# Proper Name Extraction from Non-Journalistic Texts

*Thierry Poibeau*[*] *and Leila Kosseim*[**]
[*]*Laboratoire Central de Recherches, Thales/LCR, and*
*Laboratoire d'Informatique de Paris-Nord, Institut Galilée, Université Paris-Nord.*
*Thierry.Poibeau@lcr.thomson-csf.com*
[**]*RALI, Université de Montréal, kosseim@iro.umontreal.ca*

March 9, 2001

## Abstract

This paper discusses the influence of the corpus on the automatic identification of proper names in texts. Techniques developed for the newswire genre are generally not sufficient to deal with larger corpora containing texts that do not follow strict writing constraints (for example, e-mail messages, transcriptions of oral conversations, etc). After a brief review of the research performed on news texts, we present some of the problems involved in the analysis of two different corpora: e-mails and hand-transcribed telephone conversations. Once the sources of errors have been presented, we then describe an approach to adapt a proper name extraction system developed for newspaper texts to the analysis of e-mail messages.

**Key-words:** Proper Name Extraction, Corpus, Information Extraction

## 1 Introduction

The identification of proper nouns in written or oral documents is an important task in natural language processing. This type of expression holds an important place in many corpora (newspapers, corporate documents, e-mails ...). It is therefore important to be able to identify these expressions either for specific applications (eg. to index documents by proper names or to build mailing lists) or for general research purposes (eg. to improve the syntactic analysis of a text).

Many research projects have addressed the issue of proper names identification in newspaper texts; in particular, the Message Understanding Conferences (MUC) [1, 2, 3]. In these conferences, the first task to achieve is to identify named entities, i.e. proper names and also temporal and numerical expressions. This task is generally viewed as being *generic*, in the sense that all texts use such expressions and their identification seems a priori independent of the discourse domain or textual genre. However, the experiences performed within the MUC framework have all used homogeneous corpora constituted primarily of newspaper articles. This type of text respects strict writing guidelines which facilitates the identification task. Sequences like *Mr.* for *Mister* or *Ms* precedes proper names rather systematically. These strategies are however insufficient to analyse other types of texts such as electronic mail ou minutes from a meeting because writing guidelines are either different or are much less strict. However, with the explosion of documents in electronic format, it is precisely these types of documents that need to be processed automatically.

This paper tries to determine, through two experiments on non-journalistic corpora, the weaknesses of rule-based systems and the necessary modifications to these systems in order to achieve acceptable performance. After a brief overview of the literature on named entity extraction on newspaper texts, we evaluate the performances of some systems developed for the newspaper genre on 2 types of informal texts (e-mails and manual transcriptions of dialogues). We will then present the difficulties associated with these types of texts et propose strategies to adapt rule-based system on non-journalistic texts to maintain reasonable performances in non-journalistic texts. Finally, a typology of existing errors will be presented.

## 2 Previous Work

Influenced by the MUC conferences, work on named-entity extraction have traditionally been performed on news texts. This task tries to identify 3 types of expressions:

ENAMEX: Proper names, including names of persons, locations and organizations.

TIMEX: Temporal expressions such as dates and time.

NUMEX: Numerical expressions such as money and percentages.

In this work, we have concentrated our efforts on the first type of expressions: ENAMEX. Two main approaches are generally followed for their identification: a surface linguistic approach or a probabilistic approach. The probabilistic approach uses a language model trained on large pre-tagged corpora to learn patterns of identification [10]. The IdentiFinder system [17, 18], for example, uses such an approach. Studies have shown that this type of method yields good results if the training corpora are large enough. The Hub evaluation series on speech recognition includes a named-entity extraction task from automatic transcripts of news bulletins [14, 15]. These transcripts, generated from speech recognition systems, contain properties that render extraction difficult: the texts are in one case, they lack punctuation marks and the word-error rate is not insignificant. For these reasons, most systems that work on transcripts of oral adopt a probabilistic approach[1].

The linguistic approach is based on a syntactic and lexical description of the expressions that are seeked. Here, the text is tokenised and tagged with grammatical tags. A full syntactic analysis of the sentences is usually not performed as it is both an expensive and not necessary task; only chunking is usually performed. The linguistic approach typically uses several resources:

1. Lists of trigger words – eg. *Mr* for *Mister* or *inc.* for *incorporated*

[1] This is why, most research in this area are dedicated no so much on linguistic aspects but on voice recognition aspects such as the effect of word error rate on entity extraction [18], the use of prosody to increase recognition scores [12] or the effect of the size of the training corpus [17].

2. Gazetteers – large dictionaries of known proper names
3. Dictionaries of the general language, essentially to identify unknown words

Grammar rules are then applied to combine these informations to tag the expressions that are identified with the most appropriate semantic tag. Alembic [4], Proteus [11], and TextPro [7] (a descendant of Fastus [6, 5]) are examples of systems that use this approach.

This paper will only analyse rule-based systems. Each author had developed their own rule-based named-entity extractor (Exibum and Lexis) and wanted to see how these systems performed on texts for with they were not developed (ie. non journalistic texts). Exibum [16] is a system developed as part of an bilingual (English-French) information extraction system; while Lexis [21] was developed as part of a technology watch system.

Regardless of the approach used, named-entity extraction from written documents is currently the most successful task in information extraction. Combined scores of precision and recall are comparable to human scores (in the order of 0.9 P&R[2] on news texts).

The high performances obtained with written documents from newspaper genre demonstrates that the technology is ripe to attract commercial attention, to serve as basis to higher-level NLP tools or to be tested on other types of texts.

## 3 The use of information extraction system on non-journalistic corpora

The recognition of named entities from journalistic corpora is a task in which systems achieve good performances. However, other types of text exhibit different characteristics. Companies as well as individuals are facing a huge amount of electronic texts like e-mails, news messages and so on. The texts do not follow strict redactional constraints: they use a vocabulary and a syntax that is variable and relaxed compared to journalistic texts. This idea has been validated through two experiments over informal corpora.

### 3.1 Description of the corpora

Two corpora were used for the experiments. We shall call these corpora: the Valcartier and the Communication corpora. The Valcartier corpus is made of manual transcriptions of telephone conversations in English provided by the Search and Rescue Division of the Canadian Armed Forces. This corpus will be used in the future to develop an Information Extraction application[3]. Even if these transcrip-

[2] Precision measures the ratio of correct answer over all answers given by the system. Recall measures the ration of correct answers given by the system over all correct answers. The F-score combines precision and recall into one single measure using this formula: $F = \frac{(\beta^2+1.0)\times P\times R}{\beta^2\times P+R}$. When $\beta = 1$, precision and recall have the same relative importance and the F-score is called P&R.

[3] Typically, the dialogs involve controller from a coordination center who is performing an inquiry on the disappearance of a person or an airplane. The controller discusses by telephone with an investigator on location or anyone able to help in the inquiry.

tions are made from oral conversations, their quality is very closed to that of written texts. These transcriptions contain letter in mixed cases, contain very few word recognition errors[4]. The remaining errors are mainly typos or homonymic confusions and contain punctuation marks added by the transcription agents from the intonation and the silences of interlocutors. Due to these main features, we have a very accurate transcription from oral conversations and we are able to focus on their content. The Valcartier corpus contains about 25 000 tokens and 2 200 token types.

The Communication corpus is made of e-mails in English. This corpus has been established by experts in the field of technological survey, who had to elaborate a report on telecommunications. This kind of corpus is often made of heterogeneous pieces: technical documents, product announcements, messages from concerned newsgroups, e-mails. The recognition of named entities, especially person and company names, is a major added value for analysts facing these texts. Corpus processing can be boosted by such techniques: decisions on document relevance go faster and experts can focus on the analysis of the sole relevant documents. For formatting reasons, we chose to only study the part of the corpus made of electronic mails (technical documents are often in the PDF format and thus, are not directly available). The corpus contains letter in mixed cases, is written in an informal manner: sentences can be incomplete and written in telegraphic style. We will subsequently precise the notion of informal style with the description of a grammar made of construction variations. Finally, the number of typos in the corpus is limited (architecture for architecture) compared to other experiments on electronic informal corpora.

The corpus is made of 300 000 tokens, that approximately correspond to 50 000 token types. The reference corpus, which is distinct from the training corpus is made of 85 000 tokens that correspond to 12 000 types. The reference was established by a human annotator and corrected by an expert in the fields of telecommunications.

### 3.2 A drop in performances

The use of analysis principles developed for journalistic texts over other kind of texts without any change leads to an important decrease of performances. Systems analyzing correctly about 90% of the sequences from a journalistic corpus can sustain a decrease of performance up to 50% on more informal texts. Journalistic redactional constraints often introduce person names with titles (*President Chirac*) or trigger words (*Mister Chirac*) [22]. This way of writing is not systematic in informal texts. Performances largely decrease if one analyses various texts with a too normative grammar. Incomplete sentences and telegraphic style, very frequent in informal texts, hinder syntactic analysis and an accurate tagging of proper names. This fact has been established independently by the two authors in the framework of two different applications. Exibum is using a linguistic rule-based approach to identify proper names in the texts. A first experiment was made

[4] Contrary to texts coming from automatic transcriptions that contain a 30 to 40% error rate.

to evaluate the results of Exibum on the recognition of named entities on the Valcartier corpus. The performances the system obtained went from 0.69 P&R on the MUC-6 corpus to 0.44 on the Valcartier corpus. The first experiments with Lexis concerned texts from journalistic corpora with performances included between 0.50 and 0.70 in function of the text. Once this decrease of performance was established, it appeared that we had to identify its real cause by developing a precise and rigorous evaluation using systems that have already been tested in larger evaluation campaigns.

### 3.3 Validation of the initial results

To verify that the poor performances achieved by our system on various corpora were not due to the systems themselves, the evaluation has then been enlarged to two systems having participated to the MUC conferences. We analyzed the results from the Alembic [4] and TextPro systems [7] which were publicly available. The Valcartier corpus has been given as input to these two systems, and the results have been evaluated using the MUC methodology. Finally we classified the extraction errors to try to identify characteristic features belonging specifically to the change of domain. Alembic [4], developed at Mitre Corporation, is one of the pioneer systems in information extraction. It was initially developed to participate to the MUC-4 conference in 1992 and has regularly participated to subsequent competitions, taking advantage of several improvements. TextPro from SRI [7] takes its origin in the Fastus system [6] that is also a pioneer in information extraction. TextPro is a light version of Fastus, developed for the Hub-4 conference [15]. Alembic and TextPro were among the highest performing systems at the MUC conferences. Three kinds of proper names were evaluated: person names, location names and organization names. Two human annotators independently developed the key templates[5]. Disagreements between annotators were solved by joint decision. The two human annotators evaluated with the MUC protocol achieved 0.97 and 0.96 P&R. Table 2 gives an illustration of the results from the two human annotators and the three systems over the corpus. Two different measures are given for Alembic because two specific words were systematically wrongly tagged. These were being especially frequent in our two corpora; therefore the system was unfairly disadvantaged. To obtain a measure giving a more accurate image of the results, we give a first measure on the original corpus and a second one that does not take into account these two specific words.

The results in table 2 clearly show that human annotators do not seem to be influenced by the change of corpus, while automatic system obtain lower results and cannot compete with human annotator results. It is then interesting to study the reason why these systems are not consistent. Is it a problem with the dictionaries that are not tuned to the discourse domain or the informal features appearing in the

[5] One of these two annotators is one of the authors, but the other one did not take part in the experiment.

| System | P&R MUC-6 | P&R Valcartier | P&R Communication |
|---|---|---|---|
| Human annotators | 0.97[6] | 0.97 | 0.90[7] |
| Alembic[8](Mitre) | 0.86 | 0.50 - 0.57 | -[9] |
| TextPro (SRI) | 0.86 | 0.41 | - |
| Exibum | 0.69 | 0.44 | - |
| Lexis | 0.90 | - | 0.50 |

Table 1: Extraction of proper names without adaptation of systems

syntax of the sequences that have to be recognized.

### 3.4 A grammar made of variants

The variable syntax of proper nouns is responsible for most cases of silence (i.e. non detected proper nouns). In journalistic texts, person names are generally preceded by titles and trigger words (*Mr., Mrs*) whereas it is rarely the case, in the two corpora that we are studying here,. Proper nouns and especially person names belong to an open class: titles are very efficient indicators and this is the reason why the different systems achieve good performance on journalistic texts.

The person name grammar is then original and not stable but depends on the corpus. The rules that have to be applied over informal texts are sometimes not the same as the ones dedicated to journalistic texts. A rule applying very frequently in a journalistic text will be very rare in a corpus made of electronic mails, and reciprocally for some other rules. Let us see that, even inside a unique newspaper, editorial constraints do not apply uniformly. A person name can be at first introduced by a title (*Prime Minister Edouard Balladur*) and then, in a simpler way, just introduced by a trigger word (*Mr Balladur*). Some specific sections like the society or art sections can name a person by his or her name, without any trigger word.

The grammar designed to recognize organization and company names in informal texts must include more informal ways of naming entities than the one dedicated to pure journalistic texts. Trigger words like *inc.* or *ltd* are not mentioned most of the time. For example, the name of the organization *Transportation Safety Board* includes the trigger word *Board*, which denotes that the preceding sequence designates an organization. In the Valcartier corpus, this organization is

[6] Extraction of all named entities

[7] The MUC-6 score is the official one; while the score with the Valcartier corpus has been calculated using the public version of Alembic Workbench 4.12 (URL: http://www.mitre.org/resources/centers/it/g063/workbench.html)

[8] This score is an estimation. For the Communication corpus, there was one annotator and a manual validation by an expert in the domain. The contrast between the expert and the non expert explains the good quality of the final result for an audio transcription (.97). It also shows that knowledge of the domain is necessary to accurately tag the text.

[9] Alembic performed strangely bad on the Communication corpus.

often named *Transportation Safety* without any trigger word. As in the case of person names, reduced organization names are frequently not recognized by the different systems. Lastly, location names are also identified by means of keywords (for example the preposition in or the words *lake* or *city*). The omission of trigger words or of the preposition before the location name frequently causes errors (silence or wrong categorization of the sequence). It is especially the case when there is no trigger word nor any word allowing to accurately tag the sequence: the system must then deal with previously unknown names. The only operational techniques in this framework are to dynamically type unknown entities by a local analysis of the context of the entity, by means of a cooccurrence analysis.

## 4 Towards adaptive systems

Given the different possible errors examined when we were looking at the results of the named entity recognition over different corpora, a set of strategies has been defined and evaluated on the Communication corpus.

The fact that expressions are introduced without any marker leads to many isolated unknown words. To solve this problem, it is necessary to improve the coverage of the dictionaries and to add dynamic resource acquisition process to the original system.

### 4.1 Improve dictionary coverage

The person name recognition task is achieved by Lexis with a success rate of 0.90 P&R on the MUC-6 corpus, but only 0.50 on the Communication corpus (see section 3.1). Whatever is the corpus on which was tested the system, the grammar remains relatively stable and the sequence `First_name Name` is generally the most frequent one. The variant of this structure consists in an isolated person name sometimes introduced by initials of the first name (one letter or two, generally in upper case, and followed by a dot or a space). These sequences introduced by a trigger word are, unfortunately, very rare in the electronic mails of the Communication corpus. A very frequent rule in the Herald Tribune can be very rare in an electronic mail corpus, and reciprocally for other rules. It is necessary to establish very complete lists of person names. Lexis system has currently registered over 24.000 person names. To improve the performance of the system, it is essential to have a good coverage of the finite class of proper names of the concerned language, especially first names and toponyms, even if this class cannot rigorously be described as a finite class. In parallel, unknown words feed in a significant way the dictionaries of proper names. Thus, it is possible, on a corpus which was not journalistic but technical, to reach quickly a coverage of about 0.60 P&R for person names on the Communication corpus, only by the addition to the dictionary of some of the previously unknown words for the analyzer. The performance is comparable for the recognition of location names that crucially requires exhaustive geographical denomination lists that have to be acquired from existing resources or from training corpus.

### 4.2 Dynamically recognizing new entities by machine learning techniques

A limit of the Lexis system that has been presented is the fact that it does not include any dynamic process to automatically adapt a part of its resources and rules to the corpus. This point is particularly significant for the analysis of texts like electronic mails, which are made of a significant number of person names appearing without being introduced by any trigger word. Part of these names can however be correctly analyzed if the system can find elsewhere a discriminative context making it possible to correctly identify the named entity. We propose a learning method that uses the previously found elements and the recognition rules of the proper names grammar to extend the coverage of the initial system. It is then a case of EBL, explanation based learning [19].

The mechanism is based on the registration of the grammatical rules that have been applied with success to tag previously unknown words. For example, the grammar can recognize the sequence *Mr. Kassianov* as being a person name even if *Kassianov* is an unknown word. The isolated occurrences of this word can consequently be tagged as person name. The machine learning process can be seen as an inductive mechanism using the knowledge of the system (grammatical rules) and the entities beforehand found (the positive set of examples) to improve the overall performances (the global gain in performance is about 10 to 15% in function of text, that is to say 0.66 to 0.70 P&R).

### 4.3 Using discourse structures

Discourse structures are another source for knowledge acquisition. In the terminological field, [9] showed that new terms could be extracted from the analysis of particular sequences of texts. The same principle can be used for the automatic acquisition of new entities. We are particularly interested in enumeration that can be easily localized by the presence of person names, separated by connectors (commas, subordinating conjunction, etc). For example, in the following sequence:

<PERSON_NAME> Kassianov </PERSON_NAME> **,**
<UNKNOWN> Kostine </UNKNOWN> **and**
<PERSON_NAME> Primakov </PERSON_NAME>

*Kostine* is tagged as an unknown word. The system can infer from the context (the word *Kostine* appears in an enumeration of person names) that the word *Kostine* refers to a person name, even if this person name is isolated and could not be accurately tagged by a gazetteer lookup or from other occurrences in the text. Thanks to this strategy, the score of Lexis on the Communication corpus reached 0.84 P&R.

### 4.4 Resolving tagging introduced by the tagging strategy

The learning process can lead to conflicts between two types of entity, especially when dynamic typing made it possible to assign a tag to a word that is in contradiction with the tag contained in the dictionary or identified by another dynamic strategy. It is the case, for example, when a word registered as a location name in the dictionary is used as person name in a non ambiguous text sequence. Must isolated occurrences by tagged with the tag dynamically identified by the context analysis (person name) or by the tag previously stored in the dictionary (location name)? Let us consider the following sequence from a text from the MUC-6 corpus:

```
@  Washington, an Exchange Ally, Seems
@  To Be Strong Candidate to Head SEC
@  ----
<SO> WALL STREET JOURNAL (J), PAGE A2 </SO>
<DATELINE> WASHINGTON </DATELINE>
<TXT>
<p>
   Consuela Washington, a longtime House staffer and
an expert in securities laws, is a leading candidate to be
chairwoman of the Securities and Exchange Commission in the
Clinton administration.
</p>
```

It is clear that in this text *Consuela Washington* corresponds to a person name. The first occurrence of the word *Washington* is more problematic, insofar the only information in the sentence that makes it possible to disambiguate the sequence necessitates some knowledge on the world, namely that it is generally a person who manages an organization. But it would be necessary not to attach too much importance to that kind of hypotheses because a metaphorical use of the word cannot be excluded, as in the sentence *France wants to continue to manage IMF*. In fact, a proper analysis of pronoun references allows us to perfectly disambiguate the text. An automatic system has very few chances to properly analyze the text, especially if we take into account the fact that a completely isolated occurrence of *Washington* must be analyzed as a location name, between two other occurrences where *Washington* stands for a person name (the reader infers from the context that *Washington* is certainly the place where is located the journalist who emitted the news).

To circumscribe this kind of problem and to avoid propagation of errors[10], we propose to limit this dynamic tagging process to isolated text, not to an entire corpus. For example, in the previous text, the system will tag all isolated occurrences of *Washington* as person name, but in a subsequent text, if an isolated occurrence of the word *Washington* appears, the system will tag it as location name, according to the dictionary. When more than one tag is found by the dynamic process from the same text, an arbitrary choice is then carried out.

The end-user is also free to choose the discourse and linguistic structures he wants to use during the acquisition process. Indeed, it appears that this choice

[10] That is to say, when a word received from the context a tag which is in conflict with a tag previously recorded in the dictionary. It is the case of *Washington* in the above example.

largely depends on the corpus to be analyzed. For example, the structure:

<PERSON_NAME>X</PERSON_NAME> (<PERSON_NAME>Y</PERSON_NAME>)

is very frequently used in the cinema sections of newspapers to designate, between brackets, actors playing a role (in the above example, X and Y indicate variables; the first occurrence of <PERSON_NAME> often corresponds to a first name). But things will be different in other sections, where we will find, for example, the following structures:

<PERSON_NAME>X</PERSON_NAME> (<POLITIC_ORG>Y</POLITIC_ORG>)
<PERSON_NAME>X</PERSON_NAME> (<COUNTRY>Y</COUNTRY>).

In these cases, even if the system can solve certain ambiguities, contextual rules can introduce too much noise. It is then the end-user who has to choose to activate or not such rules, according to the expected performances and to his own expertise in the field.

According to the training corpus used (the MUC-6 corpus and the Communication corpus for English, The newspaper *Le Monde* or the AFP newswire for French), in spite of errors introduced by the learning mechanisms (extension of an incorrect tag over the text), the profit remains always positive (P&R measure). We estimated a 12% gain for recall that compensates a 3% precision decrease on the Communication corpus[11].

## 5 Analysis of the remaining errors

Let us now examine the remaining extraction errors. These errors can be divided into 2 classes:

**Unsolved errors:** Errors that should have been taken into account with the above mechanisms.

**Unadressed errors:** Errors that were not taken into account. Whether they arise from a general problem such as spelling mistakes or they arise from the specificity of a particular corpus such as those found in the Valcartier corpus.

### 5.1 Unsolved Errors

The three strategies presented earlier have allowed the Lexis system to go from a score of 0.50 to 0.84 P&R on the Communication corpus. Although the improvement is significant, these performances are still inferior to the average MUC-6 performances. Among the unsolved errors, we have identified:

[11] In this experiment, the system was just tagging isolated unknown words from the knowledge acquired during the first pass. No discourse structures nor enumeration were used to tag unknown entities.

**Incompletion of the grammar or the gazetteers:** Person names such as *Lloyd Bentsen* or *Strobe Talbott* are difficut to recognize if the first names (*Lloyd*, *Strobe*), or the last names (*Bentsen*, *Talbott*) are not present in the gazetteers.

**Unrecognized transformations:** The names *Robert S. Miller* extended as *Robert S. "Steve" Miller* to explicitly present the meaning of the *S* are difficult to recognize. This type of sequence is both too complex and too specific for the analyzer. Extending the grammar to account for this type of sequence would introduce more noise (false recognitions) than reduce silence (false non-recognitions).

**Ambiguous words:** This is the case with words such as *Sun* in *Sun Tzu*, which designates a person. If *Tzu* is an unknown word and if there is no clear context to disambiguate the sequence (lack of trigger word for example), then such a sequence is difficult to tag correctly.

**Ambiguous sequences:** This is often the case when distinguishing the name of an organization from the name of a person. Names like *Mary Kay* are recognized as person names, while they are organization names. When using the evaluation protocol of MUC, these extractions are considered wrong because the semantic tag is wrong. Without any information from the context to disambiguate (eg. *Ms. Mary Kay* versus *Mary Kay inc.*) the only solution is to include a priori these sequences in a gazetteer.

### 5.2 Non-addressed errors

A certain number of problem have not been taken into account. Indeed, we have concentrated our efforts on the absence of linguistic markers. Although named-entity extraction is currently the most reusable information extraction task over different corpora; the fact remains that when applied to specific types of texts (with different discourse domains, genres or modes of communications) different types of phenomena should be taken into account.

**Discourse-Specific Terminology:** The Valcartier corpus, taken form a military context, includes specific abbreviations and terminology of the military domain (eg. *Alpha, Bravo, Charlie*) that, in this domain, either do not constitute proper names at all or at least, should tagged with a different semantic tag. Here, the use of a domain-dependant dictionary and anti-dictionary seems to be necessary.

**Genre-Specific Grammar:** Different genres using different writing guidelines can create extraction errors. In particular spelling mistakes in informal documents can occur frequently and thus cause significantly extraction errors. A first though may be to pre-process the document through a speller; however, because proper names include many unknown words (because proper names are open-class words) a speller will have difficulty recognizing them and may try to correct them. The use of a speller may thus not be such an interesting solution. In addition, in the analysis of languages using diacritics, the lack of suck marker in informal documents certainly causes recognition problems. In the case of e-mails, many research have performed descriptive analysis of writing habits used in e-mails and other informal computer-mediated communications [20, 23, 13, 8]. These observations should be used as an aid in developing grammars for proper-name extraction in these texts.

**Genre-Specific Terminology:** Terminology present in texts from informal sources, but considered inappropriate in contexts where writing guidelines are more strict, bring about their share of errors. In the Valcartier corpus, for examples, interjections that also appear in proper-names gazetteers are taged wrongly. This is the case with *Ok* and *Ha*, which, without a discriminating context can be wrongly tagged as location as they are abbreviations of *Oklahoma* and *Hawaï*.

**Mode-Specific Grammar:** finally the mode of communication influences the grammar for proper name extraction. In the Valcartier corpus, for example, proper name restarts can cause errors. For example, *Hal* . . . *Halifax* may be tagged twice. This phenomenon is specific to transcripts of spontaneous communications (oral or computer chat) and require the use of specific annotation scheme (cf. [14, 15]).

## 6 Conclusion

The work presented here allowed us to identify the problems associated with proper-name extraction developed for a specific type of text when applied to different types of texts.

While in journalistic texts, the use of grammar rules in the identification of proper names is a well-mastered task that yields results that are comparable to human ones; the same task on documents from informal exchanges has received much less attention and yields less impressive results. Two independent experiences on corpora from *real applications* have shown that systems yielding acceptable results in journalistic texts have yielded much lower scores (from around 0.90 P&R to around 0.50).

By analysing the errors that were committed on two types of informal documents, we have identified the sources of errors that currently lower scores on rule-based approaches. Following these observations, we have proposed strategies to adapt rule-based systems developed for journalistic text to non-journalistic texts. The implementation of these strategies in the Lexis system has increased the extraction scores of Lexis on the Communication corpus from 0.50 to 0.84 P&R. The analysis of the remaining errors has allowed us to highlight certain types of errors that are dependant on the discourse domain, the textual genre and the mode of communication. Named-entity extraction, although the most re-usable task in information extraction needs, nonetheless, to taken into account specific characteristics of the corpus in order to achieve human-level scores regardless of the corpus.

## References

[1]*Proceedings of the Fourth Message Understanding Conference*, San Francisco, 1992. Morgan Kaufmann Publishers.

[2]*Proceedings of the Fifth Message Understanding Conference*, San Francisco, 1993. Morgan Kaufmann Publishers.

[3]*Proceedings of the Sixth Message Understanding Conference*, San Francisco, 1995. Morgan Kaufmann Publishers.

[4]J. Aberdeen, J. Burger, D. Day, L. Hirschman, P. Robinson, and M. Vilain. MITRE: Description of the Alembic System as Used for MUC6. [3].

[5]Douglas Appelt, Jerry Hobbs, John Bear, David Israel, Megumi Kameyana, Andy Kehler, David Martin, Karen Myers, and Mabri Tyson. SRI International FASTUS system: MUC-6 test results and analysis. [3].

[6]Douglas Appelt, Jerry Hobbs, John Bear, David Israel, Megumi Kameyana, and Mabri Tyson. FASTUS: a Finite-state Processor for Information Extraction from Real-world Text. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI'93*, 1993.

[7]Douglas Appelt and David Martin. Named Entity Recognition in Speech: Approach and results using the TextPro System. [15].

[8]M. Collot and N. Belmore. Electronic Language: A New Variety of English. In S. Herring, editor, *Computer-Mediated Communications: Linguistic, Social and Corss-Cultural Perspertives*, pages 13–28. John Benjamins, Amsterdam/Philadelphia, 1996.

[9]Béatrice Daille, Benoît Habert, Christian Jacquemin, and Jean Royaute. Empirical observation of term variations and principles for their description. *Terminology*, 3(2):197–258, 1996.

[10]Lin Dekan. Using Collocation Statistics in Information Extraction. San Francisco, 1998. Morgan Kaufmann Publishers. http://www.muc.saic.com.

[11]Ralph Grishman. Where's the Syntax? The NYU MUC-6 System. [3].

[12]Dilek Hakkani-Tür, Tür Gökhan, Andreas Stolcke, and Elizabeth Shriberg. Combining Words and Prosody for Information Extraction From Speech. In *Proceedings of Eurospeech-99*, Budapest, Hongrie, 1999.

[13]S. Herring. Introduction. In S. Herring, editor, *Computer-Mediated Communications: Linguistic, Social and Corss-Cultural Perspertives*, pages 1–10. John Benjamins, Amsterdam/Philadelphia, 1996.

[14]HUB. Proceedings of the DARPA Broadcast Transcription and Understanding Workshop. Lansdowne, Virginia, 1998.

[15]HUB. Proceedings of the DARPA Broadcast News Workshop. Herndon, Virginia, 1999.

[16]Leila Kosseim and Guy Lapalme. Exibum: Un système expérimental d'extraction d'information bilingue. In *Actes de la Rencontre Internationale sur l'extraction, le Filtrage et le Résumé Automatique (RIFRA-98)*, pages 129–140, Sfax, Tunisia, 1998.

[17]Francis Kubala, Richard Schwartz, Rebecca Stone, and Ralph Weischedel. Named Entity Extraction from Speech. [14].

[18]David Miller, Richard Schwartz, Ralph Weischedel, and Rebecca Stone. Named Entity Extraction from Broadcast News. [15].

[19]Tom Mitchell. *Machine Learning*. McGraw-Hill, New York, 1997.

[20]D. Murray. The context of oral and written language: A framework for mode and medium switching. *Language in Society*, 17:351–373, 1988.

[21]Thierry Poibeau. Le repérage des entités nommées: un enjeu pour les systèmes de veille. In *Terminologies Nouvelles (actes du colloque Terminologie et Intelligence Artificielle, TIA'99, Nantes)*, number 19, pages 43–51, Nantes, France, 1999.

[22]Jean Senellart. Locating noun phrases with finite state transducers. In *Proceedings of the 17th International Conference on Computational Linguistics (COLING'98)*, Montréal, Canada, 1998.

[23]S. Yates. Oral and Written Linguistic Aspects of Computer Conferencing: A Corpus Based Study. In S. Herring, editor, *Computer-Mediated Communications: Linguistic, Social and Corss-Cultural Perspertives*, pages 29–46. John Benjamins, Amsterdam/Philadelphia, 1996.