Projektseminar: NLP/Text Mining for Historical Texts

Winter Term 2008/09

Torsten Kopp, Katrin Schmidt, Hüseyin Mergan,
Alexander Volokh

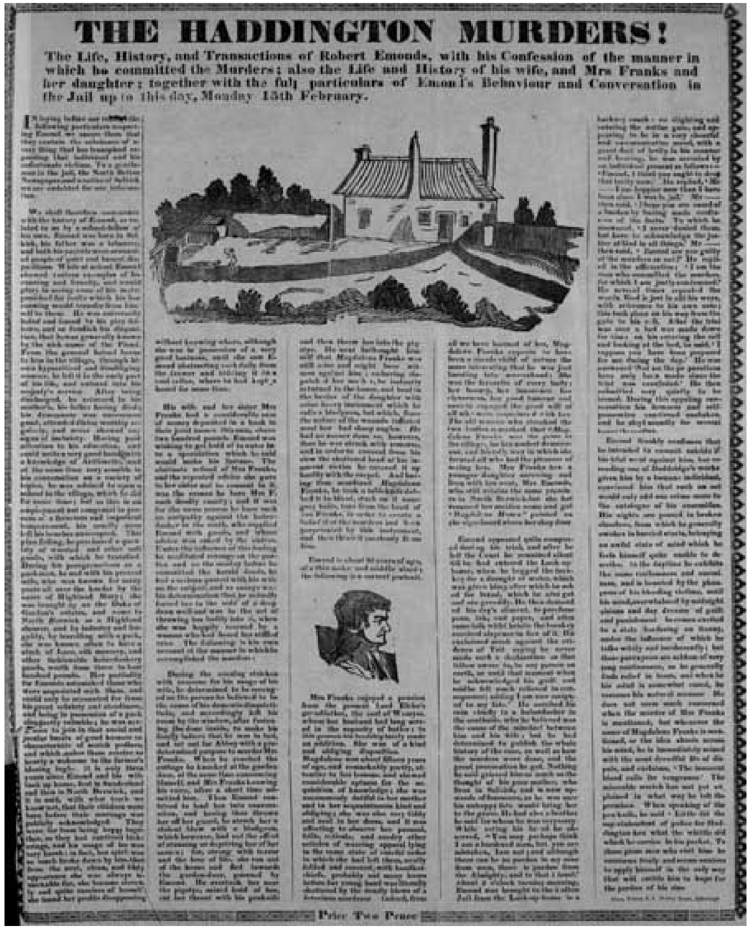
Information Extraction from Broadsides

Results of the Group Work

What are the *Broadsides*?

- Single sheet
- Street Literature
- Similar to modern tabloid press

Example:



THE HADDINGTON MURDERS!

The Life, History, and Transactions of Robert Emonds, with his Confession of the manner in which he committed the Murders; also the Life and History of his wife, and Mrs Franks and her daughter; together with the full particulars of Emond's Behaviour and Conversation in the Jail up to this day, Monday 15th February.

[illegible]

Price Two Pence

http://www.nls.uk/broadsides/broadside.cfm/id/15552/criteria/haddington

Material:

- 240 articles concerning *murder*
- Transcriptions in Html
- Uploaded at www.nls.uk/broadsides/
- Number of texts: 157
- Number of words: 125,139
- Size of the files: 1 MB
- Most from 1820 to 1837

Questions:

- Which instruments of murder are mentioned?
- Which places are mentioned?
- Why are most articles from 1820 to 1837?

Historical Context:

- Economical Crisis in Europe
- Political Crisis in England and Scotland

Also:

- Printing becomes easier and cheaper
- Rise of the newspapers

George IV. (1762-1830)

- Prince Regent since 1811
- 1820-1830 King of UK and Hannover
- Very young
- Many scandals
- Inable in politics

http://commons.wikimedia.org/wiki/File:George_IV_van_het_Verenigd_Koninkrijk.jpg

William IV. (1765-1837)

- 1830-1837 King of UK and Hannover
- *Sailor King*
- Not prepared for reign

http://en.wikipedia.org/wiki/File:Martin_Archer_Shee_-_King_William_IV_-_c.1800.jpg

Victoria (1819-1901)

- 1837-1901 Queen of UK, since 1876 Empress of India
- Young and unexperienced
- Political situation consolidated

http://commons.wikimedia.org/wiki/File:Victoria_in_her_Coronation.jpg

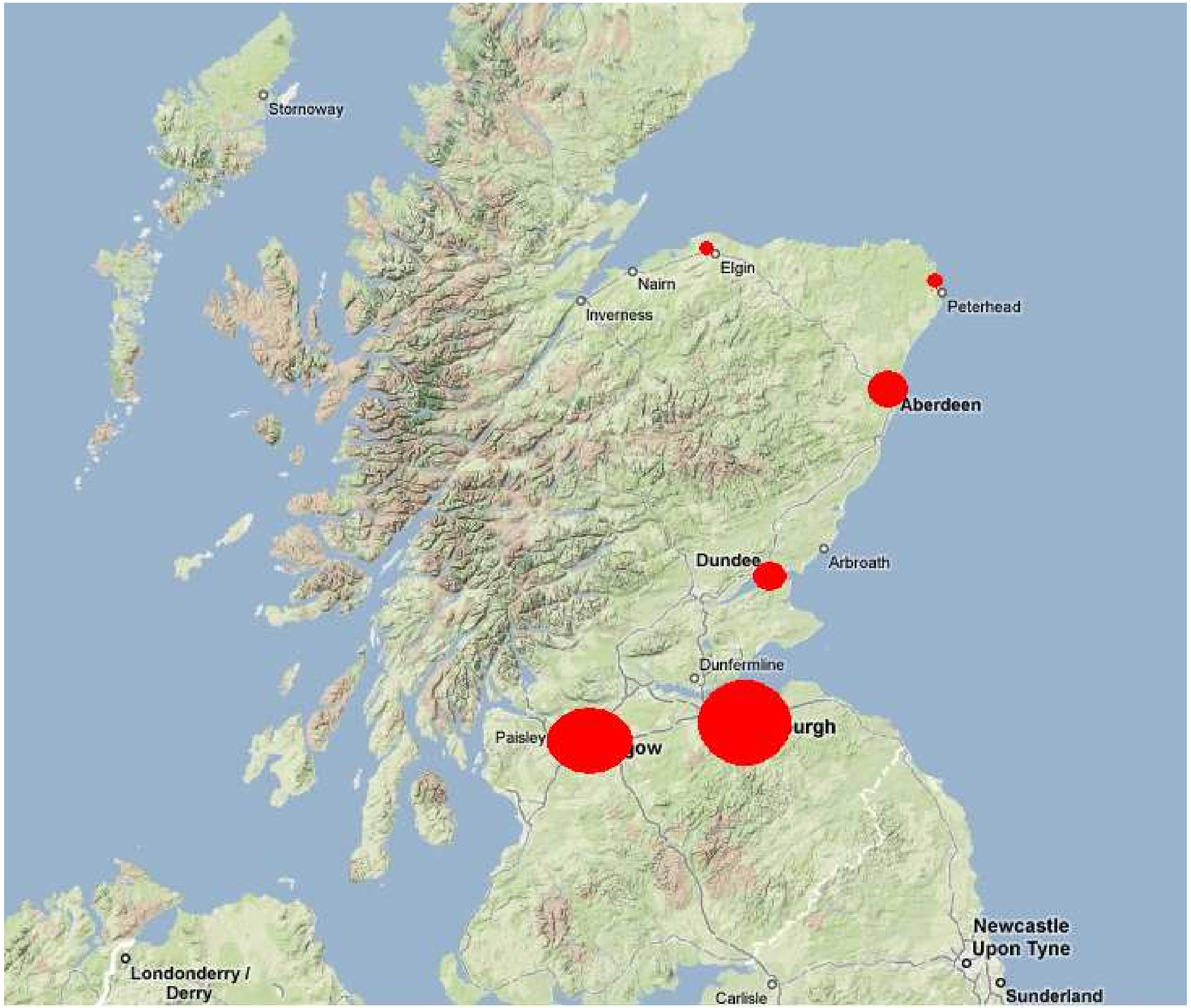
Opportunities:

- List of popular murder tools
- ‚Murder Map‘ of Scotland
- Comparison to English / Scottish fiction of that time
- Comparison to Scottish newspapers of that time

…and how can CoLi help?

- Extraction of articles correlating with your task :
  - For example „How many articles mention Edinburgh as a place of murder in the *Broadsides*?“
  - Image of certain places in the public media in that period
- *„murder-map“* of Scotland in 19th century

Example:


Stornoway
Elgin
Nairn
Inverness
Peterhead
Aberdeen
Dundee
Arbroath
Dunfermline
urgh
Paisley
gow
Newcastle Upon Tyne
Sunderland
Carlisle
Londonderry / Derry

…and how can CoLi help?

- Extracting Names
  - Gender of victims and murderers
  - Search for a certain person
  - Surnames in case of migration stories
- Listing all weapons and kinds of murder

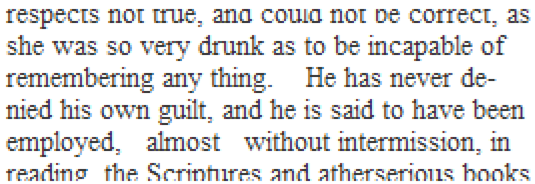
Task

- Information Extraction
- - manners of murdering
- Domain-specific
- - only “Murder“ domain
- Unsupervised
- - no annotations and other human intervention
- Open challenge
- - all possible external resources can be used
- Limited in time
- - Only 10 days for the whole task

Preparation

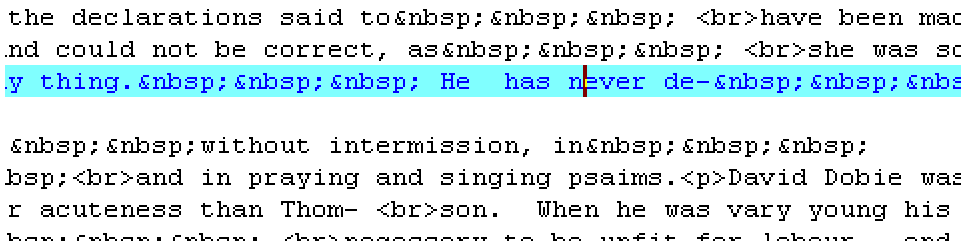
- Extraction of texts and conversion into best format for the task:

  html → normalized text
- Done with scripts developed by group



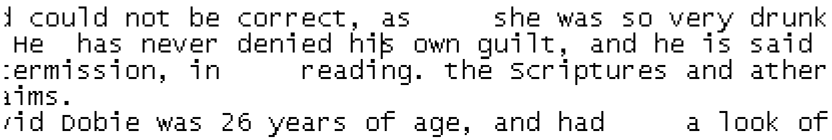
respects not true, and could not be correct, as
she was so very drunk as to be incapable of
remembering any thing.   He has never de-
nied his own guilt, and he is said to have been
employed,   almost   without intermission, in
reading the Scriptures and atherserious books

HTML



```
the declarations said to&nbsp;&nbsp;&nbsp; <br>have been mac
.nd could not be correct, as&nbsp;&nbsp;&nbsp; <br>she was sc
.y thing.&nbsp;&nbsp;&nbsp; He  has never de-&nbsp;&nbsp;&nbs

&nbsp;&nbsp;without intermission, in&nbsp;&nbsp;&nbsp;
bsp;<br>and in praying and singing psaims.<p>David Dobie was
r acuteness than Thom- <br>son.  When he was vary young his
```



```
d could not be correct, as     she was so very drunk
 He  has never denied his own guilt, and he is said
ermission, in     reading. the Scriptures and ather
aims.
vid Dobie was 26 years of age, and had     a look of
```

Processed Format

Data

- Problems:
- Broadsides:
- Language of the 19th century
- Stylistically and statistically different from resources based on WSJ or the like
- Very diverse content
- Large amount of data:
- Difficult to preprocess
- Difficult to make an impression of the data

Approach

- General Idea of the Approach:
- Shorten the data by deleting irrelevant parts
- Preprocess the shortened version
- Use syntactic and semantic information to retrieve relevant relations

Extracting Synonyms

- First strategy (offline):
- Define few seed words that you think are relevant (e.g. to murder, knife)
- Look at the patterns in which seed words occur in the data
- Query a search engine with this pattern in order to find other words that can occur in the same context
- Keep semantically related and filter out semantically unrelated results by using WordNet

Exctracting Synonyms

- Second strategy (online):
- Judging the relevance of words according to WordNet is done online during reading in the data

- Advantages: easier to get the desired list of synonyms, since one inevitably considers all the relevant words (whereas search engines definitely won‘t find some of the peculiar words)
- Disadvantages: poor runtime performance

Examples

- We take “to murder, to stab, to kill“ as seeds
- We find the following occurrences in the data:

  …of them was killed on the spot…

  …after he had murdered the deceased he…
- We substitute the seed word by a wildcard:

  “of them was * on the spot”

  “after he had * the deceased he”
- And query Yahoo!:

  [fired, put, killed, fixed, shot, jenny, executed, arrested, trampolening, stopped, there, caught, dead, built, standing]

  [hit, stabbed, shot, assaulted]

  sets of possible words out of first 100 results occurring in place of the wildcard

WordNet

- WordNet® is a large lexical database of English
- Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept
- Synsets are interlinked by means of conceptual-semantic and lexical relations

Semantic Similarity

- We use WordNet in order to judge whether a word is similar to our seeds or not:
- We use two different similarity measures:

- Resnik (Resnik)

Similarity between two concepts is the information content of their lowest super-ordinate taking into account the probability of encoutering an instance of a synset of the concept

murder#v#1 enjoy#v#3 0.061

murder#v#1 execute#v#2 0.33

murder#v#1 slay#v#1 2.02

- Jiang and Conrath (JC)

Information content of both concept nodes and of their lowest super-ordinate is taken into account

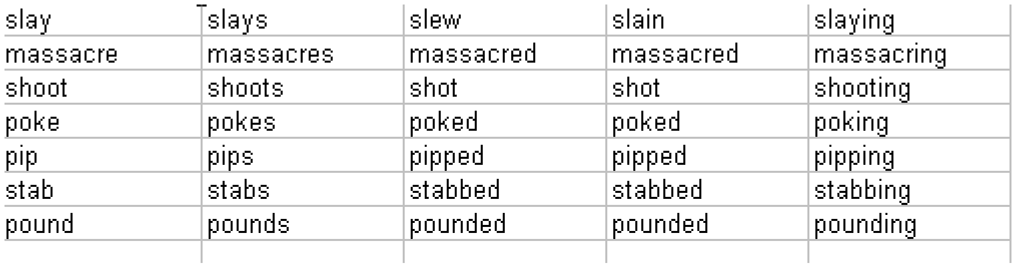
murder#v#1 enjoy#v#3 0.0

murder#v#1 execute#v#2 0.84

murder#v#1 slay#v#1 1.0

Morphological Variants

- Morphological analysis tool: RASP
  - http://www.informatics.susx.ac.uk/research/groups/nlp/rasp/
- Output format: inflected forms of given verbs (third person, past simple, past participle, present participle).



slay	slays	slew	slain	slaying
massacre	massacres	massacred	massacred	massacring
shoot	shoots	shot	shot	shooting
poke	pokes	poked	poked	poking
pip	pips	pipped	pipped	pipping
stab	stabs	stabbed	stabbed	stabbing
pound	pounds	pounded	pounded	pounding

Sample Result

- Original (contains 28 sentences)
- Processed version contains only 6 sentences
- Remaining sentences are parsed with a dependency parser
- A lot of fine distinctions of different objects are obtained

Preprocessing

- Done by external tools:
  - Tools used:
    - POS tagging script (Python NLTK)
    - Named Entitiy Recognition (Stanford NER Tool)
    - Parsing (Stanford Parser)

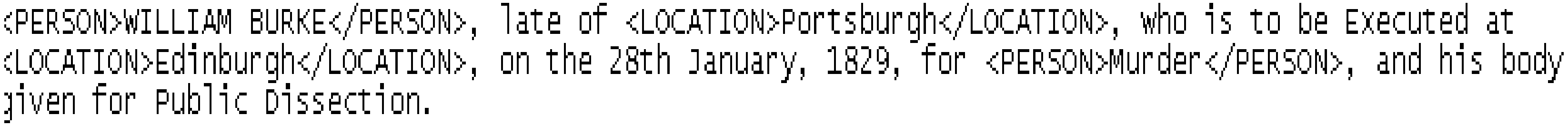
POS Tagging:

- Tools: Python Script based on Natural Language Processing Tool Kit
  - www.nltk.org
- Output: tagged tokens:

  At_IN this_DT reproach_NN he_PRP called_VBD her_PRP by_IN the_DT most_RBS horrid_JJ names_NNS ,_, and_CC beat_VBD and_CC kicked_VBD her_PRP shockingly_RB ._. She_PRP had_VBD indeed_RB a_DT miserable_JJ time_NN of_IN it_PRP ._.

Preprocessing: Named Entity Recognition

- Tools: Stanford Named Entity Recognizer
  - http://nlp.stanford.edu/software/CRF-NER.shtml
  - *Nota Bene:* ANNIE and Open Calais were tested, but they do not return efficient result for our task.
- Output: Tagged txt (Person, Location).



```
<PERSON>WILLIAM BURKE</PERSON>, late of <LOCATION>Portsburgh</LOCATION>, who is to be Executed at
<LOCATION>Edinburgh</LOCATION>, on the 28th January, 1829, for <PERSON>Murder</PERSON>, and his body
given for Public Dissection.
```

Preprocessing: Parsing

- Tool: Stanford Parser (statistical parser)
  - http://nlp.stanford.edu/software/lex-parser.shtml
- Motivation: “Dependencies representation was designed to provide a simple description of the grammatical relationships in a sentence that can easily be understood and effectively used by people without linguistic expertise who want to extract textual relations“
- Output: dependency-based parses of sentences

Dependency Parse

- … he/PRP appointed/VBN to/TO meet/VB her/PRP at/IN Lime/NNP Wood/NNP ,/, where/WRB ,/, in/IN a/DT lonely/JJ spot/NN ,/, he/PRP stabbed/VBD her/PRP with/IN a/DT knife/NN …

  nsubj(appointed-47, he-46)
  conj_and(п-28, appointed-47)
  xcomp(appointed-47, meet-49)
  nn(Wood-53, Lime-52)
  advmod(stabbed-63, where-55)
  dep(stabbed-63, in-57)
  amod(spot-60, lonely-59)
  nsubj(stabbed-63, he-62)
  rcmod(Wood-53, stabbed-63)
  det(knife-67, a-66)

  xsubj(meet-49, he-46)
  aux(meet-49, to-48)
  dobj(meet-49, her-50)
  prep_at(meet-49, Wood-53)
  rel(stabbed-63, where-55)
  det(spot-60, a-58)
  pobj(in-57, spot-60)
  nsubj(threw-71, he-62)
  dobj(stabbed-63, her-64)
  prep_with(stabbed-63, knife-67)

Using WordNet Definitions

- In case no potential weapon can be found as object of the verb(e.g. verb implicitly contains the weapon), we could back off to WordNet definitions of the verb:

- knife, stab (use a knife on)
- kick (strike with the foot)
- poison (kill with poison)

Example

- The/DT deceased/JJ Roger/NNP Haldon/NNP ,/, a/DT labourer/NN ,/, residing/VBD in/IN a/DT small/JJ cottage/NN near/IN Blackburn/NNP was/VBD poisoned/VBN by/IN his/PRP$ wife/NN ./.

- In this case one has to back off to the word definition.

Further Problems

- Several victims in the same article:
- Use NEs(and reference resolution) together with dependency relations: (fortunately victims are almost always named in the header of the article)
- E.g. Murder of Mrs. Franks and her daughter Magdalene

- Two different NEs appearing as objects:

poss(throat-15, her-14) (her standing for Mrs. Franks)

dobj(cut-13, throat-15)

prep_with(cut-13, penknife-18)

amod(Franks-6, murdered-4)

nn(Franks-6, Magdalene-5) (another weapon has to be found)