

# Sublanguage Theory for Leveraging the Unrealized Value in Trouble Tickets

Elizabeth D. Liddy

Center for Natural Language Processing  
School of Information Studies  
Syracuse University

December 12, 2005

# Center for Natural Language Processing

- Multidisciplinary team of information scientists, computer scientists, and linguists
  - Many with substantial commercial software experience
- Research, development and licensing of NLP-based technology for government, industry, and foundations
  - ARDA, DARPA, NSA, CIA, DHS, DOJ, NIH, NSF
  - Raytheon, SAIC, Boeing, ConEd, MySentient, Unilever, ModSpec
  - Robert Wood Johnson Foundation, OCLC
- For numerous applications:
  - Document Retrieval
  - Question-Answering
  - Information Extraction / Text Mining
  - Automatic Metadata Generation
  - Cross-Language Information Retrieval
- In a broad range of domains:
  - CRM, public health, aerospace engineering, math & science education, intelligence, patents, alternative medicine, security
  - 65 projects to-date

# Sublanguage Methodology

- Sublanguage Theory predicts that speech & texts produced within a community engaged in a specialized, common activity:
  - Deal with a circumscribed subject area
  - Share a common vocabulary
  - Exhibit common habits of word usage
  - Use deviant rules of grammar
    - A subset of the rules of the standard language
    - High frequency of certain, odd constructions
  - Fairly predictable text structure
  - Extensive use of special symbols & abbreviations
- Sublanguage Methodology is highly successful for developing the models for specializing the core NLP technology

# Sublanguages Studied

- Pharmacology reports
- Weather reports
- Technical manuals
- Cooking recipes
- Patents
- Stock market reports
- Patient medical histories
- Legal documents
- University catalogs
- Journal abstracts
- Life insurance applications
- Web pages

# Prototypical Sublanguage Methodology

1. Select a representative **SAMPLE** of texts from a specialized community.
2. Conduct a **DISTRIBUTIONAL ANALYSIS** of words in sample texts.
3. Determine **SEMANTIC WORD CLASSES** depending on their similarities of occurrence.
4. Define sublanguage **GRAMMAR** based on co occurrence patterns of sublanguage word classes.
5. Establish an **SUBLANGUAGE MODEL** based on sublanguage lexicon & grammar.
6. Specialize **NLP TECHNOLOGY** to accurately interpret new texts based on the sublanguage model.

# Trouble Ticket Problem

- Companies have large numbers of field reports of the problems that customers encounter with their products, services, or systems
  - ***Trouble Tickets***
- But companies are not fully leveraging the value of the data contained in these reports, including their company's responses
  - Fail to analyze & learn from trends that are only obvious when large repositories can be represented for rich data mining
  - Miss out on proactive insights vital to their business interests

# Trouble Ticket Problem (cont'd)

- Trouble Tickets are typically a combination of structured and unstructured sections
  - Structured portions - the range of language used is still quite varied
  - Unstructured portions of the reports – complaints, comment fields, remarks – exhibit even freer natural expression
  - Variety results from multiple inputers
- Current access to tickets is either keyword searches or SQL style database queries which under perform
  - Without NLP, system fails to capture multiple ways a common issue or solution occurs, thus resulting in under-counting
- Customer is largest utility provider in New York City
  - ConEdison

# Sample Raw Trouble Ticket

ME00007923

|001| CONST MGMT REPORTS SPARKING WIRE IN MH N/S SPRING ST  
|002| 55' E/O 12TH AVE (ON WALK) - CONTRACTORS ON LOCATION-MC  
|003| 06/08/00 23:16 MDERWILLIM DISPATCHED BY 48414  
|004| 06/08/00 23:17 MDERWILLIM ARRIVED BY 48414  
|005| 06/08/00 23:17 CREW PULLED OFF FOR OUTAGE.....CG  
|006| 06/08/00 23:18 MDERWILLIM UNFINISHED BY 48414  
|007| 06/09/00 15:00 MDEDONOHUE DISPATCHED BY 48414  
|008| 06/09/00 16:00 MDEDONOHUE ARRIVED BY 44729  
|009| 06/09/00 18:20 MDEDONOHUE REPORTS CLEARED MULTIPLE B/O'S  
|010| IN SB#46977 N/S SPRING ST 55'E/O 12TH AV READY FOR C.A.I -  
|011| 06/09/00 18:34 MDEDONOHUE COMPLETE BY 44729  
|012| 06/09/00 18:34 REFERRED TO: CAI EDSWBR FYI BY 44729  
|013| 06/10/00 14:10 NO C.M. ACTION REQD.===== BY 44979

# Example: "Complaint"



COMPLAINT

**ME00007923**

|001| CONST MGMT REPORTS SPARKING WIRE IN MH N/S SPRING ST  
|002| 55' E/O 12TH AVE (ON WALK) - CONTRACTORS ON LOCATION-MC  
|003| 06/08/00 23:16 MDERWILLIM DISPATCHED BY 48414  
|004| 06/08/00 23:17 MDERWILLIM ARRIVED BY 48414  
|005| 06/08/00 23:17 CREW PULLED OFF FOR OUTAGE.....CG  
|006| 06/08/00 23:18 MDERWILLIM UNFINISHED BY 48414  
|007| 06/09/00 15:00 MDEDONOHUE DISPATCHED BY 48414  
|008| 06/09/00 16:00 MDEDONOHUE ARRIVED BY 44729  
|009| 06/09/00 18:20 MDEDONOHUE REPORTS CLEARED MULTIPLE B/O'S  
|010| IN SB#46977 N/S SPRING ST 55'E/O 12TH AV READY FOR C.A.I -  
|011| 06/09/00 18:34 MDEDONOHUE COMPLETE BY 44729  
|012| 06/09/00 18:34 REFERRED TO: CAI EDSWBR FYI BY 44729  
|013| 06/10/00 14:10 NO C.M. ACTION REQD.===== BY 44979

# Example: "Office Action"

OFFICE\_ACTION

**ME00007923**


|001| CONST MGMT REPORTS SPARKING WIRE IN MH N/S SPRING ST  
|002| 55' E/O 12TH AVE (ON WALK) - CONTRACTORS ON LOCATION-MC  
|003| 06/08/00 23:16 MDERWILLIM DISPATCHED BY 48414  
|004| 06/08/00 23:17 MDERWILLIM ARRIVED BY 48414  
|005| 06/08/00 23:17 CREW PULLED OFF FOR OUTAGE.....CG  
|006| 06/08/00 23:18 MDERWILLIM UNFINISHED BY 48414  
|007| 06/09/00 15:00 MDEDONOHUE DISPATCHED BY 48414  
|008| 06/09/00 16:00 MDEDONOHUE ARRIVED BY 44729  
|009| 06/09/00 18:20 MDEDONOHUE REPORTS CLEARED MULTIPLE B/O'S  
|010| IN SB#46977 N/S SPRING ST 55'E/O 12TH AV READY FOR C.A.I -  
|011| 06/09/00 18:34 MDEDONOHUE COMPLETE BY 44729  
|012| 06/09/00 18:34 REFERRED TO: CAI EDSWBR FYI BY 44729  
|013| 06/10/00 14:10 NO C.M. ACTION REQD.===== BY 44979

# Example: "Office\_Note"



**OFFICE\_NOTE**

**ME00007923**

|001| CONST MGMT REPORTS SPARKING WIRE IN MH ~~N/S~~ SPRING ST  
|002| 55' E/O 12TH AVE (ON WALK) - CONTRACTORS ON LOCATION-MC  
|003| 06/08/00 23:16 MDERWILLIM DISPATCHED  BY 48414  
|004| 06/08/00 23:17 MDERWILLIM ARRIVED BY 48414  
|005| 06/08/00 23:17 CREW PULLED OFF FOR OUTAGE.....CG  
|006| 06/08/00 23:18 MDERWILLIM UNFINISHED BY 48414  
|007| 06/09/00 15:00 MDEDONOHUE DISPATCHED BY 48414  
|008| 06/09/00 16:00 MDEDONOHUE ARRIVED BY 44729  
|009| 06/09/00 18:20 MDEDONOHUE REPORTS CLEARED MULTIPLE B/O'S  
|010| IN SB#46977 N/S SPRING ST 55'E/O 12TH AV READY FOR C.A.I -  
|011| 06/09/00 18:34 MDEDONOHUE COMPLETE BY 44729  
|012| 06/09/00 18:34 REFERRED TO: CAI EDSWBR FYI BY 44729  
|013| 06/10/00 14:10 NO C.M. ACTION REQD.===== BY 44979

# Example: "Field\_Report"

## FIELD\_REPORT

**ME00007923**

|001| CONST MGMT REPORTS SPARKING WIRE IN MH N/S SPRING ST  
|002| 55' E/O 12TH AVE (ON WALK) - CONTRACTORS ON LOCATION-MC  
|003| 06/08/00 23:16 MDERWILLIM DISPATCHED BY 48414  
|004| 06/08/00 23:17 MDERWILLIM ARRIVED BY 48414  
|005| 06/08/00 23:17 CREW PULLED OFF FOR OUTAGE.....CG  
|006| 06/08/00 23:18 MDERWILLIM UNFINISHED BY 48414  
|007| 06/09/00 15:00 MDEDONOHUE DISPATCHED BY 48414  
|008| 06/09/00 16:00 MDEDONOHUE ARRIVED BY 44729  
|009| 06/09/00 18:20 MDEDONOHUE REPORTS CLEARED MULTIPLE B/O'S  
|010| IN SB#46977 N/S SPRING ST 55'E/O 12TH AV READY FOR C.A.I -  
|011| 06/09/00 18:34 MDEDONOHUE COMPLETE BY 44729  
|012| 06/09/00 18:34 REFERRED TO: CAI EDSWBR FYI BY 44729  
|013| 06/10/00 14:10 NO C.M. ACTION REQD.===== BY 44979

# Example: "Job\_Completion"

## JOB\_COMPLETION

**ME00007923**

|001| CONST MGMT REPORTS SPARKING WIRE IN MH N/S SPRING ST  
|002| 55' E/O 12TH AVE (ON WALK) - CONTRACTORS ON LOCATION-MC  
|003| 06/08/00 23:16 MDERWILLIM DISPATCHED BY 48414  
|004| 06/08/00 23:17 MDERWILLIM ARRIVED BY 48414  
|005| 06/08/00 23:17 CREW PULLED OFF FOR OUTAGE.....CG  
|006| 06/08/00 23:18 MDERWILLIM UNFINISHED BY 48414  
|007| 06/09/00 15:00 MDEDONOHUE DISPATCHED BY 48414  
|008| 06/09/00 16:00 MDEDONOHUE ARRIVED BY 44729  
|009| 06/09/00 18:20 MDEDONOHUE REPORTS CLEARED MULTIPLE B/O'S  
|010| IN SB#46977 N/S SPRING ST 55'E/O 12TH AV READY FOR C.A.I -  
|011| 06/09/00 18:34 MDEDONOHUE COMPLETE BY 44729  
|012| 06/09/00 18:34 REFERRED TO: CAI EDSWBR FYI BY 44729  
|013| 06/10/00 14:10 NO C.M. ACTION REQD.===== BY 44979

# Example: "Job\_Referral"

## JOB\_REFERRAL

**ME00007923**

|001| CONST MGMT REPORTS SPARKING WIRE IN MH N/S SPRING ST  
|002| 55' E/O 12TH AVE (ON WALK) - CONTRACTORS ON LOCATION-MC  
|003| 06/08/00 23:16 MDERWILLIM DISPATCHED BY 48414  
|004| 06/08/00 23:17 MDERWILLIM ARRIVED BY 48414  
|005| 06/08/00 23:17 CREW PULLED OFF FOR OUTAGE.....CG  
|006| 06/08/00 23:18 MDERWILLIM UNFINISHED BY 48414  
|007| 06/09/00 15:00 MDEDONOHUE DISPATCHED BY 48414  
|008| 06/09/00 16:00 MDEDONOHUE ARRIVED BY 44729  
|009| 06/09/00 18:20 MDEDONOHUE REPORTS CLEARED MULTIPLE B/O'S  
|010| IN SB#46977 N/S SPRING ST 55'E/O 12TH AV READY FOR C.A.I -  
|011| 06/09/00 18:34 MDEDONOHUE COMPLETE BY 44729  
|012| 06/09/00 18:34 REFERRED TO: CAI EDSWBR FYI BY 44729  
|013| 06/10/00 14:10 NO C.M. ACTION REQD.===== BY 44979

# Study Stages

1. Data Cleanup / Pre-processing
2. Annotation of sample of Trouble Tickets
3. Sublanguage Model development of Trouble Ticket content
4. Testing & analysis of automatic ticket representation
5. Machine Learning of model of Trouble Ticket content

# Stage 1: Data Cleanup / Pre-processing

- Dataset provided- 162,105 tickets:
  - *Data* file (structured data, generated semi-automatically)
  - *Remarks* file (free-text data, entered by the Call Center Operator)
  - From 2000 to 2005
- Pre processing steps:
  - Stripped TicketID and line # from each line
  - Converted non-ASCII characters
  - Added, to each ticket from *Remarks* file, selected data components (Original and Actual Trouble Types) from *Data* file
  - Converted tickets to XML format
    - XML markup for ticket sections & section components

# Stage 1: Developing the Tokenizer

- Adapted our tokenizer for the Con Edison data:
  - To cover the identified special features of Con Edison language, such as common misspellings and name variants, abbreviations, fixed phrases, and so on
    - Token = a term, including acronyms and fixed phrases, e.g.:  
*MANHOLE, NO ACCESS, I & A*
- Example:

26991 F/O 41-45 **BROADWAY** & CREW REPORTS THERE IS MULTIPLE  
26991|F/O|41-45|**BROADWAY** |&|CREW|REPORTS|THERE|IS|MULTIPLE|

04/28/03 10:30 DRISKILL REPORTS IN SB-29146 F/O 4146 **B'WAY**.  
04/28/2003|10:30|DRISKILL|REPORTS|IN|SB-29146|F/O|4146|**BROADWAY**|

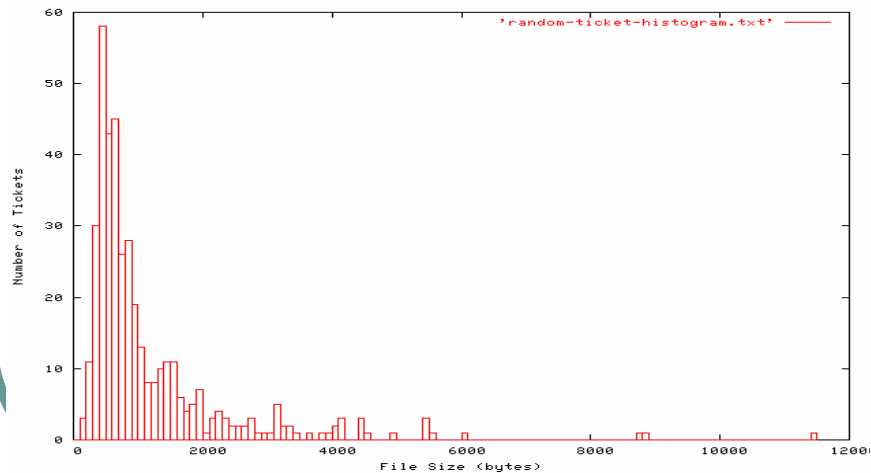
W 29 ST & **B/WAY** - AFTER TROUBLE SECTION WAS C/F/R WE  
WEST 29TH STREET|&|**BROADWAY**|-|AFTER|TROUBLE|SECTION|WAS|C/F/R|WE|

ST -METAL POLE-TAG#:389204- COLUMBUS AV, **BWAY**, SOUTH SIDE O  
ST|METAL|POLE-TAG|#|389204|COLUMBUS AVENUE|**BROADWAY**|SOUTH|SIDE|O|

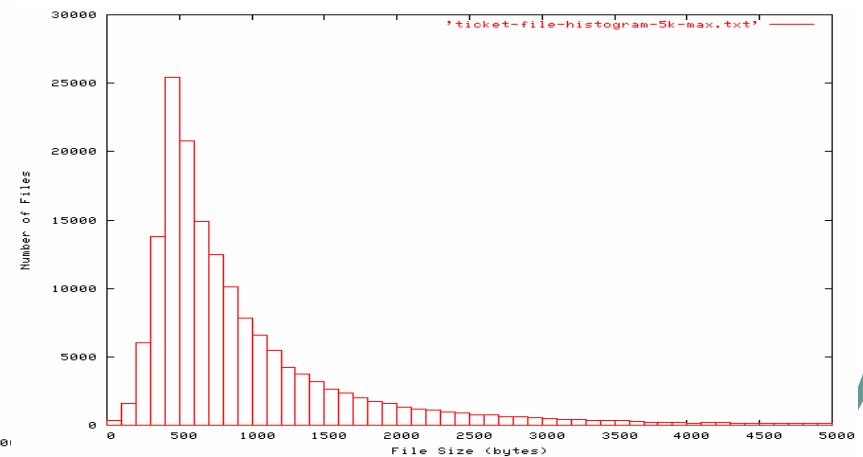
## Stage 2: Annotation of Sample of Tickets

- Used a subset of 400 randomly selected tickets + 6 largest (over 23Kb each) tickets, all created in 2000-2005
- Annotated & analyzed a sample (70 from the 400 subset +3 from the subset of largest 6 tickets)
- Ticket distribution by size:

*Subset (400 tickets)*



*Entire Dataset (162K tickets)*



## Stage 2: Identifying Explicit Ticket Sections

Section Name	Data
Complaint ( <i>Initial Remarks</i> )	Free-text
Office Action	Structured text (produced by filling out formatted screens)
Office Note	
Office Note – Additional Field Information ( <i>Ongoing Remarks</i> )	Free-text
Field Report	
Job Referral	Structured text (produced by filling out formatted screens)
Job Completion	
Job Cancelled	

# Resulting Annotation: Ticket Sections

```
<complaint>
    CONST MGMT REPORTS SPARKING WIRE IN MH N/S SPRING ST
    55' E/O 12TH AVE (ON WALK) - CONTRACTORS ON LOCATION-MC
</complaint>
<office_action> 06/08/00 23:16 MDERWILLIM DISPATCHED BY 48414 </office_action>
<office_note>
    06/08/00 23:17 MDERWILLIM ARRIVED BY 48414
    06/08/00 23:17 CREW PULLED OFF FOR OUTAGE.....CG
    06/08/00 23:18 MDERWILLIM UNFINISHED BY 48414
</office_note>
<office_action> 06/09/00 15:00 MDEDONOHUE DISPATCHED BY 48414 </office_action>
<office_note> 06/09/00 16:00 MDEDONOHUE ARRIVED BY 44729 </office_note>
<field_report>
    06/09/00 18:20 MDEDONOHUE REPORTS CLEARED MULTIPLE B/O'S
    IN SB#46977 N/S SPRING ST 55'E/O 12TH AV READY FOR C.A.I -
</field_report>
<job_completion>
    06/09/00 18:34 MDEDONOHUE COMPLETE BY 44729
</job_completion>
<job_referral>
    06/09/00 18:34 REFERRED TO: CAI EDSWBR FYI BY 44729
</job_referral>
<office_note> 06/10/00 14:10 NO C.M. ACTION REQD.===== BY 44979 </office_note>
```

# Stage 3: Sublanguage Model Development

## Steps Taken:

- Identify core vocabulary, including:
  - Abbreviations & acronyms
  - Special terms & phrases
  - Most common misspellings & typos
- Identify ticket structure (major sections; section components)
- Identify core domain concepts (Trouble Types; Location; Time; Person; ConEd Department, etc)
  - Explicit & Implicit
- Map vocabulary to concepts

## Data used:

- A subset (400 randomly selected + 6 largest tickets)

## Stage 3: Sublanguage Model Development (contd)

- Identifying Core Vocabulary
  - Acronyms
    - Trouble Types - *SMH*
    - Departments – *EDS, S/S/C*
    - Directions, locations - *N/W/C, S/S/C*
    - Other - *PACM, LSE*
  - Abbreviations
    - *BSMNT, BLDG, F/UP, B/O*
  - Special terms & phrases
    - *FEEDER, WHITE HAT, CO FRCES*

## Stage 4: Automatic Ticket Representation: Ticket Sections

- Developed & implemented rules for automatic identification of ticket sections
- Ran automatic section identification on the entire dataset
- Manually evaluated a sample of the system output (73 annotated and 80 unseen tickets)
  - < 1.5 % error rate
- Conclusion: Predictable sublanguage reveals structure

## Stage 4: Automatic Ticket Representation: Semantic Components

- Implemented patterns for some semantic components:
  - Time; Location; Entry\_Person; Feeder; ECS\_structure; Hazard; Urgent
- Compared system output on 70 'gold standard' manually tagged tickets
  - Showed the system accuracy to be 90% or higher
    - Varied based on particular component
  - Proving that automatic identification of semantic components can be done effectively
- Tagging of semantic components helps bring together different expressions of same meaning

## Semantic Components - Examples

- *hazard*:

<INFO type="time" normalized="05/07/2004 @ 08:01">05/07/04  
08:01</INFO> <INFO type="*hazard*"> **UNSAFE LADDER** </INFO>

CUST SAID THAT WIRES /CABLES FROM <INFO  
type="ECS\_structure" normalized="MANHOLE">  
MANHOLE</INFO> TO BLDG BSMT ,,IS RUNNING ALONG THE  
SIDE WALK,,CREATING A <INFO type="*hazard*"> **HAZ** </INFO>  
,,WALKING,,,AND...

<INFO type="time" normalized="07/30/2004 @ 16:04">07/30/04  
16:04</INFO> <INFO type="*hazard*"> **PACM** </INFO> NEAR GAS  
SERVICE ON WATER <INFO type="entry\_person"> BY  
12137</INFO>

# Semantic Components - Location

- *location* - **BROADWAY**:

REPORT AT THE CORNER OF <INFO type="*location*" normalized="**BROADWAY**"> **BROADWAY** </INFO> AND <INFO type="location" normalized="BATTERY PLACE">BATTERY PL</INFO>

SHUNTS RUN FROM THE W.<INFO type="*location*" normalized="**BROADWAY**">**BWAY**</INFO> SIDE OF BUILDING.....

DRISKILL REPORTS IN <INFO type="ECS\_structure" normalized="SB-29146">SB-29146</INFO> F/O 4146 <INFO type="*location*" normalized="**BROADWAY**">**B'WAY**</INFO>.

# Semantic Components - Complaint Section

**<complaint\_source>** CONST MGMT **</complaint\_source>**

REPORTS

**<problem>** SPARKING WIRE IN

**<ECS\_structure>** MH **</ECS\_structure>**

**<location>** N/S SPRING ST 55' E/O 12TH AVE  
(ON WALK)

**</location >**

CONTRACTORS ON LOCATION

**</problem >**

**<entry\_person>** MC **</entry\_person>**

## Stage 5: Applying Knowledge Discovery Approaches

- In order to mine for frequency-based patterns:
  - By field (in the *Data* file)
  - By ticket section (in the *Remarks* file)
  - By Trouble Type (combining *Data* & *Remarks* files)
- Apply Machine Learning to:
  - Assist with Trouble Type assignment
- Mine for associations:
  - Group related tickets for a broader picture of the past and present utility problems in NYC

# Mining Frequency-Based Patterns: *Remarks File*

- Mined for n-grams of n consecutive terms using the Ngram Statistic Package tool
  - Log likelihood algorithm
- Bi-grams and tri-grams were generated for particular ticket sections & Trouble Types
- Analysis of bigrams & trigrams:
  - Shows that Trouble Types differ in their vocabulary in both *complaint* and *field\_report* sections
  - Appears useful for purpose of mining domain specific “phrases”

# Top 10 bi-grams - *Complaint Section*

<b>WL</b>	<b>NL</b>	<b>ACB</b>
<b>WATER LEAKING</b>	<b>FUSES CHECKED</b>	<b>B/O S</b>
<b>WATER LEAK</b>	PART SUPPLIED	<b>DUCT EDGE</b>
ASST ASAP	<b>NO LIGHTS</b>	<b>AC BURNOUT</b>
<b>WATER COMING</b>	<b>LIC #</b>	NO PARKING
REQ ASST	<b>- RMKS</b>	ACCESS ANYTIME
ELEC CONDUIT	<b>ENTIRE BLDG</b>	CONST MGMT
ASAP ETS	CUSTOMER END	WEST WALL
CO ASST	ASST ASAP	EAST WALL
CHECK FIX	800-752-6633 BREAKERS	<b>AC B/O</b>
SWITCH GEAR	SUPPLIED ENTIRE	<b>FLUSH REQUIRED</b>

# Top 10 bi-grams - *Field-Report Section*

<b>WL</b>	<b>NL</b>	<b>ACB</b>
<b>WATER LEAK</b>	<b>PSC MADE</b>	<b>B/O S</b>
<b>SUMP PUMP</b>	<b>B/O S</b>	<b>S CLEARED</b>
<b>SERVICE DUCT</b>	<b># 6</b>	<b>3-500 2-4/0</b>
<b>DYE TEST</b>	<b>BLD #</b>	<b>NO PARKING</b>
<b>WATER LEAKING</b>	<b>3 PHASES</b>	<b>FLUSH ORDERED</b>
<b>NO WATER</b>	<b>1 PHASE</b>	<b>DUCT EDGE</b>
<b>FOUNDATION WALL</b>	<b>CUT CLEARED</b>	<b>3 1/2</b>
<b>WATER COMING</b>	<b>FULL SERVICE</b>	<b>3-200 1-4/0</b>
<b>FOUNDATION LEAK</b>	<b>REMOVED BRIDGE</b>	<b># 6</b>
<b>I-A3 FOUND</b>	<b>3-500 2-4/0</b>	<b>7-4/0 3</b>

# Mining Ticket Section Patterns

- Analysis of section type sequences, i.e. discourse analysis:

ME00000205: complaint office\_action office-note office-note-add-info field-report office-note-add-info field-report job\_completion job\_referral

ME00000206: complaint office\_action office-note job\_completion office-note-add-info field-report job\_referral field-rep

ME00000214: complaint office\_action office-note field-report office-note office\_action office-note office-note-add-info office\_action office-note field-report office\_action office-note field-report job\_referral office-note-add-info job\_completion office-note-add-info job\_referral office-note-add-info

ME00000305: complaint office\_action office-note field-report job\_completion office-note-add-info job\_referral office-not add-info job\_referral

ME00000370: complaint office\_action office-note job\_completion job\_referral field-report

ME00000395: complaint office\_action job\_referral office-note-add-info job\_referral office-note field-report office-note office\_action office-note field-report office-note office-note-add-in office\_action office-note field-report office-note office\_action office-note field-report office-note office\_action off: note field-report job\_completion job\_referral office-note-add-info job\_referral

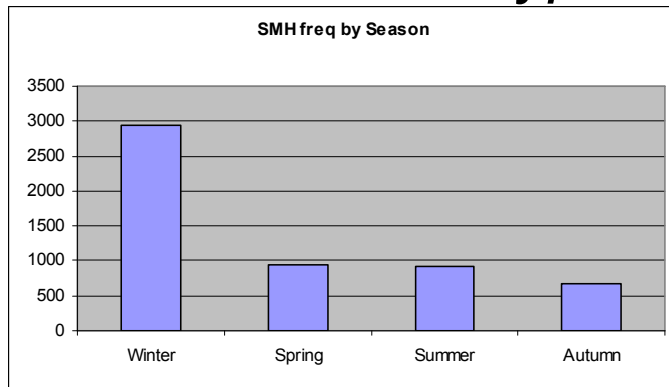
ME00000409: complaint office\_action office-note job\_completion job\_referral

ME00000572: complaint office\_action office-note job\_completion office-note-add-info

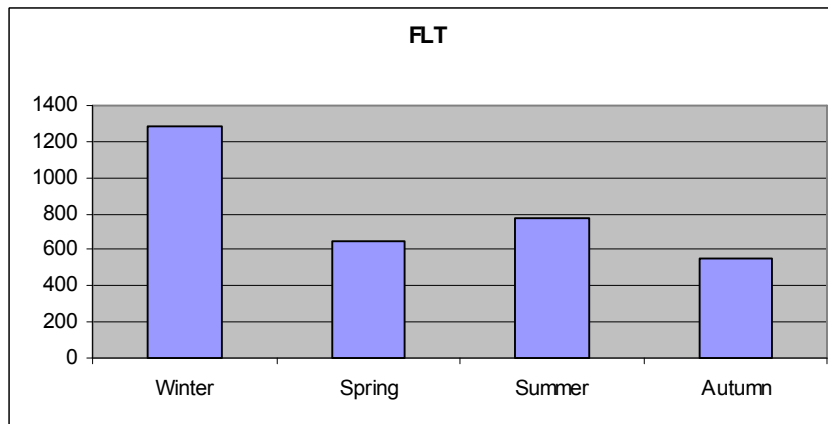
ME00000573: complaint office\_action office-note job\_completion

# Seasonal Patterns of Trouble Types - 1

- Some *Trouble Types* show distinct seasonal patterns:

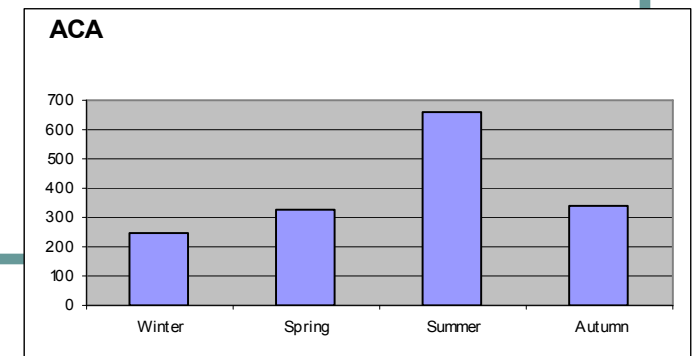


Smoking Manhole



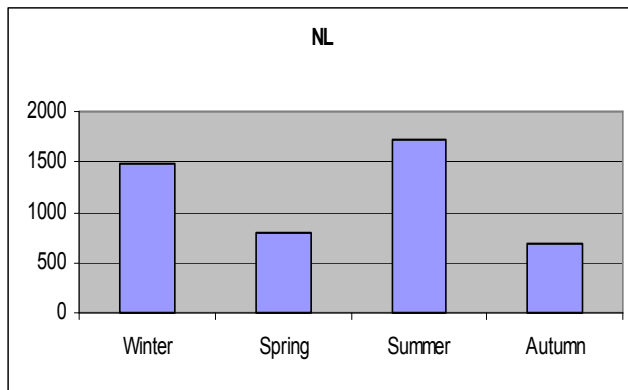
Flickering Lights

Asbestos Clear Access

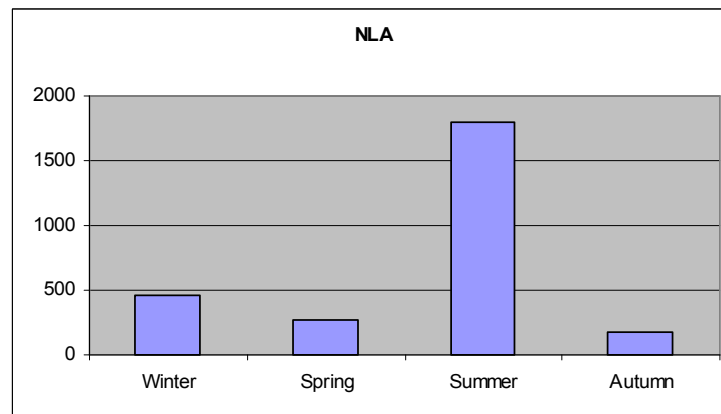


# Seasonal Patterns of Trouble Types - 2

- Noticeable differences in seasonal patterns of similar *Trouble Types*:



No Lights - Individual



No Lights - Area

# Top 10 All-Season Locations

Frequency	Street / Ave		Cross Street / Ave
760	YORKVILLE		HELLGATE SUBSTATION
611	BROADWAY		
202	57	ST	5 AV
188	WATER	ST	COENTIES SLIP
162	1	AV	E 40 ST
158	125	ST	5 AV
155	47	ST	5 AV
151	BROADWAY		SPRING ST
150	57	ST	AMERICAS AV
148	54	ST	5 AV

# Machine Learning for Assigning *Trouble Type*

- MSE is the most frequent *Trouble Type* – 18%, which means:
  - Almost 1/5 of all tickets cannot be effectively mined for associations between a *Trouble Type* and other ticket components
- We demonstrated that in many cases, more specific *Trouble Types* could be assigned:

```
<TICKET id="ME05003448" original-code="EDSMSE" actual-code="EDSWL">  
<SECTION type="complaint">  
    WATER LEAKING INTO TRANSFORMER  
    BOX IN  
    BASEMENT OF DORM; PLS CHECK FOR  
    SAFETY
```

# Patterns in Trouble Code Changes

## *Top 10 Original Trouble Codes*

29678	18.3%	EDSMSE
10996	6.8%	EDSHCE
9909	6.1%	EDSSMH
7898	4.9%	EDSWL
5579	3.4%	EDSNL
5242	3.2%	EDSHME
5216	3.2%	EDSUDC
5066	3.1%	EDSACB
5001	3.1%	EDSOA
4912	3.0%	EDSSO

# Patterns in Trouble Code Changes

## Top 10 *Original* Trouble Codes

29678	18.3%	<b>EDSMSE</b>
10996	6.8%	EDSHCE
9909	6.1%	EDSSMH
7898	4.9%	EDSWL
5579	3.4%	EDSNL
5242	3.2%	EDSHME
5216	3.2%	EDSUDC
5066	3.1%	EDSACB
5001	3.1%	EDSOA
4912	3.0%	EDSSO

# Patterns in Trouble Code Changes

*Top 10 Original Trouble Codes*

29678	18.3%	<b>EDSMSE</b>
10996	6.8%	EDSHCE
9909	6.1%	EDSSMH
7898	4.9%	EDSWL
5579	3.4%	EDSNL
5242	3.2%	EDSHME
5216	3.2%	EDSUDC
5066	3.1%	EDSACB
5001	3.1%	EDSOA
4912	3.0%	EDSSO

*Top 10 Actual Trouble Codes*

29032	17.9%	<b>EDSMSE</b>
9111	5.6%	EDST9X
7928	4.9%	EDSWL
4780	2.9%	EDSOA
4691	2.9%	EDSACB
4378	2.7%	EDSNL
4288	2.6%	EDSUDC
4203	2.6%	EDSSMH
4014	2.5%	EDSOPN
3589	2.2%	EDSUAC

# Learning Patterns for Trouble Type Assignment

- Using Machine Learning:
  - Trained a system on ‘problem descriptions’ for known classes of *Trouble Types*
  - Then, for a new ticket, the system either assigns the top ranked *Trouble Type*, or suggests list of possible *Trouble Types*, based on its NLP-based learning of ‘problem descriptions’ for operator to select from
  - Could also be done “off-line” for pre-existing tickets to clear up the number of MSE tickets and increase the number of tickets with real *Trouble Types* for data mining
  - Highly promising results on 6,500 unseen *Trouble Tickets*

# Machine Learning Experiments

- Operational Scenario
  - An NLP-enabled system, based on information in the *Initial Remarks* ('*complaint*') section, suggests to a Call Center Operator a list of potentially relevant Trouble Types
- Tool: *Extended LibSVM*
- Experimental design:
  - Multi-label classification task
  - System was trained on problem descriptions from '*complaint*' section of specific Trouble Type tickets
  - Experiments
    1. Classifier is tested on tickets with known Trouble Types
    2. Classifier is tested on miscellaneous tickets

# Machine Learning: Experiment 1

- Train & test classifier on known Trouble Types
  - Using “gold standard” available from the client
  - 5 most frequent Original Trouble Types selected
  - ‘Complaint’ section used because it contains only the information available to the Call Center Operator at the time the ticket is created and assigned a Trouble Type

- *Dataset:*

<b>TT</b>	<b>Training</b>	<b>Test</b>
<b>EDSSMH</b>	7432	2477
<b>EDSWL</b>	5924	1974
<b>EDSNL</b>	4184	1395
<b>EDSOA</b>	3751	1250
<b>EDSACB</b>	3800	1266

# Machine Learning: Experiment 1 Results

Trouble Type	Precision P	Precision N	Recall P	Recall N
SMH	91.8	98.2	90.8	98.4
WL	98.4	99.1	97.9	99.3
NL	92.8	98.7	93.8	98.5
OA	99.7	99.7	98.7	99.9
ACB	93.2	97.5	88.6	98.6

- P – Positive Class (“target” Type)
- N – Negative Class (the rest)

## Machine Learning: Experiment 2

- Trained classifier on 5 known Trouble Types from Experiment 1
- Ran MSE tickets through the classifier
- Manually evaluated the results
- Challenges:
  - Involves manual evaluation (time + effort)
  - If evaluation is done by developers, need to be validated by SMEs

# Machine Learning: Experiment 2 Dataset

	<b>Training</b>	<b>Test</b>
<b>EDSHCE</b>	8247	0
<b>EDSSMH</b>	7432	0
<b>EDSWL</b>	5924	0
<b>EDSNL</b>	4184	0
<b>EDSHME</b>	3932	0
<b>EDSUDC</b>	3912	0
<b>EDSACB</b>	3800	0
<b>EDSOA</b>	3751	0
<b>EDSSO</b>	3684	0
<b>EDSOPN</b>	3036	0
<b>EDSFLT</b>	2621	0
<b>EDSUAC</b>	2612	0
<b>EDSSOP</b>	2545	0
<b>EDSSLT</b>	2409	0
<b>EDSOOE</b>	2300	0
<b>EDSNLA</b>	2291	0
<b>EDSLV</b>	2268	0
<b>EDSTRF</b>	2050	0
<b>EDSSPD</b>	2014	0
<b>EDSWBR</b>	1842	0
<b>EDSMSE</b>		7420

# Machine Learning: Experiment 2 Results

- Of 7420 MSE tickets in the Test set, system classified:
  - 181 as SMH
  - 330 as WL
- For each Type, 50 tickets were manually evaluated
- Of 50 MSE tickets classified into SMH:
  - 25 were judged by CNLP analyst as “correct”
  - 14 of these had their original MSE Type later changed to SMH
  - Of remaining 11 tickets validated with the SME, only 1 was not confirmed as SMH
- Of 50 MSE tickets classified into WL:
  - 35 were judged by CNLP analyst as “correct”
  - 23 of these had their original MSE Type later changed to WL
  - Of remaining 12 tickets validated with the SME, only one was not confirmed as WL
- Results are promising, but not exhaustive nor conclusive

# Findings & Conclusions

- Application shows utility of sublanguage approach
- Language of Con Edison's Trouble Tickets demonstrates typical sublanguage characteristics
  - Tickets' linguistic patterns are consistent & can be utilized to support semi- or fully-automated identification of important ticket components (events, organizations, equipment, urgency)
    - Assisting analysts
    - Bringing together lexical & syntactic variants streamlines and expands coverage of subsequent data analysis.
- Utilizing identified components & knowledge of sub language patterns, can analyze data more effectively:
  - Using annotated data as input to sophisticated statistical analyses packages
  - Applying developed prototype system to mine for relevant specific Trouble Types for an MSE ticket

## Findings & Conclusions *(cont-d)*

- N-gram co-occurrence approach showed potential for generating a domain-specific lexicon
- Temporal & seasonal data analysis may lead to insights into factors affecting Con Edison's proactive plans
- High level performance in automatic assignment of specific Trouble Codes
- Combining information from *Data & Remarks* (with newly identified components) supports discovery of interesting associations, e.g.
  - Between Trouble Type and severity (urgency) of the case, actions undertaken, possible impact, etc

# Possible Indicators of 'Severity' of Case

- Trouble Type
- # of *field-persons* involved;
- Whether an additional crew is requested
- Whether external agencies and companies are involved
- Length of text:
  - Entire ticket
  - *Field-report* sections
- Duration (time span) of a case
- Whether the ticket has been re-opened
- Max # of clients interrupted
- # of calls received for the ticket
- Presence of certain clues (e.g. *Urgency* or *Hazard* section components)

# Outcomes

- Automatic identification of trouble ticket sections & section components
- Detection of seasonal patterns of *Trouble Types*
- Contrary to customer's belief, MSC Trouble Codes do not get changed to more specific *Trouble Type* during human review
  - 18.3% of original trouble codes were MSC
  - 17.9% of final codes were MSC
- High level performance in automatic assignment of specific Trouble Codes
- Successful application of Sublanguage Methodology

# Acknowledgements

- CNLPers
  - Svetlana Symonenko
  - Steve Rowe
- Con Edison's SMEs