Chapter 7

# Certainty Identification in Texts: Categorization Model and Manual Tagging Results

**Victoria L. Rubin and Elizabeth D. Liddy**
*Syracuse University*
*School of Information Studies*
*Center for Natural Language Processing*
*Syracuse University*
*Syracuse, NY 13244-1190, U.S.A*
Email: {vlrubin, liddy}@syr.edu

**Noriko Kando**
*National Institute of Informatics*
*2-1-2 Hitotsubashi, Chiyoda-ku*
*Tokyo 101-8430, Japan*
Email: kando@nii.ac.jp

**Abstract**

This chapter presents a theoretical framework and preliminary results for manual categorization of explicit certainty information in 32 English newspaper articles. Our contribution is in a proposed categorization model and analytical framework for certainty identification. Certainty is presented as a type of subjective information available in texts. Statements with explicit certainty markers were identified and categorized according to four hypothesized dimensions – level, perspective, focus, and time of certainty. The preliminary results reveal an overall promising picture of the presence of certainty information in texts, and establish its susceptibility to manual identification within the proposed four-dimensional certainty categorization analytical framework. Our findings are that the editorial sample group had a significantly higher frequency of markers per sentence than did the sample group of the news stories. For editorials, high level of certainty, writer's point of view, and future and present time were the most populated categories. For news stories, the most common categories were high and moderate levels, directly involved third party's point of view, and past time. These patterns have positive practical implications for automation.

**Keywords**: Subjectivity, manual tagging, natural language processing, uncertainty, epistemic comments, evidentials, hedges, certainty expressions; point of view, annotating opinions.

# 1. Analytical Framework

## 1.1 Introduction: What is Certainty Identification and Why is it Important?

The fields of Information Extraction (IE) and Natural Language Processing (NLP) have not yet addressed the task of certainty identification. It presents an ongoing theoretical and implementation challenge. Even though the linguistics literature has abundant intellectual investigations of closely related concepts, it has not yet provided NLP with a holistic certainty identification approach that would include clear definitions, theoretical underpinnings, validated analysis results, and a vision for practical applications. Unravelling the potential and demonstrating the usefulness of certainty analysis in an information-seeking situation is the driving force behind this preliminary research effort.

Certainty identification is defined here as an automated process of extracting information from certainty-qualified texts or individual statements along four hypothesized dimensions of certainty, namely:

- what degree of certainty is indicated (**LEVEL**),
- whose certainty is involved (**PERSPECTIVE**),
- what the object of certainty is (**FOCUS**), and
- what time the certainty is expressed (**TIME**).

Some writers consciously strive to produce a particular effect of certainty due to training or overt instructions. Others may do it inadvertently. A writer's certainty level may remain constant in a text and be unnoticed by the reader, or it may fluctuate from statement to statement and blatantly attract readers' attention. There may be evident traces of such writers' behavior that may become apparent upon a closer examination with a systematic theoretical framework. The difficulty is to discern such traces at the discourse, syntactic, semantic, and lexical levels, wherever such explicit information is available and to be able to recognize these explicit markers with a series of NLP algorithms.

The importance of assessing how certain writers are about their statements is evident, especially in the stream of constantly updated news reports. Readers want to know, for instance, how sure writers or experts might be about public policy changes, about a possibility of a political or a financial turmoil, about what the government's intentions are regarding interest rates or about chances of coup d'etats versus peaceful transfers of power.

Recognizing such certainty assessments would traditionally be considered a task for humans. While humans may rely to some extent on the big picture as obtained from world knowledge and prior experience, much certainty information comes from linguistic coding in texts and may be accessible to a systematic analysis with the help of NLP algorithms. Combined with the capabilities of an IE system, the task of linguistic de-coding of certainty information could then be handled successfully automatically, and the results could be presented to users for confirmation and possible modifications.

## 1.2 Certainty, Explicit Certainty Markers, and Closely Related Concepts

A typical dictionary definition of certainty is "the quality or state of mind of being free from doubt, especially on the basis of evidence" (Merriam-Webster 2004). The notion of certainty in the context of this chapter incorporates a full spectrum of certainty states ranging from doubt to

complete conviction in the truth of a statement. There are several related concepts that have been previously addressed in NLP and linguistics literature: subjectivity, modality, evidentiality, and hedging. This section reports on how these closely related linguistic concepts are interpreted to define certainty, and concludes with a list of terms that are considered to be explicit certainty markers.

### *1.2.1 Subjectivity*

This study departs from the notion of subjectivity. Uncertainty, or certainty in terms of this chapter, is the speculative type of subjectivity (Wiebe 2000) that is analogous to the other types of subjectivity for which manual and automated tagging has proven to be a feasible NLP task (Wiebe et al. 2001). Subjectivity has been defined in NLP as "aspects of language used to express opinions and evaluations" (Wiebe 1994, 2000, Wiebe et al. 2001). Cognitive Grammar describes subjectivity as "a part of the conceptual structure of information that lies behind linguistic 'packaging'" (Mushin 2001).

Subjectivity tagging is considered particularly relevant for the news report genre (Wiebe et al. 2001). When developing news report schemata components for an automated text structurer, Liddy et al. (1993) noted that subjectivity, or objectivity, as an attribute in texts, deserved special attention. They observed that binary distinctions of statements (e.g., "+ subjective" or "- subjective") may not be sufficient to adequately represent micro-level similarities and distinctions in texts. In addition, discourse components may have multiple dimensions embedded in each of the concept labels (Liddy et al. 1995). This study further explores identifiable dimensions of certainty in written news reports and editorials.

### *1.2.2 Epistemic Comments and Modality*

Certainty can also be seen as a variety of epistemic modality expressed through epistemic comments. One type of epistemic comment is certainty expressions (e.g., *probably, perhaps, undoubtedly*) that provide clues to the writer's certainty or assessment of the truth of a statement and qualify a writer's attitude towards expressed knowledge. Epistemic comments reflect epistemic modality, which is described in Functional Linguistics as a writer's assumptions or assessments of possibilities expressed in statements, specifically regarding confidence in the truth of the expressed propositions (Coates 1983). Writer's confidence in the truth is synonymous with certainty. In other words, certainty is a writer's assessment of the truth of the statement.

### *1.2.3 Evidentials and other Reportive Means*

Certainty, in particular in languages other than English, can be expressed by means of evidentials that reveal a degree of reliability of expressed information. English resorts to other reportive means, such as attributive adverbials (e.g*., supposedly, allegedly*) and reporting verbs (e.g., *claim, suggest*).

Evidentials were originally narrowly described as "suffixes expressing subjective relations… those expressing subjective knowledge" (Mushin 2001) and later understood as a semantic category that specifies the type of the reported information source. Based on her comparison of Macedonian, Japanese and English corpora, Mushin (2001) concluded that English lacks clear grammatical markers of evidentiality, and that most types of English discourse are "faceless" in the sense of lacking epistemic evaluation in a grammatical inventory of reportive suffixes or other purely

grammatical manifestations. However, she comes to a promising conclusion that English compensates for such lack of reportive means by other identifiable means by which speakers express, for instance, that the story they are telling was the product of someone else's telling. In particular, she notes that English does have a rich inventory of adverbials of "propositional attitude" (Mushin 2001).

The choice of reporting verbs depends on how strongly the writer wants to be aligned with the reported source (Hyland 1998). Bergler et al. (2004) also hypothesize that the description of the source and the choice of the attributive or reporting verbs can, in fact, express the writer's level of confidence in the attributed material. Such verbs can be used "both, to bolster a claim made in the text already, or to distance the author from the attributed material, implicitly lowering its credibility" (Anick and Bergler, 1992, cited in Bergler et al., 2004).

In the same line of Evidential Semantics research, Chafe (1986) suggested a model that addresses reliability of knowledge expressed through evidentials. Knowledge was broadly defined as "the basic information whose status is qualified in one way or another by markers of evidentiality," where the notion of evidentiality extends beyond evidence and can be as inclusive as any "attitude toward knowledge" (Chafe 1986). In other words, he suggests different statuses that reveal the reliability of the expressed information: "People are aware, though not necessarily consciously aware, that some things they know are *surer bets for being true than others,* that not all knowledge is equally reliable." This is what is called "certainty" in this chapter. Chafe continues: "Thus, one way in which knowledge may be qualified is with an expression indicating the speaker's assessment of its *degree of reliability*, the likelihood of its being a fact." Chafe's degrees of reliability can be expressed in English through propositional attitude adverbials (in Mushin's terms), which are the same as epistemic comments (in Coates' terms), or explicit certainty markers in this chapter's terms.

### *1.2.4 Hedges and Other Terminology for Explicit Certainty Markers*

The term traditionally associated with linguistic uncertainty, especially in scientific writing, is hedging. Hedging was introduced by Lackoff (1972) and has generally been defined as "words whose job is to make things more or less fuzzy." In Hyland (1998), hedging refers to "any linguistic means used to indicate either a) a lack of commitment to the truth value of an accompanying proposition, or b) a desire not to express that commitment categorically." In research articles, hedges are "a crucial means of presenting new claims for ratification and are among the primary features which shape the research article as the principle vehicle for new knowledge" (Hyland 1998).

Hyland (1998) identifies several categories of how hedges can be expressed in everyday speech and scientific writing. The following surface lexical markers are used to attenuate strength of utterance: epistemic adjectives, epistemic adverbs, lexical verbs, auxiliary verbs, prosody, tag questions, and verbal fillers. Syntactic markers include *if*-clauses of condition and concessions, contrastive markers (e.g., *nevertheless*), and passivization (e.g., *it can be questioned*). Several other devices are classified as hedges particular to scientific writing only: hedging quantities for purposeful imprecision, admitting to lack of knowledge, citing a source, and referring to limitations (of the model, experimental conditions, or methods) (Hyland 1998).

Thus, hedging is a device that indicates a lack of commitment to the statement, reveals scepticism, expresses caution, or displays an open mind about a proposition. In this study hedges classify

statements into low or moderate levels of certainty. Several other linguistic means of a writer's assessment of knowledge such as shields, approximators (Lackoff 1972), understatements, and tentatives; as well as intensifiers (Cappon 2000), emphatics (Holmes 1990), boosters and assertives (Searle 1979) are considered to be explicit certainty markers of varying certainty levels.

In summary, certainty is viewed as a type of subjective information available in texts and a form of epistemic modality expressed through explicitly-coded linguistic means. Such devices as subjectivity expressions, epistemic comments, evidentials, reporting verbs, attitudinal adverbials, hedges, shields, approximators, understatements, tentatives, intensifiers, emphatics, boosters, and assertives, often overlap in their definitions, classifications, and lexical representations in English. In essence, they perform the same role for the purpose of this study. They explicitly signal the presence of certainty information that covers a full continuum of a writer's confidence, ranging from uncertain possibility and withholding full commitment to statements to a confident necessity, reassurance, and emphasizing of the full commitment to statements. For the purpose of this study, these devices are all called explicit certainty markers.

In the remainder of the paper, we develop a certainty categorization model, report on preliminary results, and conclude with outlined challenges and applications.

## 2. Proposed Certainty Categorization Model

Expressing some degree of certainty in language is inevitable, just as one is bound to have a spatial angle of vision. By analogy with subjectivity, certainty is generally understood to be a pragmatic position rather than a grammatical feature. Banfield (1982) observed that subjectivity, a closely related concept, is a spatial notion by nature, and in language, it is taken to be located in a speaker. While it is questionable whether truly objective statements may exist, it seems even less likely that a statement may exist without a degree of certainty in the presented information. Each statement should potentially reveal a particular pragmatic position, or a level of certainty, but not all of them are explicitly marked. The commonly used declarative mood of stating facts and opinions may have an implied certainty level without any explicit indication that would be considered identifiable for Information Extraction purposes. Statements with implicit certainty levels are not discussed under the current categorization model, they are grouped into a separate pool of no identifiable explicit certainty information.

The proposed certainty categorization model distinguishes 4 dimensions for explicitly identifiable certainty. The certainty level is the first and most important dimension. The other three are perspective, focus, and time (Figure 1). Each dimension is subdivided into several categories creating 72 possible dimension-category combinations (4 levels by 3 perspectives by 2 foci by 3 times).

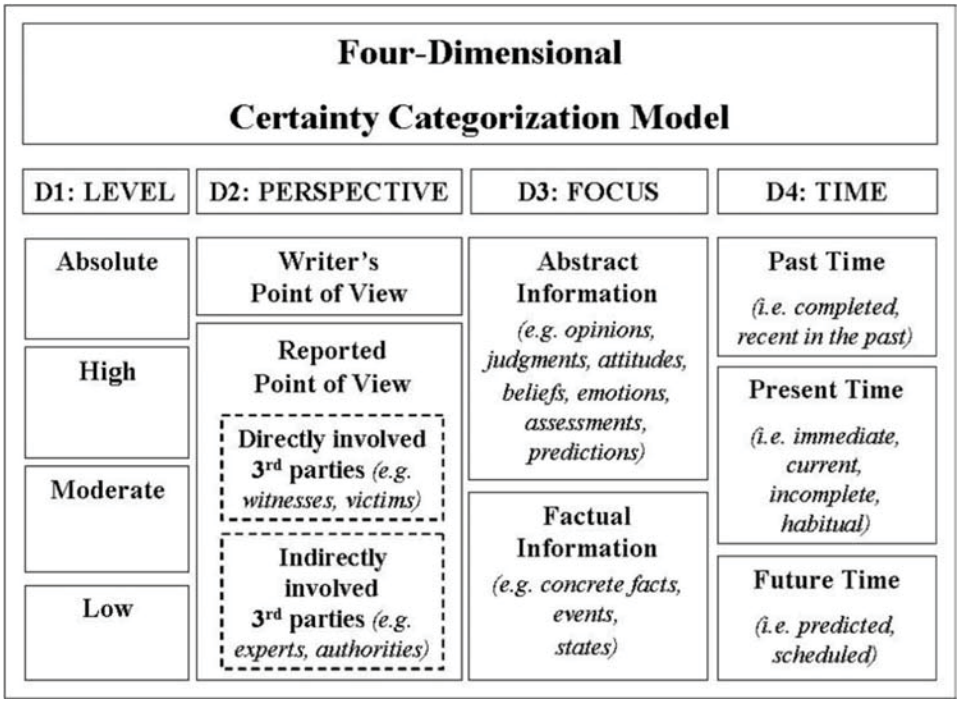


*Figure 1. Four-Dimensional Certainty Categorization Model with the Four Hypothesized Dimensions (across) and Their Categories (down).*

**2.1 First Dimension: Certainty Level**

The concept of certainty seems to fall inherently into levels. We suggest the division of the certainty level dimension into four categories - absolute, high, moderate, and low. The excerpts below exemplify a decreasing degree of certainty from absolute certainty in the first example to low certainty in the last one. The explicit certainty markers are highlighted in bold.

> *(1) An enduring lesson of the Reagan years,* ***of course****, is that* ***it really does take smoke and mirrors to*** *produce tax cuts, spending initiatives and a balanced budget at the same time. (ID=e3.28)*

> *(2) … but* ***clearly*** *an opportunity is at hand for the rest of the world to pressure both sides to devise a lasting peace based on democratic values and respect for human rights. (ID=e22.6)*

> *(3) That fear* ***now seems*** *exaggerated, but it was not entirely fanciful. (ID=e4.8)*

> *(4) So far the presidential candidates are more interested in talking about what a surplus* ***might buy*** *than in the painful choices that lie ahead. (ID=e3.7)*

Having an explicit certainty marker that places a statement into one of the levels of certainty is what distinguishes a certainty-qualified statement from a non-marked one. Each certainty-qualified statement is further hypothesized to contain some information regarding a perspective, a focus, and a time of certainty.

### 2.2 Second Dimension: Perspective

The second dimension in Figure 1, the certainty perspective, separates the certainty point of view into the writer's and the reported points of view. The writer's certainty refers to the experiencer of certainty at the time of writing a statement as exemplified below.

> *(5) More evenhanded coverage of the presidential race would help enhance the legitimacy of the eventual winner, which* ***now appears likely to be*** *Putin. (ID=e8.14)*

The certainty is clearly attributed to the writer in the example above. A practical question is whether third party voices can be isolated from the author's since they are presented through the author's prism.

Reported point of view can refer to either individuals or organizations. It is divided into two sub-categories. First, those of directly involved third parties, such as victims, witnesses, and survivors, are direct event participants, who are either present at the described event or whose life is directly affected by the events. Second, those of indirectly involved third parties, such as experts, authorities, and analysts, are tangentially related to the event in professional or other capacities.

> *(6) The Dutch recruited settlers with an advertisement that* ***promised*** *to provide them with slaves who "****would accomplish*** *more work for their masters, …" (ID=e27.13)*

> *(7) The historian Ira Berlin, author of "Many Thousands Gone,"* ***estimates*** *that one slave perished for every one who survived capture in the African interior… (ID=e27.8)*

In the first example the writer reports on the certainty of the group of direct participants, the Dutch; and in the second example, the writer refers to the expert historian's opinion.

The writer's certainty as expressed in text should not be confused with the reader's certainty that the text is believable. The writer's certainty about his or her own and others' assertions is captured in texts. The reader's certainty is related to numerous factors that inform his or her own subjectivity, or point of view. The former is accessible for analysis since it has a written record, but the latter is less tangible and may reflect high inter-personal variability. Thus, the reader's certainty is out of scope for this study, as this inquiry focuses on the writer's certainty model and its multi-dimensional complexity in the newspaper context.

### 2.3 Third Dimension: Certainty Focus

The third dimension, the certainty focus, is divided into abstract and factual information in the narrative. The term focus is used in van Dijk's (1981) localized selection sense as the referent, or the object, subject, or topic of conversation that is being talked about, or predicated upon, in a particular localized syntactic unit, such as a sentence or a clause.

Abstract information may include judgments, opinions, attitudes, beliefs, moral principles, and emotions. Usually such statements, as in the example below, reflect an idea that does not represent an external reality, but rather a hypothesized world, existing only in someone's mind, and separated from embodiment or object of nature.

> *(8) In Iraq, the first steps **must be taken** to put a hard-won new security council resolution on arms inspections into effect. (ID=e8.12)*

Factual information contains reports of states or events, evidence, and known facts. It is usually based on facts that have an actual existence in the world of events.

> *(9) The settlement **may not fully compensate** survivors for the delay in justice, ... (ID=e14.19)*

### 2.4 Fourth Dimension: Time

The fourth dimension accounts for the relevance of time (past, present and future) to the moment when the statement was written. The past naturally includes completed or recent states or events; the present consists of current, immediate, and incomplete states of affairs; and the future contains predictions, plans, warnings, and suggested actions. The time dimension is relevant since certainty of predictions into the future, for instance, may alter an action plan for someone who is reviewing certainty analysis information in a systematic way, as business or intelligence analysts may do.

## 3. Empirical Study

In order to obtain a preliminary sense of the nature and frequency of certainty markers in text, we conducted a pilot study.

### 3.1 Research Questions

The goals of the study were to empirically determine:

1. if the sample data support the hypothesized four dimensional categorization model,
2. if so, which categories are most and least frequent for a sample of English news articles,
3. if the data do not support the model, how the categorization might be enhanced,
4. whether there are differences in certainty distributions between editorials and news stories, overall and per hypothesized category,
5. how many perceived categories of certainty can be distinguished within each dimension.

### 3.2 Data

We manually analyzed 32 articles published in *The New York Times* during the first week of January 2000 (from the AQUAINT Corpus of English Texts). This constitutes a total of 685 sentences, excluding headlines. The topics of the sample articles varied – the editorials included discussions of political leaders, presidential and state government campaigns, the economic and financial situations in US, Croatia, and Angola, recent historical discoveries, pharmaceutical consumer alerts, and the role of the Internet and computers in everyday lives. The news included reports on the misnumbering of *The New York Times* issues, on the controversy around the millennium and Y2K bug, and on women's basketball.

### 3.3 Analysis Methods

The data were analyzed manually at the sentence-level by one coder, the first author. If a sentence contained explicit certainty information markers, it was decomposed along each certainty dimension by answering questions such as "What is the certainty level?" and "Whose perspective is being presented?" The number of occurrences of markers per article were totalled and adjusted for article sentence length, resulting in one frequency score per article. The length of explicit certainty markers was not pre-determined.

First, we were interested in an overall frequency of occurrence of explicit certainty markers across all of the data. Second, we identified whether the two sample groups, editorials and news stories, had significantly different means. Third, we looked at the overall distribution of frequency scores (in markers per sentence) per category within each dimension. For instance, were there more occurrences of high or low levels of certainty on average? Fourth, for the editorial sample, we identified the least and most frequent combinations out of 72 possible dimension-category combinations. And last, we assessed whether the data easily fell into the hypothesized categories.

### 3.4 Results and Discussion

#### *3.4.1 Certainty Markers Frequency Distributions in Two Sample Groups*

In the total set of 32 articles (685 sentences), an average of 0.53 explicit certainty markers per sentence were identified. Identified certainty markers included but were not limited to *it was not even clear that*, *remains to be seen*, *don't believe they will*, *not necessarily*, *we thought*, *estimated*, *seems exaggerated, would probably have to, is expected to,* and *will almost certainly have to.*

The sample group of 28 editorials (564 sentences) contained more explicit certainty markers per sentence ($M$=0.6, $SD$ = 0.26) than the sample group of 4 news stories (121 sentences; $M$=0.46, $SD$ =0.04 markers per sentence). This difference was statistically significant, $p$ = 0.0056, two-tailed heteroscedastic t-test.

Within three dimensions, level, perspective, and time, average frequencies of occurrence of explicit certainty markers per sentence differed from category to category, as well as between sample groups.

#### *3.4.2 Certainty Frequency Distributions in the Level Dimension*

Table 1 shows that, of all possible level categories, the high certainty level contained most markers per sentence (0.33). Here is an example sentence from an editorial that falls into the category of high certainty:

> *(10) The crowd cheering the opening of the Erie Canal in 1824* ***knew*** *that the city* ***would forever be*** *transformed, Wallace notes. (ID=e28.19,* certainty level = high*)*

In an automated implementation, the Information Extraction frame would receive a value of "high" in the certainty level slot, as shown above. The explicit certainty markers are in bold as in the rest of the data samples throughout the chapter.

| | Editorials Sample Group | | News Stories Sample Group | |
|---|---|---|---|---|
| **LEVEL** | Mean, markers per sentence | St. dev., markers per sentence | Mean, markers per sentence | St. dev., markers per sentence |
| **Absolute** | 0.07 | 0.09 | 0.03 | 0.05 |
| **High** | **0.33** | 0.17 | **0.19** | 0.09 |
| **Moderate** | 0.17 | 0.14 | **0.20** | 0.15 |
| **Low** | 0.04 | 0.06 | 0.04 | 0.05 |

*Table 1. Distribution of Markers Per Sentence in the Four Level Categories.*

In news stories, both high and moderate levels of certainty were the two most prominent levels (approximately 0.2 markers per sentence). An example of the moderate level of certainty follows:

> *(11) But as midnight closed in, the streets teemed with people and **there seemed to be** little left of the anxiety over terrorist attacks that prompted the mayor of Seattle last week to cancel a major outdoor celebration around the city's famed Space Needle. (ID=n3.9,* certainty level = moderate*)*

### *3.4.3 Certainty Frequency Distributions in the Perspective Dimension*

Table 2 demonstrates that in editorials certainty from the writers' points of view is expressed more commonly (0.43 mean markers per sentence) than certainty of third parties (0.13 and 0.04), as is expected.

| | Editorials Sample Group | | News Stories Sample Group | |
|---|---|---|---|---|
| **PERSPECTIVE** | Mean, markers per sentence | St. dev., markers per sentence | Mean, markers per sentence | St. dev., markers per sentence |
| **Writer's** | **0.43** | 0.23 | 0.16 | 0.10 |
| 3rd **directly** involved party's | 0.13 | 0.13 | **0.24** | 0.11 |
| 3rd **indirectly** Involved party's | 0.04 | 0.06 | 0.05 | 0.06 |

*Table 2. Distribution of Markers Per Sentence in the Three Perspective Categories.*

Consider that even though this example sentence talks about a third party, the expressed certainty actually belongs to the writer:

> *(12) He also **ought to** urge France and Russia to persuade Saddam Hussein to accept the resolution. (ID=e8.14,* perspective = writer's point of view*)*

We also observed that in news stories attention shifts to the certainty of the directly involved third parties (0.24) such as presidential candidates, political leaders, a Cuban orphan and his family, and just a person waiting for a flight at the airport whose direct words are cited below:

> *(13) "I **think** it **will probably be** OK..." (ID=n4.23,* perspective = directly involved third party's point of view*)*

The indirectly involved third parties are rather rare and usually occur in the form of experts' opinions, sometimes cited as well. For instance, the economists' points of view rendered below reflect their certainty, and the writer may or may not be sure about that statement:

> *(14) **Most** economists **believe** Alan Greenspan is more responsible for the economy's spectacular performance than Congress, Presidents Bush and Clinton or any other*

*identifiable factor. (ID=e9.1,* perspective = indirectly involved third party's point of view*)*

The difficulty for automation will likely be in correctly interpreting the writer's intended use of the experts' opinions. Sometimes the reference to the source is vague but it is quite clear that the expressed certainty is the writer's:

*(15) Although **some research suggests** that **some** supplements **can** produce positive health effects, there have also been cases where people have been made ill by supplements, or their conditions have become worse… (ID=e28.3,* perspective = writer's point of view*)*

### *3.4.4 Certainty Frequency Distributions in the Focus and Time Dimensions*

Table 3 demonstrates that abstract and factual foci of certainty were approximately evenly distributed in editorials (0.33 and 0.27) and in news stories (0.23), even though the editorial sample group had a larger deviation from the mean compared to the news stories.

| | **Editorials Sample Group** | | **News Stories Sample Group** | |
|---|---|---|---|---|
| **FOCUS** | Mean, markers per sentence | St. dev., markers per sentence | Mean, markers per sentence | St. dev., markers per sentence |
| **Abstract** | 0.33 | 0.20 | 0.23 | 0.05 |
| **Factual** | 0.27 | 0.19 | 0.23 | 0.09 |

*Table 3. Distribution of Markers Per Sentence in the Two Focus Categories.*

As for the time dimension, it is not surprising that certainty analysis captures the news stories' tendency to report events in the past (0.20, as opposed to 0.11 and 0.14, as seen in Table 4).

| | **Editorials Sample Group** | | **News Stories Sample Group** | |
|---|---|---|---|---|
| **TIME** | Mean, markers per sentence | St. dev., markers per sentence | Mean, markers per sentence | St. dev., markers per sentence |
| **Past** | 0.14 | 0.12 | **0.20** | 0.11 |
| **Present** | **0.24** | 0.18 | 0.11 | 0.05 |
| **Future** | **0.22** | 0.16 | 0.14 | 0.09 |

*Table 4. Distribution of Markers Per Sentence in the Three Time Categories.*

The editorials' tendency to state opinions about current and predicted events also becomes apparent. An example from a news story about millennium flight cancellations refers to a piece of factual information with certainty expressed by the experts in the past:

*(16) The failure lasted only about 30 minutes and had no operational effect, the FAA said, adding that **it was not even clear that** the problem was caused by the date change. (ID=n4.19,* time = past*)*

An example from editorials demonstrates an abstract writer's assessment in the present:

*(17) **Whatever happens next,** these candidates have shown that one-on-one debates **really can** give voters a choice on issues and on leadership temperament as well. (ID=e16.18,* time = present*)*

Also, many editorials had a closing statement in the last paragraph that contained some certainty markers that either urged action or expressed an overall opinion statement in the form of a prediction, such as shown below:

*(18) There will be problems along the way, but the Internet **will likely** change the way America does business far beyond the habits of holiday shoppers. (ID=e2.22,* time = future*)*

On the whole, the presence of data in each category suggests that the categorization model is viable when applied manually. Now a gold standard and a codebook of rules can be created for an inter-coder agreement study and further automation of the process. High frequency of explicit certainty markers in some categories emphasizes where linguistic analysis should be concentrated to cover the majority of certainty expression cases.

### *3.4.5 Certainty Marker Occurrences in Dimension-Category Combinations*

Table 5 shows the distribution of occurrences of explicit certainty markers in dimension-category combinations for editorials, the larger of the two sample groups. The table is to be read by cross-referencing the two dimensions in columns (focus and level) with the two dimensions in rows (perspective and time). For instance, the absolute level of a writer's certainty about abstract information in the past only happened once, while in the present it occurred 18 times. The table forms 72 possible combinations (4 levels by 3 perspectives by 2 foci by 3 times), plus an additional group ('None') that recorded the occurrence of statements containing no explicit certainty information (289 sentences). In total, there were 624 occurrences of certainty-qualified sentences and non-qualified sentences, while the editorial sample group contained 564 sentences. This means that 335 certainty markers were assigned to 275 sentences. The difference of 60 occurrences is explained by a special treatment of complex sentences. The unit of analysis was generally a sentence; but complex sentences were split into two or more simple sentences, if each simple sentence expressed a different idea qualified by a distinct certainty marker.

| **FOCUS** | | Abstract Information | | | | Factual Information | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | **LEVEL** | | | | | | | | |
| **PERSPECTIVE** | **TIME** | Abs. | High | Mod. | Low | Abs. | High | Mod. | Low | |
| | | | | | | | | | | Total |
| Writer's | Past | 1 | 8 | 10 | 1 | | 12 | 11 | | 43 |
| | Present | 18 | **29** | 16 | 8 | | 13 | 10 | 1 | 95 |
| | Future | 13 | **25** | 12 | 2 | 2 | **27** | 12 | 3 | 96 |
| 3rd Indirectly | Past | | 8 | 4 | | 1 | 11 | 2 | | 26 |
| Involved | Present | 2 | 3 | 2 | | 1 | 7 | 5 | 1 | 21 |
| Arty | Future | 1 | 8 | 1 | | 1 | 11 | 2 | 2 | 26 |
| 3rd Directly | Past | | 3 | 1 | 2 | | 4 | | | 10 |
| Involved | Present | | 2 | 4 | | | 4 | 2 | 1 | 13 |
| Party | Future | | | | | | 5 | | | 5 |
| **NONE** | | | | | | | | | | 289 |
| Total | | 35 | 86 | 50 | 13 | 5 | 94 | 44 | 8 | 624 |

*Table 5. Certainty Markers Count of Occurrences in the Editorials Sample Group.*

Of the 72 possible combinations, 15 combinations were rather typical in editorials. Three had an unusually high representation in editorials. These combinations are writer's high level of certainty about abstract information in the present or future, such as predictions and current assessments, which had 29 and 25 occurrences respectively. There were also 27 occurrences of a writer's future high level certainty factual predictions, stating with high certainty what will happen in the future. Twelve combinations accounted for the majority of occurrences (between 10 and 18).

Additionally, 35 combinations were found to be rare in editorials, with ≤ 8 occurrences in our data; for instance, the low level of a writer's certainty about present or future factual information had 1 and 3 occurrences respectively. The remaining 22 combinations did not have any representation in our data. For instance, directly involved third parties' low level of certainty about abstract information in either past, present or future were never found.

The observed distribution is consistent with the goal of editorials to state opinions, inevitably with different levels of certainty. It directs us to the combinations that cover the majority of explicit certainty markers and provide guidance in automating the categorization.

### *3.4.6 Challenges*

One criterion for deciding whether the sample data support the hypothesized model is "ease-of-fit" experienced by the coders when analyzing the data, in other words, whether the data landed naturally or had to be forced into the allotted categories within each dimension. Our coder found the easiest dimension for categorization to be time. The only adjustment that had to be made was an expansion of the notion of present time to include regular or habitual actions. It was also noted that certainty level categorization may include an additional fifth category of uncertainty in the model refinements. Currently, we have made no distinction between low certainty and uncertainty.

The dimension of perspective, on the other hand, is sufficiently granular and, depending on the application, could even be collapsed into two main categories: the writer's and third party's point of view. The benefit of distinguishing a rather rare category of third indirectly involved party's perspective is for when we are particularly interested in, let's say, experts' certainty. The experts could also be sub-divided into groups of political, economic, media-related, religious expertise and influence that can be identified with NLP and IE tools.

The distinction of focus into factual and abstract information presented the most difficulties for annotation due to fuzzy boundaries between known facts and opinions. The focus was considered factual when an event or state of affairs was clearly mentioned. Otherwise, the focus was considered abstract and further sub-categorized into a type of opinion, judgment, or emotion, such as fear, a warning, an assessment, or a conviction, the details of which are not herein reported.

As concluded in most pilot studies, the annotation could be improved with a clearer set of guidelines and definitions. All of the hypothesized categories in the model are not final and are open to further refinement, as the data analysis proceeds and the theoretical framework stabilizes. In addition, the uneven sizes of the two sample groups (editorials and news reports) presented a statistical challenge. In the future, we will distribute our manual tagging efforts evenly. The first author will incorporate some of the above-mentioned refinements into her doctoral thesis.

The proposed model makes several assumptions and raises several philosophical and practical issues. For instance, we are assuming that uncertainty is expressed due to doubt on the basis of evidence (by our definition), thus we do not make a distinction between truly being uncertain and appearing to be uncertain. There may be other desired reasons for appearing to be uncertain, such as the psychological effect of non-aggression, the social politeness effect, the humbling effect of hedged speech, and practical concerns for avoiding liabilities. Identifying these pragmatic functions of uncertainty is currently out of the scope of the study, but poses a challenge for future automated identification. Another problem is literal interpretation of the identified clues. For instance, the word "certain" itself has an alternate meaning of "definite but not specified." Our model does not include this meaning, but the issue of contextual disambiguation still persists.

## 4. Applications

The categorization, and the resulting linguistic clues and patterns for most frequent categories, will serve as a starting point for a certainty identification module in an intelligence analyst's question and answering system. This model will be applied to identifying and extracting perceived certainty of specified writers or reported third parties relative to the analyst's topics of interest.

The nature of government or business analysts' work requires time and effort to look through enormous amounts of raw textual information such as news reports or editorials in order to find answers to their questions. Traditional search systems can normally alleviate the analysts' load by retrieving texts by key words or phrases. State-of-the-art QA systems can usually localize the best answers and provide them in the form of short answers, best paragraphs, or best-fit full documents. But none of these current methods incorporate certainty of the text.

Certainty analysis, in addition to the QA-application, will add an extra level of sophistication that may assist analysts by alerting them in advance of, or at the time of, retrieval of the certainty of the information in the responses. For instance, data can be analyzed by a set of user-specified parameters from the refined and validated certainty model. An implemented system could be capable of providing users with alerts that warn the user of extreme levels of certainty, multiple levels of certainty in the same texts, absence or presence of certainty-qualified statements, and change of certainty levels. A cross-document summary could trace changes in certainty over time. The goal is to make raw data searchable by natural language certainty-oriented questions such as "*How certain were President Bush's statements about presence of weapons of mass destruction in Iraq in 2003 compared to 2004?"* The answers can be provided in QA system answer style – a flexible number of either best short answers, or most relevant paragraphs, or most relevant documents.

In addition, the collection of certainty expressions may become input data to machine learning algorithms for certainty identification and extraction. It also may suggest a new way of automating genre identification based on differences in markers per sentence frequencies and category distributions. Also, the study results capture current trends in newspaper writing, and are potentially useful as a set of suggestions on how to convey a desired level of certainty.

## 5. Conclusions and Future Work

Our contribution is in a proposed categorization model and analytical framework for certainty identification. The results of this pilot study reveal an overall promising picture of the presence of certainty information in texts, and establish the ability to manually identify and categorize individual statements according to the proposed certainty model.

Generally, our findings are that editorials had a significantly higher frequency of markers per sentence than did news stories. For editorials, high level of certainty, writer's point of view, and future and present time were the most populated categories. For news stories, the most common were high and moderate levels, directly involved third party's point of view, and past time. We are interested in conducting further data analysis per genre within newspaper articles since we have established that the frequency distribution differs depending on genre. This may have implications for automated genre identification. We will use insights from previous work on genre classification (Liddy et al. 1995, Kando 1996).

For editorials, of the possible 72 combinations, the high level of certainty from the writer's point of view expressed abstractly in the present and the future, and expressed factually in the future were most common; 12 combinations were typical; 35 were rather rare; and 22 never occurred. These results shed light on where the majority of lexical, semantic and syntactic patterns can be expected during linguistic analysis of editorials for automating categorization.

The sample data fit relatively well into the pre-defined categories. Some categories, such as the certainty level, can still be further refined with finer distinctions. The focus dimension will require further research. The study yielded a collection of explicit certainty markers which will be further grouped and analyzed in terms of lexical, semantic and syntactic patterns.

We also plan to conduct a full-scale inter-coder reliability study with multiple annotators by adapting our online data collection facility, developed for a concurrent study of emotional subjective content (Rubin et al. 2004).

## 6. Acknowledgements

This research was made possible by the National Science Foundation East Asia Summer Institutes for U.S. Graduate Students Research Grant No. 0309745. The first author extends her gratitude to her host researchers, Dr. Kando and Dr. Adachi, for welcoming this effort at the National Institute of Informatics, Tokyo, Japan. We are also grateful to the colleagues at Dr. Nakagawa's Language Informatics Laboratory, Information Technology Center at the University of Tokyo, and the researchers at Dr. Isahara's Computational Linguistics Group at the Communications Research Laboratory in Kyoto, Japan, for their comments and suggestions during the early stages of the research. We would especially like to thank Dr. Wiebe for her input in personal interactions at the 41st Annual Meeting of the Association for Computational Linguistics in Sapporo, Japan in 2003.

## 7. Bibliography

Anick, P. and Bergler, S. (1992) *Lexical structures for linguistic inference.* In Pustejovsky, J. and Bergler, S. (Eds.) Lexical Semantics and Knowledge Representation. Berlin, Springer Verlag: 121-135.

Banfield, A. (1982) *Unspeakable Sentences.* Routledge and Kegan Paul, Boston.

Bergler, S., Doandes, M., Gerard, C., and Witte, R. (2004) *Attributions.* In Qu, Y., Shanahan, J. G., Wiebe, J. (Eds.) Proceedings of AAAI Spring Symposium: Exploring Attitude and Affect in Text: Theories and Applications, Stanford, CA. AAAI Press.

Cappon, R. J. (2000) *The Associated Press Guide to News Writing.* Foster City, CA, IDG Books Worldwide Inc.

Chafe, W. (1986) *Evidentiality in English Conversation and Academic Writing.* In Chafe, W. and Nichols, J. (Eds.) Evidentiality: The Linguistic Coding of Epistemology. Norwood, New Jersey, Ablex Publishing Corporation. 20: 261-273.

Coates, J. (1983) *The Semantics of the Modal Auxiliaries.* London & Canberra, Croom Helm.

Holmes, J. (1990) *Hedges and boosters in women's and men's speech.* Language and communication 10 (3): 185-205.

Hyland, K. (1998) *Hedging in Scientific Research Articles.* Amsterdam, Philadelphia, John Benjamin Publishing Company.

Kando, N. (1996) *Text structure analysis based on human recognition: Cases of Japanese newspaper and English newspaper.* Bulletin of National Center for Science Information Systems, No. 8, pp.107-126 (Japanese)

Lackoff, G. (1972) *Hedges: a study of meaning criteria and the logic of fuzzy concepts.* Chicago Linguistic Society Papers.

Liddy, E.D., McVearry, K., Paik, W., Yu, E.S., and McKenna, M. (1993) *Development, implementation & Testing of a Discourse Model for Newspaper Texts.* Proceedings of the ARPA Workshop on Human Language Technology, Princeton, NJ, March 21-24, 1993.

Liddy, E.D., Paik, W., and McKenna, M. (1995) *Development and Implementation of a discourse model for newspaper texts.* Proceedings of the AAAI Symposium on Empirical Methods in Discourse Interpretation and Generation. Stanford, CA.

*Merriam-Webster Online Dictionary*, http://www.m-w.com/. Accessed on January 30, 2004.

Mushin, I. (2001) *Evidentiality and Epistemological Stance: Narrative Retelling.* Amsterdam, John Benjamins Publishing Co.

Rubin, V. L., Stanton, J. M., and Liddy E. D. (2004) *Discerning Emotions in Texts.* AAAI Spring Symposium: Exploring Attitude and Affect in Text: Theories and Applications, Stanford, CA.

Searle, J. R. (1979) *Expression and Meaning : Studies in the Theory of Speech Acts.* Cambridge, London, New York, Melbourne, Cambridge University Press.

van Dijk, T. A. (1981) *Studies in the Pragmatics of Discourse*, Mouton Publishers, The Hague, The Netherlands

Wiebe, J. M. (1994) *Tracking Point of View in Narrative.* Computational Linguistics 20 (2): 233-287.

Wiebe, J. M. (2000) *Learning Subjective Adjectives from Corpora.* Proceedings of the 17th National Conference on Artificial Intelligence (AAAI-2000). Austin, Texas, July 2000.

Wiebe, J., Bruce, R., Bell, M., Martin, M., and Wilson, T. (2001) *A Corpus Study of Evaluative and Speculative Language.* Proceedings of the 2nd ACL SIGdial Workshop on Discourse and Dialogue. Aalborg, Denmark, September, 2001.