Automatic Recognition of Documentary Forms

William Underwood
Sandra Laib

Technical Report ITTL/CSITD 08-02

May 2008

Computer Science and Information Technology Division
Information Technology and Telecommunications Laboratory
Georgia Tech Research Institute
Georgia Institute of Technology

The Army Research Laboratory (ARL) and the National Archives and Records Administration (NARA) sponsor this research under Army Research Office Cooperative Agreement W911NF-06-2-0050. The findings in this paper should not be construed as an official ARL or NARA position unless so indicated by other authorized documentation.
ABSTRACT

The names of documentary forms, such as correspondence, memoranda and agenda, are elements of the archival descriptions of record series, file units and individual records. Metadata; such as date, author, addressee and topic, are elements of documentary forms. This metadata can be used as access points to search for and retrieve records from archives.

The concept of documentary form is reviewed and an approach to defining documentary forms as context-free grammars is described. Grammars are defined for twenty-two documentary forms that occur in Presidential e-records. A method for automatically recognizing the intellectual elements of documentary form is described. It is explained how attribute-value context-free grammars for documentary forms can be used with a parser to recognize and interpret the documentary forms of e-records. The semantic representation of the meaning of a documentary form can be used for metadata extraction and archival description. The results of a test are described in which manually constructed grammars for documentary forms are used to recognize document types of Presidential e-records.

Keywords: archival description, documentary form, metadata extraction, computational linguistics,
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Information Memoranda</td>
<td>33</td>
</tr>
<tr>
<td>Recommended Telephone Call</td>
<td>35</td>
</tr>
<tr>
<td>Memorandum of Conversation (MEMCON)</td>
<td>36</td>
</tr>
<tr>
<td>Memorandum of Telephone Conversation (TELCOM)</td>
<td>38</td>
</tr>
<tr>
<td>White House Referral Slip</td>
<td>39</td>
</tr>
<tr>
<td>Executive Orders</td>
<td>41</td>
</tr>
<tr>
<td>National Security Directives</td>
<td>43</td>
</tr>
<tr>
<td>National Security Review</td>
<td>44</td>
</tr>
<tr>
<td>Presidential Determination</td>
<td>45</td>
</tr>
<tr>
<td>Presidential Proclamation</td>
<td>46</td>
</tr>
<tr>
<td>White House Fact Sheet</td>
<td>48</td>
</tr>
<tr>
<td>Transcript of Presidential News Conference</td>
<td>49</td>
</tr>
<tr>
<td>White House Press Releases</td>
<td>52</td>
</tr>
<tr>
<td>Press Pool Report</td>
<td>52</td>
</tr>
<tr>
<td>Resumes</td>
<td>54</td>
</tr>
<tr>
<td>Schedules</td>
<td>56</td>
</tr>
<tr>
<td>Mailing Lists</td>
<td>57</td>
</tr>
</tbody>
</table>
**TABLE OF FIGURES**

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>An Example of a White House Memorandum</td>
<td>5</td>
</tr>
<tr>
<td>2</td>
<td>A Parse Tree for a Sample Memorandum</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>The Semantics of the Sample Memorandum</td>
<td>21</td>
</tr>
<tr>
<td>4</td>
<td>Quasi-Logical Form for the Semantics of the Documents in the Test Case</td>
<td>23</td>
</tr>
<tr>
<td>5</td>
<td>A Formal Letter</td>
<td>28</td>
</tr>
<tr>
<td>6</td>
<td>An Example of a White House Letter</td>
<td>30</td>
</tr>
<tr>
<td>7</td>
<td>An Example of a Decision Memorandum</td>
<td>32</td>
</tr>
<tr>
<td>8</td>
<td>Example of Concluding Part of a Decision Memorandum</td>
<td>33</td>
</tr>
<tr>
<td>9</td>
<td>An Example of an Information Memorandum</td>
<td>34</td>
</tr>
<tr>
<td>10</td>
<td>An Example of a Recommended Telephone Call</td>
<td>35</td>
</tr>
<tr>
<td>11</td>
<td>A Memorandum of Conversation</td>
<td>37</td>
</tr>
<tr>
<td>12</td>
<td>A Memorandum of Telephone Conversation</td>
<td>39</td>
</tr>
<tr>
<td>13</td>
<td>A White House Referral Slip</td>
<td>40</td>
</tr>
<tr>
<td>14</td>
<td>An Example of an Executive Order</td>
<td>42</td>
</tr>
<tr>
<td>15</td>
<td>An Example of a National Security Directive</td>
<td>43</td>
</tr>
<tr>
<td>16</td>
<td>A National Security Review</td>
<td>45</td>
</tr>
<tr>
<td>17</td>
<td>A Presidential Determination</td>
<td>46</td>
</tr>
<tr>
<td>18</td>
<td>Example of a Presidential Proclamation</td>
<td>47</td>
</tr>
<tr>
<td>19</td>
<td>A White House Fact Sheet</td>
<td>49</td>
</tr>
<tr>
<td>20</td>
<td>A Transcript of a Presidential News Conference</td>
<td>51</td>
</tr>
<tr>
<td>21</td>
<td>An Example of a White House Press Release</td>
<td>52</td>
</tr>
<tr>
<td>22</td>
<td>A Press Pool Report</td>
<td>52</td>
</tr>
<tr>
<td>23</td>
<td>An Example of a Resume</td>
<td>55</td>
</tr>
<tr>
<td>24</td>
<td>An Example of a Schedule</td>
<td>56</td>
</tr>
<tr>
<td>25</td>
<td>An Excerpt from a Mailing List</td>
<td>57</td>
</tr>
</tbody>
</table>
1. Introduction

1.1 Background

The names of documentary forms, such as correspondence, memoranda or agenda, are key elements of archival descriptions. National standards [Describing Archives 2004] and institutional guidelines [NARA 2002] for describing archives prescribe that names of documentary forms be included in scope and content notes of record series and file units and in titles of items.

Archival description includes the identification of access points, “a name, term, phrase, or code that is used to search, identify, or locate a record, file, or document” [NARA 2002]. NARA’s Lifecycle Data Requirements Guide specifies guidelines for creating access points for records, for example, date, author, addressee, topic, and special record type (documentary form). Due to the increasing volume of records acquired by NARA and the limited human resources that can be applied to archival description, the problem arises that it may be decades or even centuries before acquisitions are manually reviewed and described in sufficient detail to be reliably retrieved with reasonable precision and recall. Since an increasing number of accessioned records are digital records, a capability to automatically recognize documentary form and to extract metadata from these records would enable the automatic indexing of records for earlier access.

In previous research, context-free grammars describing the documentary form of two document types—White House Memos and White House Correspondence—were induced from samples of these documentary forms [Underwood and Harris 2006]. It was also demonstrated how these grammars could be used with a parser to recognize the documentary forms of other records of these types.

1.2 Purpose

The purpose of this report is to describe progress in automatically

- Identifying the intellectual and physical elements of documentary forms,
- Recognizing documentary forms using grammars describing the structure of documents in terms of their intellectual elements, and
- Extracting metadata from documents whose form has been recognized.

1.3 Scope

In section 2, the concept of documentary form is reviewed and our approach to defining documentary forms as context-free grammars is described. In section 3, a method for automatically recognizing the intellectual elements of documentary form is described in detail. In section 4, it is explained how attribute-value context-free grammars for documentary forms can be used with a parser to recognize and interpret the documentary
forms of records. In section 5, the results of a test are described in which manually constructed grammars for documentary forms are used to automatically recognize document types of Presidential e-records. In section 6, related research in defining and recognizing document structures is reviewed. In section 7, the results are summarized.

2. Approach

2.1 The Concept of Documentary Form

Documentary form is a fundamental concept of the science of Diplomatics. Diplomatics originated in the need to authenticate medieval documents, but has been extended to address the nature and authenticity of modern documents including electronic records [Duranti 1998].

*Documentary form* is "the rules of representation according to which the content of a document, its administrative and documentary context, and its authority are communicated [InterPARES 2001b]. Documentary form consists of physical form and intellectual form. The *physical form* of a document is the overall appearance, configuration, or shape, derived from its material characteristics and independent of its intellectual content. The *intellectual form* of a document is "the sum of a record's formal attributes that represent and communicate the elements of the action in which the record is involved and of its immediate context, both documentary and administrative."

The elements of the former are defined as external or *extrinsic*, while the elements of the latter are defined as internal or *intrinsic* [Duranti 1998, p 134]. For electronic records, extrinsic elements include:

- overall presentation features (e.g., textual, graphic, image, sound, or some combination of these);
- specific presentation features (e.g., special layouts, hyperlinks, colors, sample rate or sound files);
- electronic signatures and electronic seals (e.g., digital signatures);
- digital time stamps;
- other special signs (e.g., digital watermarks, an organization's crest or personal logo).

Intrinsic elements are the discursive parts of the record that communicate the action in which the records participates and the immediate context. They fall into three groups:

- elements that convey aspects of the record's juridical and administrative context (e.g., the name of the author, addressee, the date);
- elements that communicate the action itself (e.g., the indication and description of the action or matter);
• elements that convey aspects of the record's documentary context and its means of validation (e.g., the name of the writer, the attestation, the corroboration). [InterPARES 2001a, p. 5]

Documentary forms for Federal and Presidential records are described in manuals such as the *US Government Correspondence Manual* and the *White House Correspondence Manual*, where they are customarily referred to as styles. Descriptions of elements of form and layout are usually by example.

Word processing applications such as Microsoft Word and WordPerfect utilize document templates or styles to facilitate creation of documents in prescribed forms. Historically, these styles have been represented with proprietary markup tags.

### 2.2 Grammars for Documentary Forms

Standard markup languages such as SGML and XML use tags to delimit elements of text, document type definitions (DTDs) to define the structure of these elements, and attributes of the elements to characterize the presentation of document elements. An XML DTD [W3C 2006] is a formal analog of the concept of the intellectual form of a document. An XSL stylesheet [W3C 2001] is formally analogous to the concept of the physical form of a document.

An XML DTD is an extended context-free grammar. This makes it plausible that the documentary form of non-XML documents could be defined with context-free grammars.

Formally, a *context-free grammar* is a 4-tuple \( G = \langle N, T, R, S \rangle \) where

- \( N \) is a set of *non-terminal symbols*
- \( T \) is a set of *terminal symbols*. \( N \) and \( T \) are disjoint sets.
- \( R \) is a set of *rules* of the form \( A \rightarrow w \) where \( A \) is a member of \( N \) and \( w \) is a string belonging to \( N \cup T \), and
- \( S \) is a member of \( N \) called the *initial symbol*

In a grammar for a documentary form, the terminal symbols correspond to the names of intellectual elements. The non-terminal symbols correspond to the names of combinations of intellectual elements. The initial symbol is the name of a documentary form. The rules of the grammar define the structure of the documentary form.

To define a documentary form of a particular record type, one must first identify the intellectual and physical elements of the documentary form. In some cases, one can accomplish this by reference to correspondence or style manuals. These manuals are also useful for identifying the terms that are customarily used to name the intellectual elements. In cases in which a documentary form is widely used but not specified in a manual, one can collect a sample of records of a particular documentary form and compare them to identify those elements that do not change from one example to another.
In the following section, a context-free grammar for the documentary form of White House memoranda is shown. Appendix A includes grammars for twenty-one other documentary forms that occur in the electronic records of the Presidential administration of George H.W. Bush.

### 2.3 A Context-Free Grammar for White House Memoranda

A memorandum is "a written communication much like a letter but having no salutation or complimentary ending. Usually used within or between offices of the same organization" [NARA 1993].

The documentary form of White House Memoranda is very similar to the memorandum style specified in the US Government Correspondence Manual. One difference is that the Address Line in a White House memo uses the caption MEMORANDUM FOR while the Government Manual specifies the use of TO. Another difference is that the Thru Line in a White House Memo uses the caption “THROUGH” while the Government Style manual specifies the use of “THRU”. Since both styles appear in memoranda created or received by the White House Staff, the grammar defining the documentary form of memoranda will incorporate both elements.

Figure 1 shows an example of a White House memorandum.¹ Memoranda such as this were printed on White House stationery with a letterhead “THE WHITE HOUSE” and “WASHINGTON” centered at the top of the page. The electronic copies of memoranda typically did not contain the letterhead, but in some cases it was typed at the top of the memorandum.

---

¹ Bush Presidential Library, Bush Presidential Records, WHORM Subject File, Disasters-Natural, ID#324869.
April 27, 1992

MEMORANDUM FOR SAM SKINNER

FROM: EDE HOLIDAY

SUBJECT: California Earthquake

Attached is a situation report from FEMA on the northern California earthquake. No deaths have been reported and 45 people are known to have suffered injuries. In addition, there has been extensive property damage. While FEMA is awaiting a request from the State before initiating any recovery activities, a joint State/Federal preliminary damage assessment is likely to begin today.

Director Stickney has requested that we forward the situation report to you.

Attachments

Figure 1. An Example of a White House Memorandum

The following context-free grammar rules define the documentary form of the memo shown in Fig. 1 as well as the forms of other memoranda in the Bush Presidential records.

MEMO \rightarrow MEMOHEAD BODY
MEMO \rightarrow MEMOHEAD BODY OPTIONAL
MEMOHEAD \rightarrow DATE ADDRLINE SNDRLINE SUBJLINE
MEMOHEAD \rightarrow DATE ADDRLINE THRU LINE SNDRLINE SUBJLINE
ADDRLINE \rightarrow FOR ENTITIES
SNDRLINE \rightarrow FROM ENTITIES
SUBJLINE \rightarrow SUBJ TOPIC
THRU LINE \rightarrow THRU ENTITY
BODY \rightarrow PARAS
OPTIONAL \rightarrow ATTACHMENT CCLIST BCCLIST
OPTIONAL \rightarrow ATTACHMENT BCCLIST
OPTIONAL \rightarrow ATTACHMENT CCLIST
OPTIONAL \rightarrow ATTACHMENT
OPTIONAL \rightarrow CCLIST BCCLIST
OPTIONAL \rightarrow BCCLIST
OPTIONAL \rightarrow CCLIST
CCLIST \rightarrow CC ENTITIES
BCCLIST \rightarrow BCC ENTITIES
PARAS \rightarrow PARA PARAS
PARAS \rightarrow PARA
ENTITIES \rightarrow ENTITIES ENTITY
ENTITIES \rightarrow ENTITY
The intellectual element FOR is “MEMORANDUM FOR” or “TO:”. The element THRU is “THRU:” or “THROUGH”. The element FROM is “FROM:”. The element SUBJ is “SUBJECT:”. The element TOPIC is the clause after SUBJ. The element ATTACHMENT is “Attachment”, “Attachments”, or Attachments (DIGIT)”.

The Grammar G₁ for White House Memoranda is <N, T, R, S> where

\[
N = \{\text{MEMO, MEMOHEAD, BODY, OPTIONAL, ADDRLINE, SNDRLINE, THRU, SUBJLINE, ENTITIES, PARAS, CCLIST, BCCLIST}\}
\]

\[
T = \{\text{DATE, FOR, FROM, SUBJ, THRU, TOPIC, ATTACHMENT, ENTITY, PARA, CC, BCC}\}, \text{the names of the intellectual elements in a memo.}
\]

R is the set of rules shown above, and

S is the symbol MEMO

### 3. Automated Method for Recognizing Intellectual Elements

Our approach to recognizing intellectual elements is based on an extension of research conducted in annotating the named entities in Presidential e-records [Underwood and Isbell 2008]. That research uses the following resources from the General Architecture for Text Engineering (GATE) architecture [Cunningham et al 2007]. The italicized terms indicate resources that we have modified or added to recognize the documentary forms of e-records.

*File Conversion to Plain Text (or HTML)*
Document Reader
English Tokenizer
Wordlist Lookup + enhanced wordlists
Sentence Splitter
Hepple POS Tagger + lexicon
Semantic Tagger + Named Entity Rules
Annotation Set Transfer
Document Element Tagger + Document Element Rules (DER)
SUPPLE Parser + mapping file + feature table + Document Type Grammars

The input to this process is an electronic record in a variety of legacy or current file formats. The output of the process is (1) the intellectual and physical elements of the documentary form of the record, (2) a parse tree indicating the form of the record, and (3)
metadata indicating the document type, the chronological date, the author(s), addressee(s) and topic of the record. In the following sections, some of these resources are described.

3.1 Converting File Formats to Plain Text or HTML

GATE has the built-in capability to process the following document formats:

- Plain Text
- HTML
- SGML
- XML
- RTF
- Email
- PDF (some documents)
- Microsoft Word (some documents)

An issue that immediately arises in considering the physical elements of form is that these are most often represented in a variety of proprietary file formats and style sheets. We need to get the physical elements of form into a common representation.

Because Presidential records may be created in a variety of proprietary or standard file formats, a file type identifier of our creation is used to identify the file format of a record. An appropriate file format converter is used to convert documents that are not in plain text or HTML formats to plain text or HTML.

Both types of conversion preserve the intellectual elements of form. Physical elements of form are preserved if the files are converted to HTML, and lost if the files are converted to plain text. Physical elements (e.g., fonts, font size, underlining, uppercase, boldface, italics, capitalization, bullets, indented paragraphs, centering, flush right, flush left, justified, columns, borders, shading and pagination) need to be considered to effectively recognize intellectual elements such as titles, section headings, and pages.

A corpus was constructed of 45 documents, spreadsheets, and database tables in 32 different file formats. Many of the files are copies of Bush Presidential records already opened to the public. Most of the file formats are those that were current during the administration of President George H. W. Bush.

Two HTML File format converters were evaluated with regard to the HTML markup that they create from different proprietary file formats. One was from Microsoft and the other from Oracle. The Microsoft converters produced much more information in the HTML formats than was needed. It particular, it included information that facilitated conversion back to Microsoft formats.

HTML Export [Oracle 2007] recognizes and converts about 250 file formats to HTML. It supports conversion to four flavors of HTML (Generic, HTML 2.0, HTML 3.0 and
HTML 4.0). Table export is supported in the Generic Flavor and HTML 3.0 and 4.0, and there are parameters for controlling the number of Table columns displayed per page. It also supports conversion to HTML tags for bulleted and numbered lists.

HTML export converts many relational database and spreadsheet file formats to HTML tables. There are only a few textual, database and spreadsheet file formats encountered in the Bush personal computer records for which there are not HTML converters. These include Advanced Revelation Databases, PlanPerfect 5.1, and WordPerfect 4.2. These are handled by converting to a file format that HTML Export can convert to HTML.

HTML Export was chosen for the conversion task because of the large number of legacy, as well as current, file formats to which it applies, and because of its rich set of options that allow a high degree of customization of the HTML output.

When a document is read by GATE, it identifies the file format, and applies the appropriate document reader. The document reader strips and converts any markup into GATE’s annotation format. When reading a document in HTML format, GATE extracts the text between tags and creates a GATE annotation. The name of the tag constitutes the annotation’s type. All the tags’ attributes are converted to annotation features and the annotation spans the text covered by the tag.

### 3.2 Wordlist Lookup

Each wordlist represents a set of names, such as male and female first names, surnames, cites, states, countries, organizations, days of the week, etc. For each list, a major type is specified and, optionally, a minor type. For instance, Atlanta has major type location and minor type city. These lists are compiled into finite-state machines. Any text tokens that are matched by these machines are annotated with features specifying the major and minor types. The GATE distribution includes a vanilla set of wordlists, primarily for demonstration of functionality. The wordlist coverage has been improved by creation of additional wordlists and wordlists that distinguish unambiguous person’s first names and surnames from those that are ambiguous, and unambiguous location names from location names that are ambiguous [Underwood and Isbell 2008]

### 3.3 Semantic Tagger

The semantic tagger consists of rules that act on annotations assigned in earlier phases to produce annotations of compound names such as person’s full names, dates, and times (e.g., 3 PM). The GATE distribution also includes a vanilla set of rules for annotating named entities. The performance of the Semantic Tagger has been improved by modifications and additions to the rules. Semantic annotation capabilities were also extended to recognize job titles, postal addresses, bills and statutes, telephone numbers, and social security numbers. An experiment was conducted in which the improved wordlists and Semantic Tagger were applied to a sample of 50 Bush personal computer records [Underwood and Isbell 2008]. The performance was:
Overall average precision 0.9178
Overall average recall 0.9286
Overall F-measure 0.9108

This performance is very good. Without this level of performance, the method recognizing intellectual elements, which are dependent on the Semantic Tagger, could not achieve an adequate level of performance.

3.4 Annotation Set Transfer

Annotation Set Transfer is a GATE plug-in that copies annotations from one Annotation Set to another. The Annotation Set Transfer is used to transfer the paragraph annotations of plain text documents (and the p annotations in the case of HTML) from the Original Markups Annotation Set to the Default Annotation Set.

3.5 The Document Element Tagger

The Document Element Tagger is a multi-phase tagger consisting of JAVA Annotation Pattern Engine (JAPE) rules. The rules identify some of the annotations created by the semantic tagger as intellectual elements of documentary forms and annotate them as elements. It also identifies and annotates additional intellectual elements.

3.5.1 PossEleStep1 Phase

The first phase of the Document Element Tagger matches strings appearing in a document to possible intellectual elements, creates a PossElement annotation for those it finds, and assigns an intellectual element name to the eletype feature of the annotation. The PossElement annotation is an intermediate annotation. Not all PossElement annotations become Element annotations.

The following is an example of a JAPE rule that tries to match the “Executive” followed by “Order” followed by a number against tokens in a document If it succeeds, it adds the annotation PossElement annotation with eletype=eoid.

```
// Executive Order [number] -> eoid
Rule: EOID
({Token.string == "Executive"}
 {Token.string == "Order"}
 {Token.kind == number}
):lookup
-->
   :lookup.PossElement = {eletype = "eoid"}
```
The following list shows some of the possible intellectual elements and their associated eletype values. In this phase, a single JAPE rule matches all strings that produce the same eletype value. Some of the rules below are grouped by the eletype values shown on the right-hand side of the rules.

Action → action
ACTION → action

Attachment → attachment
Attachments → attachment
Attachments: → attachment
Attachments({number}) → attachment

bcc: → bcc

c c → cc
cc: → cc
Copies to → cc
Copy to → cc

(C) → classabbr
(S) → classabbr
(TS) → classabbr
(U) → classabbr

CONFIDENTIAL → classification
SECRET → classification
TOP SECRET → classification
TOP SECRET → classification
UNCLASSIFIED → classification

, → conjunct
, and → conjunct
and → conjunct

DATE OF SUBMISSION: → date_of_submission
TOPICS OF DISCUSSION: → disctopics
DATE TO CALL: → dtc
-- → emdash

Enc. → enclosure
Enclosure → enclosure
Enclosures → enclosure
Enclosures ({number}) → enclosure

# # # → ends
Executive Order {number} → eoid
For Immediate Release → fir
Dear → formula_of_address
Devotedly,  →  formula_of_respect
Respectfully  →  formula_of_respect
Sincerely,  →  formula_of_respect
Sincerely yours,  →  formula_of_respect
Warmly,  →  formula_of_respect
With best regards,  →  formula_of_respect
With best regards, Sincerely,  →  formula_of_respect
With kind regards,  →  formula_of_respect

FROM:  →  from
Home:  →  home
MINUTES OF THE CABINET MEETING  →  minuteshdg
NATIONAL SECURITY DIRECTIVE  →  nsd
NATIONAL SECURITY REVIEW-  →  nsr
Office:  →  office
Office of the Press Secretary  →  office_press_secy

PS  →  postscript
P.S.  →  postscript

PURPOSE:  →  purpose
RECOMMENDED BY:  →  RCmdby
Recommended Telephone Call  →  rcmdtelecall
SUBJECT:  →  subj
TELEPHONE:  →  telephone
THE WHITE HOUSE  →  THE_WHITE_HOUSE

THROUGH:  →  thru
THRU:  →  thru

MEMORANDUM FOR  →  to
TO:  →  to

The White House  →  twh
_______  →  uline (this rule identifies a variable length underscore).

3.5.2 PossEleStep2 Phase

The PossEleStep2 phase matches some of the annotations previously identified by the Semantic Tagger. For each match it creates a PossElement annotation and gives it an eletype feature value.

If a Date annotation is found, create a PossElement annotation with eletype=docdate.

If a Phone annotation is found, create a PossElement annotation with eletype=phoneno.
If a Person annotation is found, create a PossElement annotation with eletype=person.

If a JobTitle annotation is found, create a PossElement annotation with eletype=jobtitle.

If an Address annotation is found, create a PossElement annotation with eletype=address.

3.5.3 PossEleStep3 Phase

The PossEleStep3 phase matches strings following previously identified PossElement annotations. For each match, it creates a PossElement annotation and assigns a value to the annotations eletype feature..

   If there is a PossElement annotation with the eletype formula_of_respect, and it is followed by person [jobtitle] annotations; and there are two or three lower case letters, or two or three uppercase letters followed by a colon or forward slash followed by two or three lowercase letters, or two or three uppercase letters followed by a colon or forward slash, followed by two three uppercase letters, followed by a slash or colon, followed by two or three lowercase letters,
then create a PossElement annotation with eletype=refinitials.

   If there is a PossElement annotation with eletype formula_of_address, and It is followed by a person and a colon,
Then create a PossElement annotation with eletype=salutation.

3.5.4 Para Phase

The GATE Document Reader uses a weak method for segmentation of plain text documents. It annotates any sequence of characters that are followed by a carriage return, line feed as a paragraph. Hence title, page numbers, subject lines in memoranda, etc. are all annotated as paragraphs. The Para phase adjusts for this weakness by re-annotating paragraphs to be paragraphs (or p annotations in the case of HTML) just in case they are a sequence of one or more sentences.

The Para phase differentiates between what was originally marked with a paragraph or p annotation and what constitutes an intellectual element annotation with eletype=para. The Element annotation must end in a period. This is also were the Element annotation with eletype=chrondate is identified. The Element annotation with eletype=chrondate is different from the Element annotation with eletype=docdate. The Element annotation with eletype=chrondate must be the only annotation in the paragraph. The following rules
show the intellectual Element annotations and their associated eletypes that have been implemented. The rule identifying chrondate is run first. This prevents the date followed by a period from being given the eletype=para.

If there is a PossElement annotation with eletype=docdate, and
   it is the only annotation on its line,
Then create an Element annotation with eletype=chrondate, and
   remove the PossElement annotation with eletype=docdate.

If there is a paragraph annotation, and
   it has a PossElement annotation with eletype=classabbr at the end,
then create an Element annotation with eletype=para_class that spans
   from the beginning of the paragraph through the last token annotation.

If there is a paragraph annotation, and
   the paragraph begins with "By the authority vested in me as President",
then create an Element annotation with eletype=eo_authority_para.

If there is a paragraph annotation, and
   the last token annotation in the paragraph has string="".",
then create an Element annotation with eletype=para that spans
   from the beginning of the paragraph through the last token annotation.

3.5.5 NotInPara Phase

The NotInPara phase has a single rule. This rule matches PossElement annotations created in earlier phases and checks if the annotation is contained inside an Element annotation with eletype=para or eletype=eo_authority_para. If the matched PossElement annotation is not contained inside these Element annotations, the rule changes the PossElement annotation into an Element annotation and keeps the eletype from the PossElement annotation.

3.5.6 Element Phase

The Element phase has a single rule. This rule is based on Element annotations created in the previous phase. The rule matches Element annotations with eletype equal to subj, purpose, disctopics, declassifyon, postscript, or reference followed by a Token. It also matches Element annotation with eletype=eoid followed by Element annotation with eletype=emdash followed by a Token annotation. The rule then finds the containing paragraph annotation and checks if it has an Element annotation with eletype=classabbr at the end of it. If so it gets the annotation set spanning the matched Token annotation up to but not including the Element annotation with eletype=classabbr otherwise it gets the annotation set spanning the matched Token annotation through the end of the paragraph.
The rule then creates an Element annotation that spans the annotation set. It sets the eletype of the new Element annotation based on the matched Element annotation or annotations. If the matched eletype equals subj, then new eletype equals TOPIC. If the matched eletype equals purpose, then the new eletype equals purpose_para. If the matched eletype equals disctopics, then the new eletype equals para. If the matched eletype equals declassifyon, then the new eletype equals declass_instr. If the matched eletype equals postscript, then new eletype equals postscript. The span increases to include the matched Element, and the original matched Element is removed. If the matched eletype equals reference, then new eletype equals refers_to. Otherwise, the new eletype equals eotitle.

3.5.7 RemovePossElements Phase

The RemovePossElements phase has a single rule that removes all intermediate PossElements annotations created in the first two phases. These annotations are no longer necessary.

3.5.8 Documentum Phase

The Documentum phase has a single rule that creates a Documentum annotation. If the document is a text document, the annotation spans the entire document. If the document is an HTLM document, the annotation spans the body of the HTLM markup. This annotation indicates to the SUPPLE parser the span of the elements that it should parse.

4. Recognizing and Interpreting Documentary Forms

The Sheffield University Prolog Parser for Language Engineering (SUPPLE) provided with GATE is a bottom-up chart parser. The English grammar for sentences that is provided with SUPPLE is an attribute-value context-free grammar. As SUPPLE parses a sentence, it uses this grammar to construct a parse trees tree for the sentence and a logical representation of the meaning of the sentence.

The semantics of attribute-value context-free grammars is described as follows [Gaizauskas et al 2005]: “The semantic representation of a sentence is constructed compositionally from the semantics of the sentence’s constituents when the syntactic analysis takes place. Each noun phrase and each verb phrase in the sentence leads to the introduction of a unique identifier, e_k, which is used as a representation of the “entity” or “event” referred to by the noun/verb phrase. The semantics of a sentence is a conjunction of unary and binary predicates represented as a list of Prolog terms. We refer to this representation as a simplified quasi-logical form, or SQLF.”
The SUPPLE parser is written in Prolog and uses a GATE wrapper to run inside GATE. SUPPLE uses a layered chart parsing architecture. The set of rules at the first level is used to parse the input from the document. When the chart parsing algorithm has applied the rules in a layer, the resulting best-parses from that layer become the input to the next layer. This continues until all layers have been parsed. Each layer has a set of best-parse categories specified. These best-parse categories define what is passed on to the next layer.

The GATE wrapper requires two configuration files to create the initial chart for the SUPPLE chart parser. This chart is the representation of what is to be parsed and is the input for the first layer.

The original SUPPLE wrapper used the sentence annotation to collect the information used in creating the initial chart. It has been modified to use a parameter to pass the name of the annotation type to be used for collecting the information necessary to create the initial chart. The annotation used to identify the documentary form of a document is the Documentum annotation.

4.1 Configuration Files

The mapping.config file is a bridge between GATE annotations and SUPPLE categories. The SUPPLE wrapper uses this file to create SUPPLE categories and set features corresponding to the associated GATE annotation features. In this file, a “%” is used to specify a comment and “&” is used to specify a variable. Using “&” allows the value of the string feature of the GATE annotation to be passed to the SUPPLE category as the value of its text and name features. The SUPPLE wrapper matches the specified annotation type and any features whose value is not a variable to create the SUPPLE categories used in the initial chart.

Shown below are few of the entries in the mapping.config file that is used to parse the intellectual elements of documents.

```
% ADDRESS %
Gate:AnnotationType=Element;etype=address;string=&S
SUPPLE:category=sem_cat;type=Address;text=&S;kind=address;name=&S

% CHRONDATE %
Gate:AnnotationType=Element;etype=chrondate;string=&S
SUPPLE:category=sem_cat;type=Chrondate;text=&S;kind=date;name=&S

% FORMULA_OF_RESPECT %
Gate:AnnotationType=Element;etype=formula_of_respect;string=&S
SUPPLE:category=sem_cat;type=Formula_Of_Respect;text=&S;kind=formula_of_respect;name=&S

% PARA %
Gate:AnnotationType=Element;etype=para;string=&S
SUPPLE:category=sem_cat;type=Para;text=&S;kind=para;name=&S
```
The other configuration file, `feature_table.config`, specifies the lexical categories of the grammar and their features. The only entry in the table is the sem_cat category.

```
sem_cat;s_form;text;type;kind;name
```

### 4.2 Grammars for Recognizing and Interpreting Documentary Forms

Just as sentence form and content convey meaning, documentary form and content convey meaning. For instance, the meaning of the documentary form of a memo is something like “a memorandum with chronological date d addressed to x from author y is about topic z.”

SUPPLE rules are specified in Prolog with the predicate `rule(LHS,RHS)` where LHS is a syntactic category and RHS is a list of syntactic categories that need to be available to make the rule fire. A rule $A \rightarrow A_1 \ldots A_n$ is written as follows:

```
rule(A, [A_1,\ldots A_n]).
```

SUPPLE rules support not only parsing, but the interpretation of the content of the documentary form. Rules represent semantic information as values of the sem feature of categories. Semantic information is passed to the sem feature of the category on the left-hand side of a rule by combining elements of the semantic representation found in the sem feature of the list of categories on the right-hand side of the rule. The following example shows how person and job title categories pass semantic information up to the entity category. The hat (^) operator is used to create identifiers that are available for combining semantic information when a rule is fired. Any feature value that starts with a capital letter is a variable that can be used to pass the value. In this way the right-hand side of a rule passes values to the left-hand side of the rule.

```
rule(entity(s_form:F,sem:E^[name,E,PERSO],[jobtitle,E,JOBTITLE])),
person(s_form:F,sem:PERSON), jobtitle(s_form:F,sem:JOBTITLE)).
```
Declarative, interrogative and imperative sentences have different forms, but they share constituents, for example, noun phrase, prepositional phrases and verb phrases. Grammars for each of the sentence types share grammar rules for the common constituents. Similarly, memoranda, correspondence and reports have different forms, but share constituents, for example, paragraphs and lists of person’s names. Hence, grammars for different documentary forms can be merged so that they share rules, and each specific form is a subtype of documentary form.

The SUPPLE parser is a layered, bottom up, chart parser. Bottom-up chart parsers have the advantage that they can generate all possible parses of sentence, which is important because sentences with multiple parsers are likely to be ambiguous. They have the disadvantage that for large grammars, they are slow. The SUPPLE parser overcomes this difficulty by layering the application of grammar rules. All lowest level constituents, e.g., ENTITY, are identified before attempting to identify higher-level constituents, e.g., ENTITIES. All the levels except the final level identify sub-forms. The rules in the top layer are used for the final identification of the document form. The following sections describe some of the rules in each layer, specifically the rules that are necessary to recognize the documentary form of memoranda.

### 4.2.1 The semantic_rules layer

The semantic_rules layer is the first and lowest layer. It contains rules that take a SUPPLE category from the initial chart and assigns a semantic value to it. The rules create facts in the Prolog database. Below are the rules of this layer that are necessary to identify a simple memo. Some of these rules are also needed to identify other documentary forms.

```prolog
rule(chrondate(s_form:F, sem:NAME),
     [sem_cat(s_form:F,text:TEXT,type:'Chrondate',kind:KIND,name:NAME)]).
rule(for(s_form:F, sem:NAME),
     [sem_cat(s_form:F,text:TEXT,type:'To',kind:KIND,name:NAME)]).
rule(person(s_form:F, sem:NAME),
     [sem_cat(s_form:F,text:TEXT,type:'Person',kind:KIND,name:NAME)]).
rule(jobtitle(s_form:F, sem:NAME),
     [sem_cat(s_form:F,text:TEXT,type:'JobTitle',kind:KIND,name:NAME)]).
rule(from(s_form:F, sem:NAME),
     [sem_cat(s_form:F,text:TEXT,type:'From',kind:KIND,name:NAME)]).
rule(subj(s_form:F, sem:NAME),
     [sem_cat(s_form:F,text:TEXT,type:'Subj',kind:KIND,name:NAME)]).
rule(topic(s_form:F, sem:NAME),
     [sem_cat(s_form:F,text:TEXT,type:'Topic',kind:KIND,name:NAME)]).
rule(para(s_form:F, sem:NAME),
     [sem_cat(s_form:F,text:TEXT,type:'Para',kind:KIND,name:NAME)]).
rule(attachment(s_form:F, sem:NAME),
     [sem_cat(s_form:F,text:TEXT,type:'Attachment',kind:KIND,name:NAME)]).
```
4.2.2 The entity_rules layer

The entity_rules layer is used to pass the best-parse of the entity category to the next higher layer. Below are shown the rules of this layer. The rules of this layer are needed to identify a simple memo as well as other documentary forms. The symbol %% is used to denote comments. The comment above each rule shows the context-free grammar rule.

%% ENTITY--->PERSON JOBTITLE
rule(entity(s_form:F,sem:E^[name,E,PERSON],[jobtitle,E,JOBTITLE]),
   [person(s_form:F,sem:PERSON),
    jobtitle(s_form:F,sem:JOBTITLE)]).

%% ENTITY--->JOBTITLE
rule(entity(s_form:F,sem:E^[jobtitle,E,JOBTITLE]),
   [jobtitle(s_form:F,sem:JOBTITLE)]).

%% ENTITY--->PERSON
rule(entity(s_form:F,sem:E^[name,E,PERSON]),
   [person(s_form:F,sem:PERSON)]).

4.2.3 The entities_rules layer

The entities_rules layer is used to pass the best-parse of an entities category to the next layer. Shown below are the rules of this layer.

%%ENTITIES--->ENTITIES ENTITY
rule(entities(s_form:F,sem:E1^X^E2^[member,E1,E2],S1),
   [entities(s_form:F,sem:E1^X^S1),
    entity(s_form:F,sem:E2^S2)]).

%% ENTITIES--->ENTITY
rule(entities(s_form:F,sem:E^[SEM]),
   [entity(s_form:F,sem:E^[SEM])]).

4.2.4 The para_rules layer

The paras category is a simple group of one or more of the para category elements. The category paras do not collect any semantic information from the para category to pass on to the next layer.

%% PARAS-->PARA PARAS
rule(paras(s_form:F),
   [para(s_form:F),
    paras(s_form:F)]).

%% PARAS-->PARA
rule(paras(s_form:F),
   [para(s_form:F)]).
4.2.5 The optional_rules layer

The optional_rules layer handles optional intellectual elements that can appear close to or at the end of memoranda.

```mermaid
%OPTIONAL--->ATTACHMENT CCLIST BCCLIST
rule(optional(s_form:F),
 [attachment(s_form:F),
  cclist(s_form:F)],
 bcclist(s_form:F)));

%OPTIONAL--->ATTACHMENT BCCLIST
rule(optional(s_form:F),
 [attachment(s_form:F),
  bcclist(s_form:F)));

%OPTIONAL--->ATTACHMENT CCLIST
rule(optional(s_form:F),
 [attachment(s_form:F),
  cclist(s_form:F)));

%OPTIONAL--->ATTACHMENT
rule(optional(s_form:F),
 [attachment(s_form:F)));

%OPTIONAL-->CCLIST BCCLIST
rule(optional(s_form:F),
 [cclist(s_form:F),
  bcclist(s_form:F)));

%OPTIONAL-->BCCLIST
rule(optional(s_form:F),
 [bcclist(s_form:F)));

%OPTIONAL-->CCLIST
rule(optional(s_form:F),
 [cclist(s_form:F)));
```

4.2.6 The subform_rules layer

Sub-form refers to the fact that these categories are built from other categories and are used by the final rules that identify the documentary form of a document. The following rules identify constituents of a simple memo.

```mermaid
%MEMOHEAD-->CHRONDATE ADDRLINE SNDRLINE SUBJLINE
rule(memohead(s_form:F,sem:E1^E2^DATE,ADDRList,SNDRList,TOPIC)),
 [chrondate(s_form:F,sem:DATE),
  addrline(s_form:F,sem:E1^ADDRList),
  sndrline(s_form:F,sem:E2^SNDRList),
  subjline(s_form:F,sem:TOPIC))}.
```
4.2.7 The docform_rules layer

The rules in this layer are the rules that complete the identification of a documentary form. These rules also construct a logical form representing the meaning of each documentary form. Currently, nine documentary forms are recognized. The nine forms are eo (Executive Order), letter (Formal letter), memo (White House Memorandum), minutes (Minutes of Cabinet Meeting), natsecdir (National Security Directive), natsecrev (National Security Review), rtcmemo (Recommended Telephone Call), whletter (White House Letter), and whpressrelease (White House Press Release). The following two rules define the form of a White House Memorandum.

%% MEMO--->MEMOHEAD BODY
rule(memo(s_form:F,sem:D\^E2\^E1\^[document,D],
  [document_form,D,'White House Memorandum'],[author,D,E2],
  SNDRList,[addressee,D,E1],ADDRList,[topic,D,TOPIC], [date,D,DATE]),
  [memohead(s_form:F,sem:E1\^E2\^[DATE,ADDRList,SNDRList,TOPIC]),
  body(s_form:F))).

%% MEMO--->MEMOHEAD BODY OPTIONAL
rule(memo(s_form:F,sem:D\^E2\^E1\^[document,D],
  [document_form,D,'White House Memorandum'],[author,D,E2],
  SNDRList,[addressee,D,E1],ADDRList,[topic,D,TOPIC], [date,D,DATE]),
  [memohead(s_form:F,sem:E1\^E2\^[DATE,ADDRList,SNDRList,TOPIC]),
  body(s_form:F),
  optional(s_form:F))}.
5. Test of the Documentary Form Recognizer and Metadata Extraction

The automated method for recognizing documentary forms was tested by applying it to nine documents of different documentary forms. The Document Element Tagger was used to recognize the intellectual elements of the documents. Rules for nine of the context-free grammars developed for documentary forms were translated to attribute-value grammars and used by the SUPPLE parser to parse and recognize the documentary form of the sample documents. The parser also generated quasi-logical forms representing the meanings of the documents as expressed by the documentary form.

Figure 2 shows the list representation created by SUPPLE of the parse tree of the sample memorandum shown in Figure 1. A parse tree represents the intellectual form (structure) of a document.

```
{best_parse=(memo
  (head (chrondate (sem_cat "April 27, 1992")))
   (addrline (for (sem_cat "MEMORANDUM FOR"))
     (entities (entity (person (sem_cat "SAM SKINNER"))))
   (sndrline (from (sem_cat "FROM:"))
     (entities (entity (person (sem_cat "EDE HOLIDAY"))))
   (subjline (sub) (sem_cat "SUBJECT:"))
     (topic (sem_cat "California Earthquake"))))
  (body (paras (para
     (sem_cat "Attached is a situation report from FEMA on the northern California earthquake. No deaths have been reported and 45 people are known to have suffered injuries. In addition, there has been extensive property damage. While FEMA is awaiting a request from the State before initiating any recovery activities, a joint State/Federal preliminary damage assessment is likely to begin today.")
   (paras (para
     (sem_cat "Director Stickney has requested that we forward the situation report to you.")
   ))(optional (attachment (sem_cat "Attachments"))))

Figure 2. A Parse Tree for a Sample Memorandum
```

Figure 3 shows the quasi-logical form (qlf) created by SUPPLE of the semantics of the memorandum.

```
{qlf=[document(e1),
document_form(e1, memo),
author(e1, ‘EDE HOLIDAY’),
addressee(e1, ‘SAM SKINNER’),
topic(e1, ‘California Earthquake’),
date(e1, ‘April 27, 1992’)]}

Figure 3. The Semantics of the Sample Memorandum
The following table shows the documentary form recognize, and quasi-logical semantic representation of the documents in the test case. The documents are from a corpus constructed specifically for tests and experiments with Presidential e-records [Underwood and Hayslett-Keck 2004].

<table>
<thead>
<tr>
<th>File Name</th>
<th>Documentary Form</th>
<th>Quasi-Logical Semantic Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>048.txt</td>
<td>Formal letter</td>
<td>{qlf=[document(e1), document_form(e1, 'Formal Letter'), author(e1, e2), name(e2, 'Paula Sweeney'), topic(e1, 'Darrell Panethiere'), addressee(e1, e3), name(e3, 'The Honorable Lloyd Bentsen'), date(e1, 'August 25, 1992')]}</td>
</tr>
<tr>
<td>051.txt</td>
<td>White House Press Release</td>
<td>{qlf=[document(e1), document_form(e1, 'White House Press Release'), date(e1, 'October 2, 1990')]}</td>
</tr>
<tr>
<td>053.txt</td>
<td>Recommended Telephone Call</td>
<td>{qlf=[document(e1), document_form(e1, 'Recommended Telephone Call'), author(e1, e2), set(e2), member(e2, e5), name(e5, 'The Attorney General'), member(e2, e9), name(e9, 'Samuel K. Skinner'), member(e2, e8), name(e8, 'Nicholas E. Calio'), member(e2, e7), name(e7, 'C. Boyden Gray'), member(e2, e6), name(e6, 'Constance Horner'), addressee(e1, e3), name(e3, 'The President'), date(e1, 'April 27, 1992')]}</td>
</tr>
<tr>
<td>078.txt</td>
<td>White House Memorandum</td>
<td>{qlf=[document(e1), document_form(e1, 'White House Memorandum'), author(e1, e2), name(e2, 'EDE HOLIDAY'), addressee(e1, e3), name(e3, 'SAM SKINNER'), topic(e1, 'California Earthquake'), date(e1, 'April 27, 1992')]}</td>
</tr>
<tr>
<td>086.txt</td>
<td>White House Letter</td>
<td>{qlf=[document(e1), document_form(e1, 'White House Letter'), author(e1, e2), name(e2, 'Doug Wead'), jobtitle(e2, 'Special Assistant to the President for Public Liaison'), addressee(e1, e3), name(e3, 'Ms. Lucille Anderson'), date(e1, 'April 25, 1990')]}</td>
</tr>
<tr>
<td>119.txt</td>
<td>Executive Order</td>
<td>{qlf=[document(e1), document_form(e1, 'Executive Order'), author(e1, e2), name(e2, 'George Bush'), addressee(e1, e3), name(e3, 'Public'), topic(e1, 'Implementation of Section 311(a) of the Comprehensive Anti-Apartheid Act'), date(e1, 'July 10, 1991')]}</td>
</tr>
<tr>
<td>151.txt</td>
<td>National Security Review</td>
<td>{qlf=[document(e1), document_form(e1, 'National Security Review'), nsr_id(e1, 'NATIONAL SECURITY REVIEW-2'), classification(e1, 'UNCLASSIFIED'), addressee(e1, e2), set(e2), member(e2, e5), name(e5, 'THE VICE PRESIDENT'), member(e2, e10), name(e10, 'THE SECRETARY OF STATE'), member(e2, e9), name(e9, 'THE SECRETARY OF DEFENSE'), member(e2, e8), name(e8, 'THE DIRECTOR OF CENTRAL INTELLIGENCE'), member(e2, e7), name(e7, 'THE ASSISTANT TO THE PRESIDENT FOR NATIONAL SECURITY AFFAIRS'), member(e2, e6), name(e6, 'THE CHAIRMAN, JOINT CHIEFS OF STAFF'), topic(e1, 'Panama'), date(e1, 'February 13, 1989')]}</td>
</tr>
<tr>
<td>157.txt</td>
<td>Minutes of Cabinet Meeting</td>
<td>{qlf=[document(e1), document_form(e1, 'Minutes of Cabinet Meeting'), date(e1, 'WEDNESDAY, JUNE 25, 1975')]}</td>
</tr>
<tr>
<td>187.txt</td>
<td>National Security Directive</td>
<td>{qlf=[document(e1), document_form(e1, 'National Security Directive'), nsd_id(e1, 'NATIONAL SECURITY DIRECTIVE 5'), classification(e1, 'UNCLASSIFIED'), addressee(e1, e2), set(e2), member(e2, e5), name(e5, 'THE VICE PRESIDENT'), member(e2, e14), name(e14, 'THE SECRETARY OF STATE'), member(e2, e13), name(e13, 'THE SECRETARY OF THE TREASURY'), member(e2, e12), name(e12, 'THE ACTING SECRETARY OF T'))</td>
</tr>
<tr>
<td>File Name</td>
<td>Documentary Form</td>
<td>Quasi-Logical Semantic Form</td>
</tr>
<tr>
<td>-----------</td>
<td>------------------</td>
<td>-----------------------------</td>
</tr>
<tr>
<td></td>
<td></td>
<td>DEFENSE\r\v, member(e2, e11), name(e11, 'THE ATTORNEY GENERAL'), member(e2, e10), name(e10, 'THE DIRECTOR, OFFICE OF MANAGEMENT AND BUDGET'), member(e2, e9), name(e9, 'THE DIRECTOR OF CENTRAL INTELLIGENCE'), member(e2, e8), name(e8, 'THE ASSISTANT TO THE PRESIDENT FOR NATIONAL SECURITY AFFAIRS'), member(e2, e7), name(e7, 'THE CHAIRMAN, JOINT CHIEFS OF STAFF'), member(e2, e6), name(e6, 'THE ADMINISTRATOR, AGENCY FOR INTERNATIONAL DEVELOPMENT'), topic(e1, 'Legislation to Authorize the Transfer of Funds to the Agency for International Development (AID) for Humanitarian Assistance to Afghanistan'), date(e1, 'March 18, 1989'))</td>
</tr>
</tbody>
</table>

Figure 4. Quasi-logical Form for the Semantics of the Documents in the Test Case

Document metadata extraction, which is not yet implemented, will be accomplished by translating the quasi-logical form to attribute-value pairs. For example, the metadata created for the quasi-logical form shown in Figure 3 is

```
DOCTYPE = memo
AUTHOR = "EDE HOLIDAY"
ADDRESSEE = "SAM SKINNER"
TOPIC = "California Earthquake"
DATE = "April 27, 1992"
```

6. Related Research

The theory of text structure proposed by Nunberg [1990] distinguishes syntactic structure from text structure. The syntactic structure of a sentence is defined by phrase structure rules such as $S \rightarrow \text{NP} + \text{VP}$, while the text structure of a sentence is realized by a sequence of words that begins with a capitalized word and ends in a period. Nunberg refers to the latter as a text-sentence, to distinguish it from the syntactic category sentence. He introduces the concepts of text-clause (text that ends in a semicolon) and paragraph (begins on a new line, sometimes with a tab). He introduces the concept of a text grammar, in contrast to a syntactic (or lexical grammar), that allows one to define text-sentence, text-clauses and paragraphs with constituent structure rules. Nunberg’s work relates to the research reported in this paper by providing a formal framework for recognizing sentences and paragraphs and distinguishing them from other elements in a document. Nunberg’s theory of text structure might be used to improve the performance of the GATE Document Reader and Sentence Splitter.

Power et al [2003] are concerned with Natural Language Generation (NLG), namely, the generation of text documents (including graphical features and layout) from a rhetorical structure. They introduce the concept of abstract document structure as a level of representation that mediates between the discourse structure of a text and its physical presentation. Abstract document structure is analogous to the concept of intellectual form.
Our research differs from theirs in being concerned with Natural Language Understanding (NLU) of text documents, including the recognition of document structure of digital documents.

The research of Kim and Ross [2007a, 2007b] has an objective similar to that of the research reported in this paper – the recognition of genre (documentary form) for the purpose of metadata extraction from digital records ingested into digital archives or libraries. Their approach is to identify features of documents that will allow them to automatically classify documents by genre. The features they have identified include: image features, syntactic features, stylistic features, semantic structure, and domain knowledge features. These features are used with an image classifier, an n-gram model classifier and a stylo-metric classifier. Our research differs from theirs in that our approach is to recognize intellectual and physical elements of a document’s form, and to recognize the document genre by parsing the elements using a grammar describing the structure of each genre. Currently, our grammars are manually constructed, but our next step is to improve on our method of inducing the grammars for documentary form from a sample of documents of a particular genre.

Researchers in Image Analysis have addressed the analysis document images [Haralick 1994, Dori et al 1997]. The physical structure of a document image is distinguished from its logical structure. Physical units include blocks, frames, and pages. Logical units include characters, words, sentences, paragraphs, sections and headers. Document type classification can be accomplished without OCR by introducing an interval encoding that captures elements of the spatial layout of the document and then classifying the documents using Hidden Markov Models (HMMs) [Hu et al. 1999]. In the WISDOM++ system, decision trees and first-order logic learning is used to classify segments or blocks of text and inductively learn rules for layout-based classification [Esposito et al 2000]. Our research differs from the research in document image analysis in that the method for recognition of documentary form described in this paper operates on digital documents in textual file formats, not image formats. However, their insights in document image analysis may be useful in addressing the Text Segmentation Problem, especially in the case generated form OCRed document images and in plain text format.

7. Summary of Results

The documentary form of a record is a key element of archival descriptions of record series, file units and items. The ability to recognize document type also enables the capability to extract metadata from e-records that can be used as access points for record search and retrieval.

Context-free grammars have been constructed for twenty-two of the documentary forms that occur in Presidential e-records. JAPE rules were constructed for recognizing in documents the intellectual elements occurring in nine of these documentary forms. The nine grammars were translated into context-free attribute grammars that were used with the SUPPLE parser provided with GATE to parse the intellectual elements of
documentary form. With SUPPLE, we were also able to create the semantics of the documentary forms. This semantic representation can be used to extract metadata needed for archival description and for record search and retrieval.

Parsing of intellectual elements using manually constructed grammars for documentary forms will not recognize all examples of these document types because people do not always follow exactly the documentary form prescribed in style manuals. In prior research, the automatic induction of grammars for characterizing two documentary forms of e-records was investigated [Underwood and Harris 2006]. It is planned to collect samples (100 documents or more) of e-records of the document types considered in this paper, and automatically induce grammars from these samples.

The grammars that are induced from a sample of documents of a particular type often do not describe the documentary form the same way that a person would describe its form. This is because the current induction method does not consider the semantics of the documentary form, while a person does. We will explore the use of manually constructed grammars for a documentary form as a reference grammar that contains intellectual element names and rules that should be included in describing the form. This will bias the induced grammar to include some conventional terms and phrases for describing the form.

The current technique for inducing grammars for documentary forms uses only two operators—chunking and merging. These two operators are capable of producing all possible grammars for a sample, and even generalizing to include strings not in the sample. However, some of the grammars generated are non-intuitive, i.e., a person would not describe the structure of the document in that way. We intend to investigate use of other operators to see if more intuitive grammars can be induced.

A method for identification of intellectual elements might be based on identification of common elements by automatic, rather than manual, comparison of semantically annotated documents of the same documentary form. This would enhance the utility of the documentary form recognition technology.

The Text Segmentation method used in the Document Reader and Sentence Splitter provided with GATE needs to be improved. The relevance of related research in document image analysis and text grammars will be investigated.

The definition of the structure of documents including sections, lists, bulleted items, and tables needs to be investigated. The identification of physical elements of form such as bold face, italics face, underline, and different font and font sizes also needs to be investigated. The conversion of documents in legacy file formats to HTML is a first step in this direction.

Finally, experiments should be conducted to evaluate the performance of the documentary form induction, recognition and interpretation methods when applied to a representative sample of Presidential e-records.
References


Appendix A: Grammars for Documentary Forms

In this appendix, context-free grammars are shown that define the documentary forms of twenty-one types of records that occur in the Bush Presidential e-records.

**Formal Letter**

A *formal letter* is an external communication which has a date, inside address, salutation, body or message, a complimentary close, and a signature block. Optional elements beneath the signature block include reference initials, enclosure(s), courtesy copies, blind courtesy copies and a postscript. This style is specified in the US Government Correspondence Manual. Figure 5 shows an example of a formal letter.

August 25, 1992

The Honorable Lloyd Bentsen
703 Hart Senate Office Building
Washington, D.C. 20510

RE: Darrell Panethiere

Dear Senator Bentsen:

I've been told that Darrell Panethiere is being considered for appointment to the Fifth Circuit Court of Appeals.

I am pleased to heartily endorse him for that position. In the years during which we practiced together, I had the opportunity to observe Mr. Panethiere in many appellate situations. He is a dedicated scholar, a good legal analyst, a man of extremely high integrity, and is well respected by his fellow members of the Bar.

I believe that the broad scope of knowledge and experience he would bring to the Fifth Circuit would be a tremendous asset to that Court.

With kind regards,

Paula Sweeney

PS/DSM

cc: Mr. C. Boydman Gray
Council to the President

---

2 Bush Presidential Library, Bush Presidential Records, WHORM - Subject Files, Federal Government, Judicial Branch, Court of Appeals, ID#347712
Grammar rules defining the form of a Formal Letter are shown below.

LETTER → CHRONDATE INSIDE_ADDRESS REFLINE SALUTATION BODY
    COMPCLOSE SIGNBLK OPTIONAL
LETTER → CHRONDATE INSIDE_ADDRESS REFLINE SALUTATION BODY
    COMPCLOSE SIGNBLK
LETTER → CHRONDATE INSIDE_ADDRESS SALUTATION BODY COMPCLOSE
    SIGNBLK OPTIONAL
LETTER → CHRONDATE INSIDE_ADDRESS SALUTATION BODY COMPCLOSE
    SIGNBLK
INSIDE_ADDRESS → ENTITY ADDRESS
REFLINE → REFERENCE REFERS_TO
SALUTATION → FORMULA_OF_ADDRESS PERSON
BODY → PARAS
COMPCLOSE → FORMULA_OF_RESPECT
SIGNBLK → ENTITY
OPTIONAL → REFINITIALS OPTIONALE
OPTIONAL → REFINITIALS OPTIONALC
OPTIONAL → REFINITIALS OPTIONALB
OPTIONAL → REFINITIALS POSTSCRIPT
OPTIONAL → REFINITIALS OPTIONALE → ENCLOSURE OPTIONALC
OPTIONALE → ENCLOSURE OPTIONALB
OPTIONALE → ENCLOSURE POSTSCRIPT
OPTIONALE → ENCLOSURE
OPTIONALC → CCLIST OPTIONALB
OPTIONALC → CCLIST POSTSCRIPT
OPTIONALC → CCLIST
OPTIONALB → BCCLIST POSTSCRIPT
OPTIONALB → BCCLIST
OPTIONAL → POSTSCRIPT
REFINITIALS → TYPIST
REFINITIALS → SIGNER TYPIST
REFINITIALS → SIGNER WRITER TYPIST
CCLIST → COURTESY_COPY ENTITIES
BCCLIST → BCC ENTITIES
ENTITIES → ENTITIES ENTITY
ENTITIES → ENTITY
PARAS → PARA PARAS
PARAS → PARA
ENTITY → PERSON JOBTITLE
ENTITY → JOBTITLE
ENTITY→ PERSON

“Dear” is a FORMULA_OF_ADDRESS. “Sincerely.” “Sincerely yours,” and “Respectfully” are examples of a FORMULA_OF_RESPECT. POSTSCRIPT is “P.S.” or “PS. ”ENCLOSURE is “Enclosure”, “Enclosures” Enclosures (digit) or “Enc.” TYPIST is lowercase initials. SIGNER and WRITER are upper case initials. COURTESY_COPY is “cc:”.BCC is “bcc:”. REFERENCE is “Re:”.

29
White House Letter

A White House letter is an external communication which has a date, salutation, body or message, a complimentary close, and a signature block, followed by an inside address. This style is specified in the White House Correspondence Manual. The documentary form is essentially that of a formal letter printed on White House stationery, but with the inside address following the signature block. Fig. 6 shows an example of a White House Letter.³

```
April 25, 1990
Dear Ms. Anderson:
Thank you very much for your letter of April 9, 1990 regarding your concern for clean air.
As you know, the Administration is firmly committed to seeing that the environment is free of pollutants, that our air is pure and clear, that the water of our rivers and lakes is free of contaminants, and that our food supply is safe.
Thank you for writing on this most important matter.
Sincerely,
Doug Wead
Special Assistant to the President for Public Liaison

Ms. Lucille Anderson
316 Nevada Avenue
Carterville, IL 62918
```

Figure 6. An Example of a White House Letter

The following grammar rules for White House Letters were developed by referral to the White House Correspondence manual, and several examples of White House Letters.

```
WHLETTER → CHRONDATE SALUTATION BODY COMPCLOSE
SIGNBLK INSIDE_ADDRESS OPTIONAL
BODY → PARAS
COMPCLOSE → FORMULA_OF_RESPECT
SIGNBLK → ENTITY
```

³ Bush Presidential Library, Bush Presidential Records, Public Liaison, Doug Wead’s Files, [Alpha Correspondence File 2/90-6/90 “A” “B”].
INSIDE_ADDRESS → ENTITY ADDRESS
OPTIONAL → REFINITIALS OPTIONALE
OPTIONAL → REFINITIALS OPTIONALC
OPTIONAL → REFINITIALS OPTIONALB
OPTIONAL → REFINITIALS POSTSCRIPT
OPTIONAL → REFINITIALS
OPTIONALE → ENCLOSURE OPTIONALE
OPTIONALE → ENCLOSURE OPTIONALC
OPTIONALE → ENCLOSURE OPTIONALB
OPTIONALE → ENCLOSURE POSTSCRIPT
OPTIONALE → ENCLOSURE
OPTIONALC → CCLIST OPTIONAL:B
OPTIONALC → CCLIST POSTSCRIPT
OPTIONALC → CCLIST
OPTIONALB → BCCLIST POSTSCRIPT
OPTIONALB → BCCLIST
OPTIONAL → POSTSCRIPT
REFINITIALS → TYPIST
REFINITIALS → SIGNER TYPIST
REFINITIALS → SIGNER WRITER TYPIST
CCLIST → COURTESY_COPY ENTITIES
BCCLIST → BCC ENTITIES
ENTITIES → ENTITIES ENTITY
ENTITIES → ENTITY
PARAS → PARA PARAS
PARAS → PARA
ENTITY → PERSON JOBTITLE
ENTITY → PERSON


**Presidential Memoranda**

For Presidential communications, the Bush and Clinton Administrations inherited a number of documentary forms from the Reagan Administration.

Within a week of taking over as staff secretary [of the Reagan Administration], [Richard] Darman issued a lengthy memo outlining how memoranda were to be prepared and routed through the White House and on to the president. Darman outlined procedures covering six different types—briefing papers, decision memoranda, signature memoranda, information memoranda, telephone call recommendations, and schedule proposals—and different protocols and formats were set out for each and samples were provided. Briefing papers for meetings, for example, required not only names, dates, location, and time but also purpose, background, participants, press plan, sequence of events, outline of agenda, president’s role, and talking points (where appropriate). [Burke 2000, p. 155]
Decision Memoranda

Decision memoranda are a specific document type differing from other White House Memoranda. Decision memoranda are prepared for substantive matters requiring Presidential decision. Figures 7 and 8 illustrate the parts of a decision memorandum.

```
April 5, 1993

MEMORANDUM FOR THE PRESIDENT
FROM: Carol Rasco
SUBJECT: Announcing a time schedule for unveiling your national health reform plan and introducing legislation

I. ACTION-FORCING EVENT

Recent press reports have created uncertainty around the timing of introduction of your national health reform plan. This uncertainty is having a major impact on both Congress and the general public and is diminishing chances for enactment of major reforms. Therefore, this week you should announce a schedule for unveiling the plan and introducing legislation.

This memorandum proposes a time frame for meeting your original deadline in a consistent fashion. It also identifies the separate stages, products, and sub-dates that meeting this time frame entails. Finally, we set forth the factors that you need to consider in deciding whether you wish either to proceed with this plan or to alter it in one or more respects.

II. BACKGROUND/ANALYSIS

You have stated that no other area of domestic policy is in greater need of comprehensive reform than the nation's health care system. At the heart of the problem -- the millions of uninsured persons, the millions more who are on the verge of losing coverage, and the inequitable access to quality care -- is one overriding dilemma: the system's enormous cost. The cost of American health care fundamentally threatens the private economy. It threatens the vitality of the federal, state and local governments. And it undermines the security of today's insured Americans in the future of their own health coverage.

For all of these reasons, health care ranks at the top of your policy agenda. In January you identified health reform as one of the centerpieces of your first Presidential term and appointed the First Lady to head a task force that would study the problem and recommend comprehensive reforms.

Immediately following your inauguration you formed an Interagency Task Force on Health Care Reform and announced that within 100 days a comprehensive plan would be prepared. After several months of major effort, Task Force staff have developed several proposals for you. Given the nearness of their completion date, a time line for unveiling the plan to the public and to the Congress is essential.
```

Figure 7. An Example of a Decision Memorandum

---

4 Clinton Presidential Library. Health Care Interdepartmental Working Group.  
www.judicialwatch.org/clinton-library-documents
schedules outlined above. Follow-up memoranda on these decisions are now being prepared.

III. RECOMMENDATION

In light of these considerations, we recommend that you approve a three-part "unveiling" plan that assumes a Task Force release by May 10th to 15th, a major speech on May 25th, and the submission of legislation immediately following the Memorial Day recess. We also recommend that this week you publicly announce your schedule for proceeding.

IV. DECISION

___ Approve   ___ Approve as amended   ___ Reject

___ No action

Figure 8. Example of Concluding Part of a Decision Memorandum

The context-free grammar rules below define the documentary form of a Decision Memorandum.

DECISIONMEMO → DECHEAD DECBODY
DECHEAD → DATE ADDRLINE SNDRLINE SUBJLINE
ADDRLINE → FOR ENTITY
SNDRRLINE → FROM ENTITIES
SUBJLINE → SUBJ TOPIC
DECBOdy → EVENTSECT BKGRDSECT RCMDSECT DECSECT
ENTITIES → ENTITIES ENTITY
ENTITIES → ENTITY
EVENTSECT → EVENT PARAS
BKGRDSECT → BKGRND PARAS
RCMDSECT → RCMD PARAS
PARAS → PARAS PARA
PARAS → PARA
DECSECT → DECISION ULINE APPROVE ULINE APPROVEASAMENDED ULINE REJECT ULINE NOACTION

Intellectual element EVENT is “I. ACTION-FORCING EVENT”.
BKGRND is “II. BACKGROUND/ANALYSIS”.
RCMD is “III. RECOMMENDATION”.
DECISION is “IV. DECISION”.

Information Memoranda

Information memoranda are intended to convey information, not to elicit Presidential action. The specification for an information memo requires that it have the line
"INFORMATION MEMORANDUM FOR THE PRESIDENT" and that INFORMATION be underlined. Figure 9 shows an example of an Information Memo.5

May 4, 1989

INFORMATION MEMORANDUM FOR THE PRESIDENT

FROM: MICHAEL J. BOSKIN

SUBJECT: Employment and Unemployment in April, Labor Department Release, Tomorrow Morning, 6:30 a.m

According to the household survey, the unemployment rate for civilian workers rose 0.3 percentage point to 5.3 percent in April. The unemployment rate for all workers, including the armed forces, also rose 0.3 percentage point to 5.2 percent. The unemployment rates for teenagers and Hispanics recorded the largest increases. Nevertheless, the employment to population ratios for all workers and for civilian workers remained at their all-time highs of 63.3 percent and 63.0 percent, respectively.

Nonfarm payroll employment rose modestly in April. Nonfarm employment increased by 117,000 jobs in April according to the survey of business establishments. The April increase was less than the revised March increase of 171,000 jobs, which was depressed by the strike at Eastern. The payroll employment gains in April were the smallest since June 1986.

According to the payroll survey, employment in the goods-producing sector rose by 5,000 jobs, while employment in the service-producing sector rose by 112,000 jobs. Manufacturing jobs fell by 9,000 after rising a revised 6,000 jobs in March. Average weekly hours in manufacturing rose by 0.3 hours to 41.3 and manufacturing overtime hours rose by 0.1 hour to 4.0 hours, but these gains may reflect seasonal adjustment problems. Average hourly earnings in the nonfarm sector rose by 0.7 percent to $9.59 from a revised $9.52 in March.

Figure 9. An Example of an Information Memorandum

The following grammar rules define the intellectual form of an Information Memorandum.

INFOMEMO → INFOHEAD BODY
INFOHEAD → DATE INFOFOR ENTITY SNDRLINE SUBJLINE
SNDRRLINE → FROM ENTITIES
SUBJLINE → SUBJECT TOPIC
BODY → PARAS

5 Bush Presidential Library, Bush Presidential Records, Council of Economic Advisors, Michael J. Boskin's Files
The intellectual element INFOFOR is “INFORMATION MEMORANDUM FOR”.

**Recommended Telephone Call**

A Recommended Telephone Call is a presidential memorandum that recommends a telephone call by the President. Fig. 10 shows an example.  

<table>
<thead>
<tr>
<th>Recommended Telephone Call</th>
</tr>
</thead>
<tbody>
<tr>
<td>TO: Ronald B. Leighton</td>
</tr>
<tr>
<td>TELEPHONE: Home: (206) 588-8389 Office: (206) 572-5050</td>
</tr>
<tr>
<td>DATE TO CALL: April 28, 1992</td>
</tr>
<tr>
<td>RECOMMENDED BY: The Attorney General, Samuel K. Skinner, Nicholas E. Calio, C. Boyden Gray, and Constance Horner</td>
</tr>
<tr>
<td>PURPOSE: To ask him to serve as United States District Judge for the Western District of Washington</td>
</tr>
<tr>
<td>TOPICS OF DISCUSSION: 1. Since 1976, Mr. Leighton has been affiliated with the law firm of Gordon, Thomas, Honeywell, Raianca, Peterson &amp; Dahlin as an associate and then as a partner. 2. He received a B.A. degree, cum laude, (1973) from Whitworth College and a J.D. degree (1976) from the University of California, Hastings College of Law. 3. Mr. Leighton is 40 years old, married, his wife's name is Sally, they have two children, and reside in Tacoma, Washington.</td>
</tr>
<tr>
<td>DATE OF SUBMISSION: April 27, 1992</td>
</tr>
<tr>
<td>ACTION</td>
</tr>
</tbody>
</table>
The following grammar defines the form of a Recommended Telephone Call.

```
RTCMEMO → RTCHEAD RTCBODY
RTCHEAD → RCMTELECALL TOLINE TELELINE CALLDATELINE RCMDBYLINE
TOLINE → FOR PERSON
TELELINE → TELEPHONE HOMEPH OFFICEPH
HOMEPH → HOME PHONENO
OFFICEPH → OFFICE PHONENO
CALLDATELINE → DTC DOCDATE
RCMDBYLINE → RCMDBY ENTITIES
RTCBODY → PURPOSESECT TOPICSECT SUBMISSIONSECT ACTIONSECT
PURPOSESECT → PURPOSE PURPOSE_PARA
TOPICSECT → DISCTOPICS PARAS
SUBMISSIONSECT → DATE_OF_SUBMISSION DOCDATE
ACTIONSECT → ACTION ULINE
PARAS → PARA PARAS
PARAS → PARA
ENTITIES → ENTITY_LIST CONJUNCT ENTITY
ENTITIES → ENTITY_LIST ENTITY
ENTITIES → ENTITY CONJUNCT ENTITY
ENTITIES → ENTITY ENTITY
ENTITY_LIST → ENTITY_LIST CONJUNCT ENTITY
ENTITY_LIST → ENTITY_LIST ENTITY
ENTITY_LIST → ENTITY CONJUNCT ENTITY
ENTITY_LIST → ENTITY ENTITY
ENTITY → PERSON JOBTITLE
ENTITY → JOBTITLE
ENTITY → PERSON
```

RCMDTELECALL is “RECOMMENDED TELEPHONE CALL.” TELEPHONE is “TELEPHONE:”. HOME is “Home:”. OFFICE is “Office:”. DTC is “DATE TO CALL:”. PURPOSE is “PURPOSE:”. DISCTOPICS is “TOPICS OF DISCUSSION:”.

The prescribed form of a Recommended Telephone Call differs slightly from the form of the example. In place of the Telephone Line there is a Contacts Section, and in place of the Topics of Discussion Section there is a Background Section. The prescribed version of the Recommended Telephone Call has not yet been defined.

**Memorandum of Conversation (MEMCON)**

A Memorandum of Conversation (MEMCOM) is a transcript or a recounting of an official conversation with a foreign contact or official. A sample MEMCON is shown in Fig. 11.⁷

---

Figure 11. A Memorandum of Conversation

The following grammar rules for the documentary form of a MEMCON do not consider the security classification markings, but they can be easily added.
The intellectual element MOC is “MEMORANDUM OF CONVERSATION”. DTP is “DATE, TIME AND PLACE:”. SPEAKER_PARA is a run-in paragraph in which the run-in title is a person’s name, and EOC is “(End of Conversation)” or “—End of Conversation—“

**Memorandum of Telephone Conversation (TELCOM)**

A Memorandum of Telephone Conversation (TELCOM) is a transcript or recounting of an official telephone conversation with a foreign contact or official. A sample TELCOM is shown in Fig. 12.8

The following grammar rules define the documentary form of TELCOMs.

```
TELCOM → MOTC SUBJLINE PTSLINE DTP_LINE CONVERSATION EOC
SUBJLINE → SUBJ TOPIC
PTSLINE → PARTICIPANTS ENTITIES
ENTITIES → ENTITY ENTITIES
ENTITIES → ENTITY
ENTITY → PERSON
ENTITY → PERSON JOBTITLE
DTP_LINE → DTP DATE TIME LOCATION
CONVERSATION → SPEAKER_PARA CONVERSATION
CONVERSATION → SPEAKER_PARA
SPEAKER_PARAS → PERSON__PARA
SPEAKER_PARAS → PERSON__PARA PARAS
PARAS → PARA PARAS
PARAS → PARA
```

---

8 George Bush Presidential Library, George Bush Presidential Records: National Security Council, Presidential Telcons File, Box No 3, Folder Title: February 1990 [OA/ID CF01731]
The intellectual element MOTC is “MEMORANDUM OF TELEPHONE CONVERSATION”.

Figure 12. A Memorandum of Telephone Conversation

**White House Referral Slip**

A White House Referral is a letter sent to the President, Vice President of the United States, or the First Lady that is referred to an agency for response, for appropriate action,
or for information. The White House Referral consists of the original correspondence and a referral slip. A sample White House Referral Slip is shown in Fig. 13.9

```
THE WHITE HOUSE OFFICE
REFERRAL
MARCH 26, 1992
TO: FEDERAL EMERGENCY MANAGEMENT AGENCY
ACTION REQUESTED: APPROPRIATE ACTION

DESCRIPTION OF INCOMING:
ID: 315058
MEDIA: LETTER, DATED MARCH 12, 1992
TO: PRESIDENT BUSH
FROM: THE HONORABLE BERYL ANTHONY JR.
U.S. HOUSE OF REPRESENTATIVES
WASHINGTON DC 20515
SUBJECT: URGES THE PRESIDENT TO DECLARE UNION COUNTY
IN ARKANSAS A MAJOR FEDERAL DISASTER AREA
AS A RESULT OF DAMAGE CAUSED BY TORNADOES

PROMPT ACTION IS ESSENTIAL -- IF REQUIRED ACTION HAS NOT BEEN
TAKEN WITHIN 9 WORKING DAYS OF RECEIPT, PLEASE TELEPHONE THE
UNDERSIGNED AT 456-7486.

RETURN CORRESPONDENCE, WORKSHEET AND COPY OF RESPONSE
(OR DRAFT) TO: AGENCY LIAISON, ROOM 91, THE WHITE HOUSE, 20500

SALLY KELLEY
DIRECTOR OF AGENCY LIAISON
PRESIDENTIAL CORRESPONDENCE
```

Figure 13. A White House Referral Slip

The following grammar rules define the form of a White House Referral slip.

```
WHREFERRAL → REFERRAL DATE TOLINE ARLINE DESC
PALINE RETURNLINE PERSON JOBTITLE OFFICE
TOLINE → TO ORG
```
The intellectual element ACTREQ is “ACTION REQUESTED:”.

ACTION is “APPROPRIATE ACTION”, “DIRECT REPLY, FURNISH INFO COPY”, or “DIRECT REPLY W/COPY”.

DOI is “DESCRIPTION OF INCOMING”.

PROMPTACTION is “PROMPT ACTION IS ESSENTIAL – IF REQUIRED ACTION HAS NOT BEEN TAKEN WITHIN 9 WORKING DAYS OF RECEIPT, PLEASE TELEPHONE THE UNDERSIGNED AT”

RETURNTO is “RETURN CORRESPONDENCE, WORKSHEET AND COPY OF RESPONSE (OR DRAFT) TO:”

**Executive Orders**

An executive order is a directive from the President to federal administrative agencies. It has the force of law and does not need congressional approval to take effect. Executive orders are generally used to direct federal agencies and officials in their execution of congressionally established laws or policies. Congress can repeal or modify an executive order by passing a new law. However, it must be signed by the President or his veto overridden. Figure 14 shows an example of an Executive Order.  

The intellectual elements include subheads—headings that are run in at the beginning of a paragraph and followed by a period. In the grammar for executive orders, paragraphs with subheads are treated as simple paragraphs. The rules defining the documentary form of executive orders are shown below.

```
EO → EOID EMDASH EOTITLE EO_Authority_PARA BODY PERSON TWH CHRONDATE
    BODY → PARAS
    PARAS → PARAS PARA
    PARAS → PARA
```

---

The intellectual element EOID is “Executive Order” followed by a decimal number. EOTITLE is a TITLE after an EOID and EMDASH. EO_AUTHORITY_PARA is a paragraph beginning “By the authority vested in me as President by the Constitution and the laws of the United States of America” and ending “it is hereby ordered as follows:”. TWH is “The White House”.

Executive Order 12759 -- Implementation of Section 311(a) of the Comprehensive Anti-Apartheid Act

By the authority vested in me as President by the Constitution and the laws of the United States of America, including the Comprehensive Anti-Apartheid Act of 1986 (Public Law 99 - 440), as amended ("the Act"), and section 301 of title 3 of the United States Code, and having concluded that the Government of South Africa has taken all of the steps specified in section 311(a) of the Act, and, therefore, that title III and sections 501(c) and 504(b) of the Act have terminated, it is hereby ordered as follows:

Section 1. Implementation of Section 311(a) of the Act. All affected executive departments and agencies shall immediately take all steps necessary, consistent with the Constitution, to implement the termination of those sanctions which were imposed by title III and sections 501(c) and 504(b) of the Act.

Section 2. Status of Prior Executive Order. Except as superseded by section 1 of this order, Executive Order No. 12571 of October 27, 1986, "Implementation of the Comprehensive Anti-Apartheid Act," shall remain in effect. Pursuant to this order, the Inter-Agency Coordinating Committee established by section 12 of that order shall monitor the termination of those sanctions which were imposed by title III and sections 501(c) and 504(b) of the Act.

Section 3. Actions Taken and Proceedings Pending. This order shall not affect any action taken or proceeding pending and not finally concluded or determined on the effective date of this order, or any action or proceeding based on any act committed prior to the effective date of this order, or any rights and duties that matured or penalties that were incurred prior to the effective date of this order.

Section 4. Revocation. Executive Order No. 12332 of September 9, 1985, and Executive Order No. 12535 of October 1, 1985, which lapsed on September 9, 1987, pursuant to the provisions of sections 1622(d) and 1701 of title 50 of the United States Code are hereby revoked.

Section 5. Effective Date. This order shall be effective immediately.

George Bush

The White House,


Figure 14. An Example of an Executive Order
**National Security Directives**

On January 20, 1989, President Bush issued NSD-1 providing the charter for National Security Council (NSC) administration. NSC policy papers were named National Security Review papers (NSRs) and National Security Directives (NSDs) to distinguish them from the Reagan era documentation. The National Security Directives series was used by the Bush administration to promulgate presidential decisions relating to national security. Figure 15 shows an example of a National Security Directive.

---

**Figure 15. An Example of a National Security Directive**

---


12 Bush Presidential Library [http://bushlibrary.tamu.edu/research/nsd.php](http://bushlibrary.tamu.edu/research/nsd.php)
The following grammar rules define the form of a National Security Directive.

\[
\begin{align*}
\text{NATLSECDIR} & \rightarrow \text{CLASSIFICATION CHRONDATE NSD ADDRLINE} \\
\text{SUBJLINE_CLASS} & \rightarrow \text{BODY_CLASS DECLASSIFY_ON_LINE} \\
\text{ADDRLINE} & \rightarrow \text{FOR ENTITIES} \\
\text{SUBJLINE_CLASS} & \rightarrow \text{SUBJ TOPIC CLASSABBR} \\
\text{BODY_CLASS} & \rightarrow \text{PARAS_CLASS} \\
\text{ENTITIES} & \rightarrow \text{ENTITIES ENTITY} \\
\text{ENTITY} & \rightarrow \text{PERSON} \\
\text{ENTITY} & \rightarrow \text{PERSON ORGANIZATION} \\
\text{PARAS_CLASS} & \rightarrow \text{PARAS_CLASS PARA_CLASS} \\
\text{PARAS_CLASS} & \rightarrow \text{PARA_CLASS}
\end{align*}
\]

The intellectual element NSD is “NATIONAL SECURITY DIRECTIVE” followed by a decimal number. FOR is “MEMORANDUM FOR:”. CLASSABBR is (U), (C), (S), and (TS). CLASSIFICATION is UNCLASSIFIED, CONFIDENTIAL, SECRET, and TOP SECRET. “THE” followed by a job title is recognized as PERSON, a person’s name. SUBJ is “SUBJECT:”. PARA_CLASS is a paragraph that has at the end a security classification abbreviation enclosed in parentheses.

**National Security Review**

The National Security Review papers were used during the presidential administration of George H. W. Bush. This name was used in order to distinguish them from National Security Studies conducted by the Reagan administration. Figure 16 shows an example of a National Security Review.\(^ {13} \)

The grammar rules below define the form of a National Security Review.

\[
\begin{align*}
\text{NATLSECREV} & \rightarrow \text{CLASSIFICATION CHRONDATE NSR ADDRLINE} \\
\text{SUBJLINE_CLASS} & \rightarrow \text{BODY_CLASS DECLASSIFY_ON_LINE} \\
\text{ADDRLINE} & \rightarrow \text{FOR ENTITIES} \\
\text{SUBJLINE_CLASS} & \rightarrow \text{SUBJ TOPIC CLASSABBR} \\
\text{BODY_CLASS} & \rightarrow \text{PARAS_CLASS} \\
\text{ENTITIES} & \rightarrow \text{ENTITIES ENTITY} \\
\text{ENTITY} & \rightarrow \text{PERSON} \\
\text{ENTITY} & \rightarrow \text{PERSON ORGANIZATION} \\
\text{PARAS_CLASS} & \rightarrow \text{PARAS_CLASS PARA_CLASS} \\
\text{PARAS_CLASS} & \rightarrow \text{PARA_CLASS}
\end{align*}
\]

NSR is “National Security Review” followed by a hyphen followed by a decimal number. DECLASSIFY_ON_LINE is “Declassify on: OADR”.

\(^ {13} \) Bush Presidential Library [http://bushlibrary.tamu.edu/research/nsr.php](http://bushlibrary.tamu.edu/research/nsr.php)
**Presidential Determination**

A *Presidential Determination* is a report requested of the President by Congress concerning any finding or determination under any provision of the Foreign Assistance Act for each fiscal year. All Presidential Determinations are addressed to the Secretary of State, are released to the public in a White House Press Release, and published in the Federal Register under Presidential Documents. A *Disaster Declaration* is a Presidential
determination that a jurisdiction of the United States may receive a federal aid as a result of damage from a major disaster or emergency. Figure 17 shows an example of a Presidential Determination.\textsuperscript{14}

\begin{table}[h]
\centering
\begin{tabular}{|l|}
\hline
Presidential Determination No. 89 – 13 -- Arms Sales to Saudi Arabia \\
April 12, 1989 \\
\hline
\end{tabular}
\end{table}

\begin{table}[h]
\centering
\begin{tabular}{|l|}
\hline
Memorandum for the Secretary of State \\
\hline
Subject: Certification with Respect to Section 1307 of the National Defense 
Authorization Act, Fiscal Year 1989 \\
\hline
Pursuant to Section 1307 of the National Defense Authorization Act, Fiscal Year 
1989, I hereby certify that Saudi Arabia does not possess biological, chemical, 
or nuclear weaponized for the intermediate-range ballistic missiles purchased from 
the People's Republic of China. \\
You are hereby authorized and directed to publish this certification in the 
Federal Register. \\
\hline
George Bush \\
\end{tabular}
\end{table}

\textbf{Figure 17. A Presidential Determination}

The following rules define the form of a Presidential Determination.

\begin{verbatim}
PRESDETERM → PDNO EMDASH TITLE DATE ADDRLINE SUBJLINE PARAS PERSON 
ADDRLINE → FOR ENTITY 
SUBJLINE → SUBJ TOPIC 
PARAS → PARAS PARA 
PARA → PARA
\end{verbatim}

PDNO is “Presidential Determination No.” followed by two digits indicating year, a hyphen and a decimal number.

\textbf{Presidential Proclamation}

A \textit{presidential proclamation} is "a declaration issued under specific authority granted to the President by Congress. Generally, it relates to matters of widespread interest. Some proclamations have no legal effect but merely are appeals to the public, e.g., the observance of American Education Week."\textsuperscript{15}

Figure 18 shows an example of the initial part of a Presidential Proclamation.\textsuperscript{16}

\begin{verbatim}
PRESDETERM → PDNO EMDASH TITLE DATE ADDRLINE SUBJLINE PARAS PERSON 
ADDRLINE → FOR ENTITY 
SUBJLINE → SUBJ TOPIC 
PARAS → PARAS PARA 
PARA → PARA
\end{verbatim}

\textsuperscript{14} Bush Public Papers. April 12, 1989. \url{http://bushlibrary.tamu.edu/research/public_papers.php}
\textsuperscript{15} Westlaw, Glossary of Terms used in Legal research. \url{http://lawschool.westlaw.com/shared/marketInfoDisplay.asp?code=MI&id=155&mainpage=17}
Figure 18. Example of a Presidential Proclamation.

The following grammar rules define the form of a Presidential Proclamation.

```
PRESPROC → TITLE BYPRESIDENT PROCLAMATION PARAS
    FORMULA_OF_PROCLAMATION PERSON TWH DATE
PARAS → PARAS PARA
PARAS → PARA
FORMULA-OF-PROCLAMATION → IN_WITNESS ORDINAL DAYOF MONTH
    INYEAR CARDINAL, INDEPENDENCE ORDINAL
```

"Proclamations issued by the President shall conclude with the following-described recitation-- IN WITNESS WHEREOF, I have hereunto set my hand this ... day of ..... .., in the year of our Lord ............... , and of the Independence of the United States of America the ............. ."17

17 Executive Order 11030--Preparation, presentation, filing, and publication of Executive orders and proclamations (June 19, 1962)
Thus, the intellectual element IN_WITNESS is “IN WITNESS WHEREOF, I have hereunto set my hand this”. DAYOF is “day of”. INYEAR is “in the year of our Lord”. INDEPENDENCE is “and of the Independence of the United States of America the”.

BYTHEPRESIDENT is “BY THE PRESIDENT OF THE UNITED STATES”. PROCLAMATION is “A PROCLAMATION”.

**White House Fact Sheet**

“A fact sheet is a presentation of data on any subject in a format emphasizing brevity, key points of interest or concern, a fairly minimalist design aesthetic, and a general desire to convey the most relevant information in the least amount of space.”\(^{18}\) For instance, the CIA World Fact Book is a collection of factsheets on the countries of the world. The initial portion of a White House Fact Sheet is shown in Fig. 19.\(^{19}\)

The following rules define the form of a White House Fact Sheet.

\[
\text{WHFACTSHEET} \rightarrow \text{WH_FACTSHEET_TITLE DATE BODY} \\
\text{BODY} \rightarrow \text{PARAS} \\
\text{PARAS} \rightarrow \text{PARAS PARA} \\
\text{PARAS} \rightarrow \text{PARA}
\]

The WH_FACTSHEET_TITLE is an intellectual element that is a title and whose first four words are “White House Fact Sheet”.

---


\(^{19}\) George Bush Presidential Library, Bush Public papers. [http://bushlibrary.tamu.edu/research/public_papers.php](http://bushlibrary.tamu.edu/research/public_papers.php)
Figure 19. A White House Fact Sheet

**Transcript of Presidential News Conference**

*A presidential press conference* (news conference) is a conference at which press and TV reporters ask questions of the President. A *transcript of Presidential News Conference* is
a written copy of reporter’s questions and the President’s answers at a press conference. When the U.S. President holds a press conference, he takes questions from the press pool in a specific order: wire services, broadcast networks, national newspapers, newsmagazines, radio and regional newspapers.

Several transcripts of press conferences were examined. It was found that that they are usually titled as to who is answering questions from the Press, e.g., The President's News Conference, The Press Secretary's News Conference, The President's and Prime Minister's News Conference. This is followed by a date. Then there are opening statements by whoever is answering questions. This is followed is by a question-answer dialog. Important intellectual elements of this dialog are the indication of the speakers (The President, The Prime Minister, The Press Secretary) and questions indicated by the abbreviation Q. These elements appear at the beginning of paragraphs. They need to be recognized as intellectual elements apart from the content of the paragraphs. The transcript of the Presidents Press Conference has topics inserted before questions and answers on the same topic.

Figure 20 shows an example of the transcript of a Presidential News Conference. The following grammar rules define the form of a transcript of a Presidential News Conference.

TRANSPRESNEWSCONF → PRESNEWSCONF DATE OPENING DIALOG
OPENING → THEPRESIDENT PARAS
DIALOG → SECT DIALOG
DIALOG → SECT
SECT → TOPICHDG QANDAS
QANDAS → QANDA QANDAS
QANDAS → QANDA
QANDA → QUESTION ANSWER
QUESTION → QUES PARAS
ANSWER → THEPRESIDENT PARAS

The intellectual element PRESNEWSCONF is “The President’s News Conference.” QUES is “Q.” THEPRESIDENT is “The President.”

---

20 Bush Public Papers, http://bushlibrary.tamu.edu/research/paper.html
The President's News Conference
January 27, 1989

The President. Harmony and peace in here. [Laughter] First, I sound worse than I feel. Let me -- --

Q. What's wrong with you?

The President. Slight cold. But it's been a full week since the inauguration. I just wanted to stop by under our new policy, give you an update.

I've been talking this week about ethics and the emphasis is not, believe me, a fad or some passing fancy. We're going to be hearing more about it -- I think a lot more. In broader terms, I'm trying to set high standards for government service: duty, honor, personal sacrifice for the common good. And I want to assemble a government that the people of this nation can be proud of. That's our goal; that's our mission. We did appoint this week, as you know, a new commission on ethics. We've got two able Chairmen: Judge Wilkey and [Attorney] General Griffin Bell. That commission started a fresh, constructive dialog with Congress on both sides of the aisle. I'm pleased with the way those initial meetings have gone. So, we're ready to roll.

I think it's been officially announced that we're going to -- certainly to Tokyo, then on to China, and I'm looking forward to it very much. And then, also, we'll be stopping for a relatively brief stop in Korea on the way back from China. We've had other invitations. That's about all we can do. And in a nostalgic basis, needless to say, Barbara and I are looking forward in a very personal way to going back to Beijing.

Now, with no further ado, I'll be glad to take some questions. Terry [Terence Hunt, Associated Press]?

Federal Pay Raise

Q. Mr. President, you said before you became President that you supported Ronald Reagan's decision on a 50-percent pay raise for Members of Congress and that you'd have something more to say about it once you'd get to the Oval Office. Most of the polls show that a vast majority of Americans oppose that pay raise.

The President. Yes.

Q. Now you're President, what's your view on that pay increase?

The President. Well, I did say I supported it, and I do support it. A raise is overdue. There's no question about that. There are some good things in this. The return of the honorarium, payments for these speeches, I think, is very, very good. I think it's good government. It's a complicated formulation. There are some elements that bother people, including me. But the President did consider all this carefully. He went forward with the

Figure 20. A Transcript of a Presidential News Conference
White House Press Releases

A White House Press Release is a statement prepared by the White House for distribution to the media for the purpose of letting the public know of White House activities. Fig. 21 shows an example of a White House Press Release.  

![White House Press Release Example]

The following grammar rules define the form of a White House Press Release.

```
WHPRESSRELEASE --> PRHDG HEADLINE BODY ENDS
WHPRESSRELEASE --> PRHDG BODY ENDS
PRHDG --> THE_WHITE_HOUSE OFFICE_PRESS_SECY ULINE FIR DOCDATE
BODY --> PARAS
PARAS --> PARAS PARA
PARAS --> PARA
```

FIR is “For Immediate Release”. ENDS is “# # #”, a journalistic standard for ending a press release.

Press Pool Report

“Press pool refers to a group of news gathering organizations pooling their resources in the collection of news. A pool feed is then distributed to members of the broadcast pool.

---

who are free to edit it or use it as they see fit. In the case of print reporters, a written pool report is distributed to all members.”  

“A daily rotating pool is set up at the White House for organizations that cover the President. They are allowed special access to certain events with the understanding that the information and footage will then be shared with the rest of the White House press corps. For instance, when the president meets with a foreign dignitary, the two will sometimes make a statement together in the Oval Office. Since it is not possible to accommodate all interested journalists in that room, pool coverage is the only practical way to handle the event.”

A sample Press Pool Report is shown in Fig. 22. The following grammar rules define the documentary form of a Press Pool Report.

```
POOLREPORT → PREPORT TITLE DATE PARAS REPORTERS
PARAS → PARA PARA
PARAS → PARA
REPORTERS → REPORTERS REPORTER
REPORTERS → REPORTER
REPORTER → PERSON, ORG
```

The intellectual element PREPORT is “Pool Report”, which may or may not be followed by a number. ORG is the name of a news media organization.

---

22 http://en.wikipedia.org/wiki/Press_pool
23 ibid
Pool Report #35

Air Force One: Pease to Andrews

8/28/90

Shortly before takeoff, Marlin informed us that the President was wounded in action this morning while casting for bluefish aboard "Fidelity." President Bush took a bluefish lure in the right ear as fishing partner and son Jono, a southpaw, was attempting to cast.

One of the hooks on the lure imbedded in the fleshy part of the ear just above the earlobe. Dr. Michael Nash, bobbing nearby in an accompanying boat, motored over and clipped the barb off the hook, pulling it through the other side of the President’s ear. The President continued fishing for another hour and boated three bluefish. Says Marlin, "The President was strong and courageous."

Back on shore <deleted Dr. Lawrence Mohr> <Inserted Nurse Paula Trivette> cleaned the puncture wound and gave the President a tetanus shot. Marlin assured your pool that the President’s phoning of world leaders will not be impaired, because he uses his left ear for the phone receiver. Indeed, your pool could see no bandage as the President boarded Air Force One.

Notwithstanding the GQ-style pose in which the President is shown in today’s newspapers, he has no plans to adorn his new pierced ear with a stud or earring.

Other matters: Marlin had nothing to say about the American hostage reported dead of a heart attack in Iraq. He said Scowcroft informed the President of the reported death, but he did not know when. Marlin also had nothing to say about reports of further interdictions of ships suspected of heading to or from Iraq. Nor did Marlin have any details on the President’s schedule for Wednesday and Thursday. No one else came back.

Ellen Warren, Knight-Ridder
Dan Goodgame, Time Magazine

Figure 22. A Press Pool Report

Resumes

There is much variation in the forms of resumes that appear in the Bush Presidential e-records. Grammatical induction on a large sample of resumes is probably the best way to characterize this documentary form and its subtypes. Figure 23 shows an example of a resume of a famous (fictitious) Georgia Tech alum.
George P. Burdel
190 North Ave.,
Atlanta, Georgia 30313.
(404) 555-0538

OBJECTIVE

Seeking a position in mentoring current Georgia Tech students where over 70 years of experience will contribute to students educational and social experience at Georgia Tech.

EDUCATION

B.S. 1980 Ceramic Engineering Georgia Institute of Technology Atlanta, Georgia

M.E. 1962 Mechanical Engineering Georgia Institute of Technology Atlanta, Georgia

EMPLOYMENT

Board of Directors, Mad Magazine 1989–1991

Member, House of Uncommons, United States of Roadkills-R-Us, Austin, Texas, 1997–1998

Director, Center for Alumni Long Life Learning, 1999–2000

REFERENCES

References on Request

Figure 23. An Example of a Resume

The following grammar rules define the documentary form of the resume in Fig. 23. The grammar characterizes a limited subset of the large variety of documentary forms of resumes. However, it may aid in identifying some of the intellectual elements of resumes.

RESUME → NAMESEC OBJSEC EDUSEC EMPLOYSEC REFSEC
NAMESEC → PERSON ADDRESS TELEPHONE
OBJSEC → OBJ PARA
EDUSEC → EDU DEGREES
DEGREES → DEGREE DEGREES
DEGREES → DEGREE
DEGREE → DEGREEABBREYR MAJOR EDUORG CITY STATE
EMPLOYSEC → EMPLOY JOBLIST
JOBLIST → JOB JOBLIST
JOBLIST → JOB
JOB → JBOTTLE ORGANIZATION CITY STATE YEAR HYPHEN YEAR
REFSEC → REFER REFOPTION
REFOPTION → REFLIST
REFLIST → REF REFLIST
The intellectual element REFOPTION is “References Available Upon Request”. OBJ is “OBJECTIVE”. EDU is “EDUCATION”. EMPLOY is “EMPLOYMENT”. REFER is “REFERENCES”. EDUORG is the name of an educational institution.

**Schedules**

A *schedule* is an ordered list of times at which things are planned to occur. Schedules have a variety of documentary forms. Figure 24 shows one of the documentary forms of a schedule.25

| MRS. BUSH’S SUMMARY SCHEDULE |
| Sunday, February 28, 1988 |
| 11:00 a.m. -Depart residence en route National Airport. Casey Healy will accompany you. |
| 11:40 a.m. -Depart National Airport on board Eastern #139 en route Atlanta, Georgia. |
| 1:22 p.m. -Arrive Atlanta. |
| 2:00 p.m. -Arrive Omni Hotel and join the Vice President. (Private time:30 minutes) |
| **NOTE:** You will follow the Vice President’s remaining schedule. |
| 3:00 p.m. -Attend The Atlanta Journal and Constitution Presidential Debate (@GB). |
| 5:30 p.m. -Depart Atlanta, Georgia en route Andrews AFB (@GB). |
| 5:55 p.m. -Arrive Andrews AFB (@GB). |
| 7:10 p.m. -Arrive residence (@GB). |

Figure 24. An example of a Schedule

The following grammar rules define in part the form of a schedule.

The intellectual element SCHEDTITLE is a title ending in "SCHEDULE" or "Schedule." EVENT is a verb phrase.

### Mailing Lists

A mailing list is a collection of addresses used for sending material to multiple recipients. The traditional mailing list is a list of names and postal addresses where delivery is through the Postal System. Figure 25 shows part of a traditional mailing list.

```
Robert Frost
121 Historic Route 7A
Shaftsbury, VT 05262

Joel C. Harris
1050 Ralph David Abernathy Blvd. SW
Atlanta, GA 30310

Margaret Mitchell
990 Peachtree Street
Atlanta, GA 30309

Edgar A. Poe
203 Amity Street
Baltimore, MD 21201
```

**Figure 25. An excerpt from a Mailing List**

The following are rules for a list of mailing addresses. Address is a semantic category recognized by the Semantic Tagger.

```
MAILINGLIST → ADDRESSES
ADDRESSES → ENTITY ADDRESS ADDRESSES
ADDRESSES → ENTITY ADDRESS
ENTITY → PERSON
```