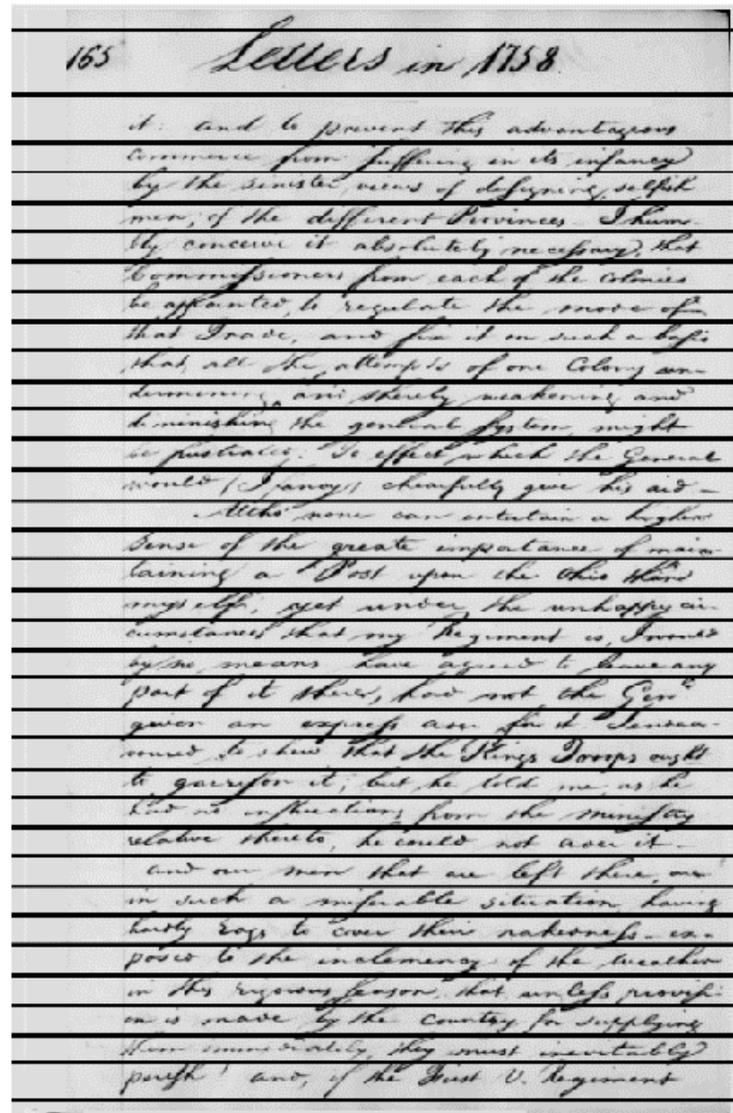
Letters in 1758.

d. and to prevent this advantageous commerce from suffering in its infancy by the sinister views of designing, selfish men; of the different Provinces - I humbly conceive it absolutely necessary, that Commissioners from each of the Colonies be appointed, to regulate the mode of that Trade, and fix it on such a basis that, all the attempts of one Colony ^{and the} diminishing, and thereby weakening and diminishing the general system, might be frustrated. To effect which the General would (I fancy) cheerfully give his aid -

Altho' none can entertain a higher sense of the great importance of maintaining a Post upon the Ohio than myself, yet under the unhappy circumstances that my Regiment is, I would by no means have agreed to leave any part of it there, had not the Governor an express order for it. I intended to shew that the Kings Troops ought to garrison it; but he told me, as he had no instructions from the Ministry relative thereto, he could not do it -

and our men that are left there, are in such a miserable situation, having hardly rags to cover their nakedness - exposed to the inclemency of the weather in this rigorous season, that unless provision is made by the Country for supplying them immediately, they must inevitably perish! and, if the First V. Regiment

Segmented Page Image





Blob Analysis

- Construct a scale space representation of a line image
 - Blob like features arise using this representation.
 - At the appropriate scale words form blobs.
- Create blobs by filtering with anisotropic second derivatives of Gaussians.

$$I(x, y; \sigma_x, \sigma_y) = G_{xx}(\cdot; \sigma_x) * f(x, y) + G_{yy}(\cdot; \sigma_y) * f(x, y)$$

$$= G_{xx}(\sigma_x, \sigma_y) * f(x, y) + G_{yy}(\sigma_x, \sigma_y) * f(x, y)$$

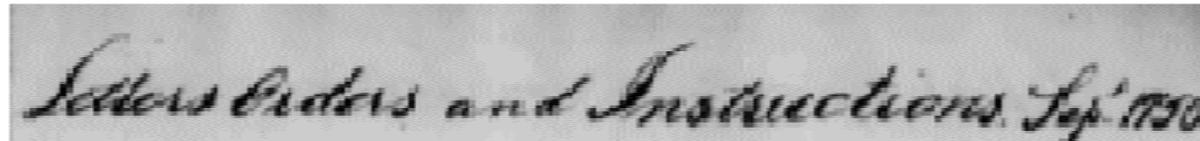
$f(x, y)$: Image

G_{xx} & G_{yy} : Second order Gaussian derivative s

σ_x, σ_y

Blob Analysis

Blobs at a few scale samples



Original
line image



?_y ????
?_x ????



?_y ????
?_x ????

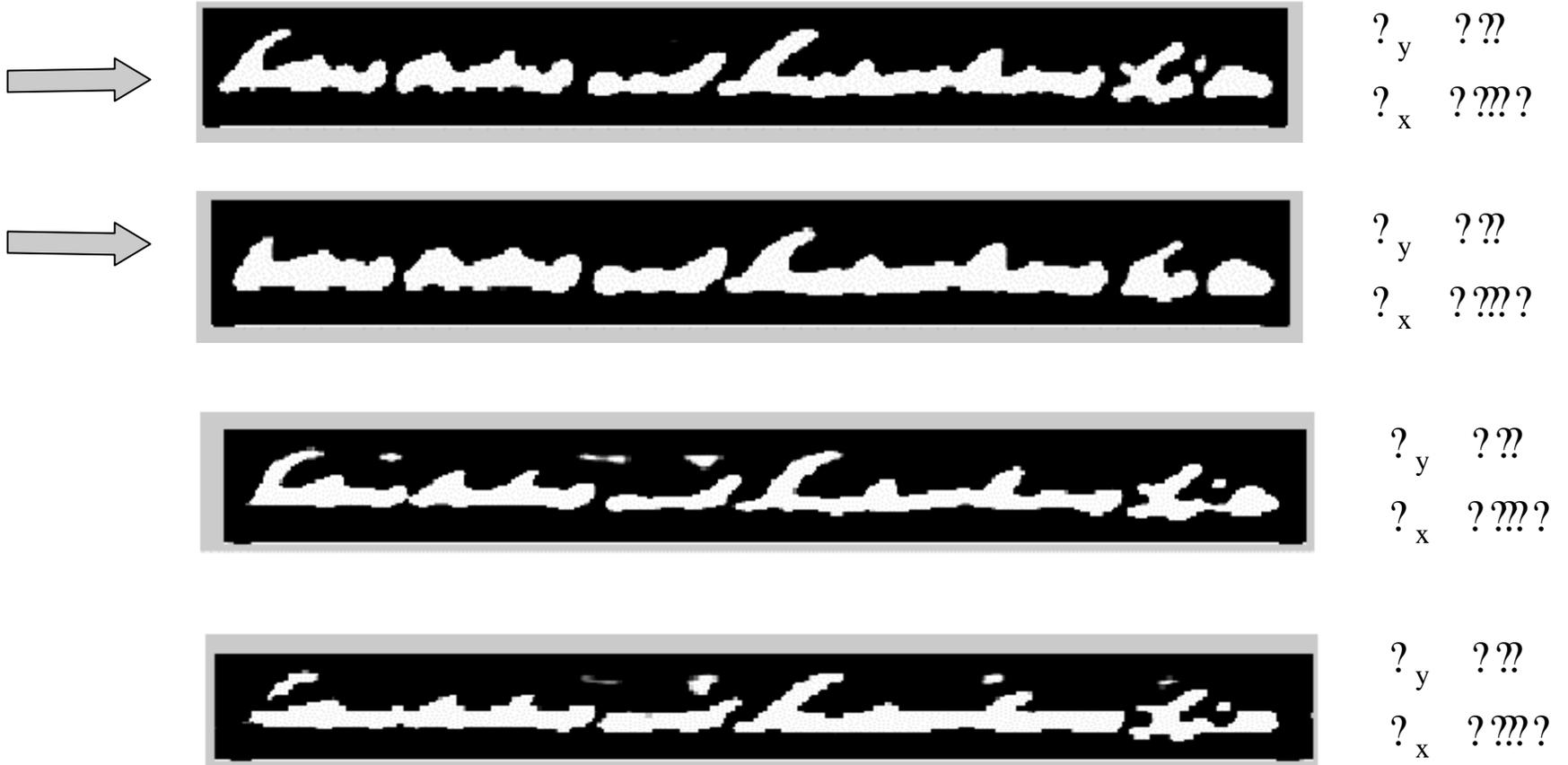


?_y ????
?_x ????

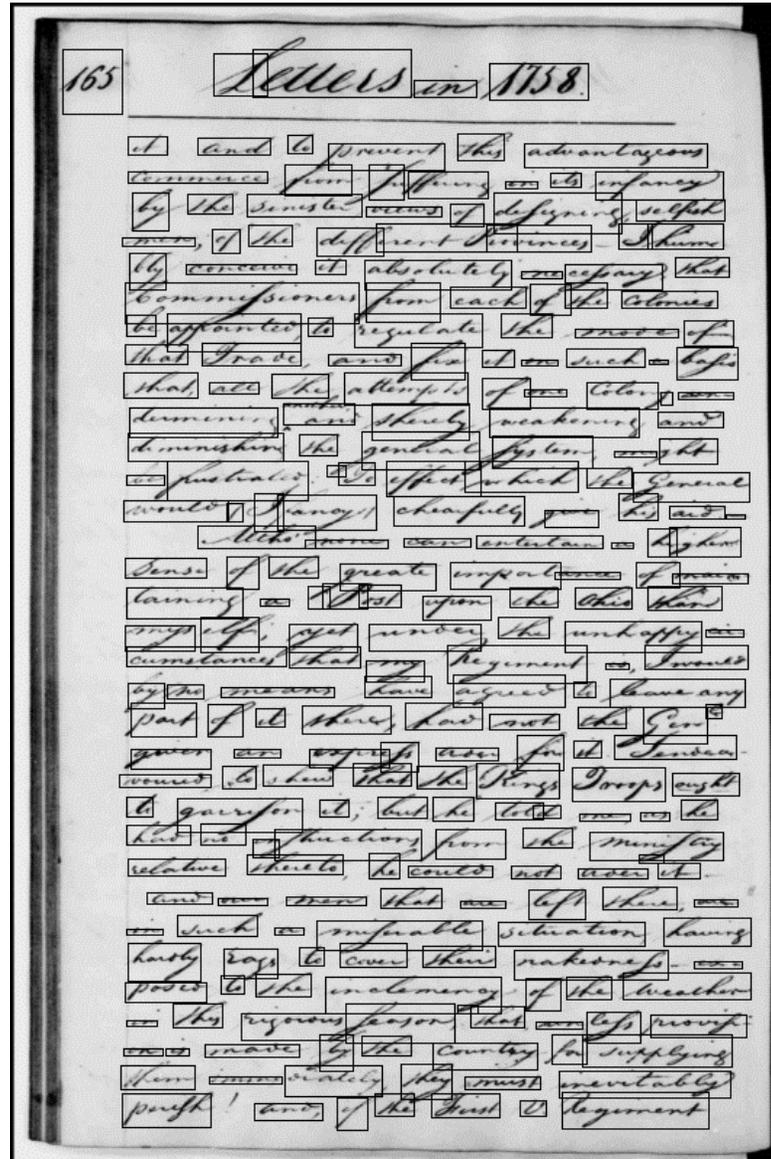


?_y ????
?_x ?????

Blob Analysis



Segmented Page





Hudson Document

Papers of:

Erasmus Darwin Hudson
Pioneer Orthopedic Surgeon
and Abolitionist

Source: Univ. of
Massachusetts @ Amherst
Library.

New York, Jan 20 '42

Dear Doctor

I have had a letter written for
me, but not knowing when to send it, kept it in
my hat. In this, I shall say over, very briefly the
substance of that; as indeed that is all I have to
write about at present. Our Standard concerns are
got into terrible embarrassed forms, if we do not employ
a special agent to attend to them. Subscriptions, when
due to a large amount, only need to be called for.
We have no system in our business, as it regards
this matter. The trusting to agent, don't, & can't be
made to meet our wants. Will you take charge
of this branch - & assume the General Agency for the
Standard? In fact, this is the only sort of General
Agent that we want & you are the only man I
know of, capable of doing justice to it. Salary ought
to be 800 dollars, or more if that amount except travel;
expenses to be paid by the Society, of course. Your
business would be to come to N. Y. & make an
arrangement with the agents of the Standard, a book or books of
all subscribers, in such order as to be able to refer
& tell immediately whether they have paid up -
what they owe, - when their subscription expires, &c.



Results

No. of pages	Avg. no. words/page	% words detected	% words fragmented	% words combined	% words correctly segmented
30	220	99.12	1.75	8.9	87.6



Recent Work - Segmentation

- Fixed bugs and some problems with algorithm.
- Recently have segmented successfully the 6400 scanned images of George Washington that we have.
 - Its impractical (too labor intensive) to compute segmentation statistics on the entire collection.
 - However, anecdotal evidence seems to imply that the segmentation is similar to that on the sample.



Matching Techniques

- Some matching methods.
 - Translation Invariant (EDM)
 - » Euclidean Distance Maps
 - Matching Under Affine Transformations (SLH)
 - » Scott and Longuet Higgins algorithm
 - EDM and SLH seem to work reasonably well.
 - Problem
 - » Requires matching every word against every other word
 - » $O(N^2)$.
 - » $N^2 = 220 \times 6400 \sim 10^{12}$ matches !!!



“Indexable” Matching Techniques

- Prune search space.
- Find matching methods which have signatures.
- Tried different techniques based on different features
 - Number and Position of Ascenders and Descenders, Projection Profiles etc
 - Each individual feature not v. good.
- Invariant Moments
 - Compute different moments of binary images invariant to similarity transformations.
 - Doesn't work well.



Other Possible Approaches

- Combine lots of features.
 - Build a classifier with many features.
 - Combine (fuse) classifiers each based on one or more features.
- Clustering
 - Cluster similar words together in some high-dimensional feature space.
 - K-means, etc
- Choice of features is still tricky.



Other Possible Approaches.

- Use a language model to constrain which word images match.
 - Language model – probabilities of word occurrence in a corpus.
 - We don't have ASCII words unlike typical OCR applications.
 - However, we can still have constraints on length, ascenders and descenders etc. by looking at machine generated fonts.
 - Some question of what corpus is appropriate for computing statistics.
 - » Washington spelt *expense* as *expence* .
 - Maybe a mixture model of a modern corpus and an old one.



Conclusion

- Have a reasonable segmentation algorithm.
- Currently working on effective matching techniques.
 - Hard Problem.
- Later, integrate into an indexing scheme.