Aspects of Uncertainty in Soft Data Fusion

Kellyn Rein
Fraunhofer FKIE
Fraunhoferstr. 20
53343 Wachtberg
GERMANY

kellyn.rein@fkie.fraunhofer.de

ABSTRACT

In today's world of asynchronous warfare, where the battle lines are not well-defined, the enemy not easily identifiable and the hunt for terrorists often means understanding how they work as groups and networks of individuals, we often need to rely on information picked up from open source such as social media and blogs, as well as informants, telephone conversations, emails, as well as military and civilian intelligence. The vast amount of natural language information being made available each day can be a tremendous source of intelligence, as long as one find and pull out the most important information. However, making sense of this is like finding a single needle in a thousand haystacks – the need for automatic processing of text-based information for intelligence purposes is essential. However, dealing with soft, natural language information is problematic in ways which can vary considerably from that of hard data from devices. While there is still much yet to be done in hard data fusion, soft data fusion continues to lag far behind, due to the fact that different processing, modelling and storage methodologies are needed to support it. In this paper, we discuss many of the hurdles which still exist in soft data fusion, with a particular focus on the types of uncertainty involved.

1.0 INTRODUCTION

The Cambridge Business English Dictionary defines soft data as "information about things that are difficult to measure such as people's opinions or feelings."[1] Objectivity defines it as "information that is susceptible to interpretation and opinion."[2] Within the fusion community soft data is usually described as data collected from humans in the form of text (natural language) rather than the "hard" data from devices (sensors).

The fusion of hard data has been well studied over the past decades; while there is still much work to be done, hard fusion is quite mature. In comparison to hard fusion, work on soft data is still quite immature. There have been numerous areas in which some success has been reached, for example, sentiment analysis of social media has proven useful in gauging the mood of large groups of individuals. Much work has been made in the further development of text analytics, in which data such as named entities, relationships, and other tidbits are extracted from sources such as news articles, documents, social media and blogs. While progress has been made on a variety of fronts, there are still a great many hurdles to be cleared before fusion systems can fully exploit soft data. Many of the problems are clearly stated in the definitions quoted above: soft data may be "difficult to measure" or "susceptible to interpretation and opinion". Other problems include the fact that human "sensors" do very "non-device" things such as passing on information from others (hearsay), speculating about the future, conjecturing about what they have seen or heard tell half-truths, lie, etc. – all of which make affect the quality of the data derived.

In this paper we will look at a number of the issues involved in extracting and using soft data. In Section 2 we will look at the overall types of uncertainty in fusion, with emphasis in the discussion on soft fusion. Section 3 focuses on the human as sensor. Section 3 presents a general discussion of uncertainty in soft data, while Sections 4 and 5 examine more deeply the linguistics aspects of uncertainty in soft data, namely uncertainty within the data and uncertainty about the data. Section 6 provides a short conclusion.

2.0 TYPES OF UNCERTAINTY

Kruger et al. [3] list five types of uncertainty in the fusion process:

1. *Source uncertainty*: how reliable is the source?
2. *Content uncertainty*: how trustworthy is the content?
3. *Correlation uncertainty*: how certain is it that various data are related?
4. *Evidential uncertainty*: how strongly is the data supportive of a specific threat (hypothesis)?
5. *Model uncertainty*: even with all factors present, how certain is it that the model mirrors reality?

The first two, *source uncertainty* and *content uncertainty*, are concerned with uncertainty at the *data level*: how much do we trust the source of the information received, and how trustworthy do we feel this information is? These two are in many cases intimately interconnected: CIA analyst Richards J. Heuer, Jr. commented: “sources are more likely to be considered reliable when they provide information that fits what we already think we know.”[4] Not only is a source considered more reliable when delivering information which fits the analyst’s conjecture, but the information itself is often considered more trustworthy when it fits the analyst’s preconceived notion. (Heuer discusses this confirmation bias in some detail in [4].) On the other hand, even the most reliable source, such as an experienced intelligence officer, may speculate or make assumptions, so even the best source still relates uncertain information.

Correlation uncertainty and *evidential uncertainty* are focused at the *fusion* level. Correlation uncertainty with soft data may involve trying to assign a degree of confidence to the belief that two items of soft data are related: for example, whether the fuzzy expression “several people” and the more specific “five combatants” are referring to the same group of individuals. Gross et al. [5] provide some excellent discussions about the problems of correlating soft data. Evidential uncertainty plays a role when it is possible that the observed action is indicative of more than one threat; for example, the purchase of 50 kg of chemical fertilizer in, say, Afghanistan may point to either the construction of an IED (a direct threat) or the cultivation of opium (an indirect threat).

Model uncertainty is focused at the “reality check” level. Much of the work that intelligence agencies, both military and civilian, can be described as the equivalent of trying to determine the shape and size of an iceberg based upon what one can observe above the water. This is because, in general, the enemy is quite skilled at obscuring his activities from observation and only very indirect clues may be observable, and that these clues can possibly reflect a different reality: using the chemical fertilizer analogy, it may well that the purchaser of a large quantity of chemical fertilizer, who happens to come from the same village as a person of interest, may not be involved in a terrorist bomb plot but instead will be fertilizing his parents’ wheat fields on the weekend. In order to reduce the chance that innocents are not mistakenly falsely targeted, there needs to be a mechanism to indicate how often the presence of most or all of the observable indicators in truth results in a true threat situation.

In the following sections we will focus on some of the ways in which natural language, and therefore soft data written in it, is uncertain, and how this uncertainty affects soft data that we use for fusion.

3.0 THE HUMAN AS SENSOR

Humans excel at various activities such as complex pattern recognition and the ability to examine information to arrive at new conclusions. However, in contrast to devices, humans as sensors are problematic on numerous levels: for example, they cannot be tested and calibrated under various conditions to gauge their reliability, they -- intentionally or unintentionally -- self-filter the information that they pass on, and the "signals" which humans do pass on (in natural language) are themselves problematic for any of a variety of reasons, as we will see.

The sources of soft data include intelligence personnel, informants, prisoners of war, and local residents in the area of operations as well as open sources such as newspapers, government documents, blogs, and social media. One of the significant differences to hard data generated by sensors is that soft data is symbolic, subject to interpretation, and often reflective of the individual's background, knowledge level, experience, cultural environment and other factors which may be difficult to accurately decode. A second significant difference is that the information is not limited to "historical" information; whereas a sensor always delivers signal based upon phenomena that have already happened, a human sensor may, for example, report on things which *might* happen. Dragos and Rein [6] discuss at some length several of the aspects of using humans as information sources, including, but not limited to the list below:

- ***Subjectivity*** - Any event observed by two or more individuals will very often result in different reporting of that event, due to differing perception and interpretation of what happened. Perception does not necessarily provide a true and accurate representation of the physical world, since sensory input passes through the individual's perceptive and cognitive filters. For example, the perception of an event differs from one source to another according to their specific set of skills, knowledge, and emotional involvement. A trained observer will report an event quite differently than an untrained passer-by: in part, this may be due to the amount and type of detail which is recorded, as well as a difference of interpretation in what exactly is happening due to previous experience. A subjective-build reality arises when several different individuals report their observations and, instead of creating only one unique "truth", several different, perhaps conflicting, interpretations. These differing "realities" may confuse the understanding of what is actually happening. The unreliability of eyewitness testimony due to these types of factors is well documented.
- ***Intention*** - Unlike hard sensors, which may provide unreliable data based upon factors such as device failure or environmental conditions, humans may deliberately alter the information based upon intention, that is, through conscious efforts to fabricate, conceal or distort evidence for some reason. Such distortion may be the result of omission of important details or outright lies, that is, false stories invented with the intent to deceive. In other cases, false statements are mixed with true statements, or ambiguity is used in order to obscure the issue or protect the individual at a later point from charges of lying or deception. The intention of an informant is not necessarily always the result of malice or disinformation: he may, for example, provide false or misleading information which he believes the hearer wishes to hear, in order to win favor or gain attention, or in order to protect himself from personal negative consequences. Whether the basis is malicious or not, the intentional distortion of affects the quality of the information.
- ***Opinion*** - humans will not only filter or process observations of events that they have witnessed according to their perceptions of the world, they may offer opinions, make assumptions or judgments about that which they have observed. Often, opinions or judgments rest on ground insufficient to produce certainty, and are completely dependent on the source providing them. Beyond the obvious subjective dimension of opinions, mixing factual information and personal impressions or estimations raises specific problems for further automatic processing of data. This is not limited to untrained individuals: even trained intelligence officers will do this: reporting an observation of a group of individuals observed in the field may be accompanied by speculation about whether this group may be hostile or not, and opinion as to what their intentions may be.
- ***Hearsay*** - Another inherent characteristic of human communication which contributes to uncertainty of information is hearsay. Rather than offering personal observation or opinion, the source passes information which he (claims) to have received from another source. Hearsay is problematic from a variety of perspectives. For example, hearsay information may have been

(intentionally) removed from its original context, thereby subtly or dramatically changing the original meaning of the information; this "cherry picking" of tidbits of the truth may be done to distort. A second problem is that the reporter may pass on hearsay information that has come through a chain of individuals before arriving at the reporter: "my brother told me that his wife's sister's husband talked to a guy who attended the meeting and said…". With each further telling, distortion, misunderstanding, (mis)interpretation and distance from the original context may increase, with the result that the resulting hearsay information may, as in Chinese whispers, be significantly altered from the original.

It should also be noted here that in particular with open sources such as online news sources, blogs and social media, the hearsay problem can be acute. This is not just due to obvious phenomena such as "going viral" by re-tweeting or re-posting from other sites. On-line news sources often reference and link to articles on other sites, often with accompanying commentary. Should, however, the original source print a retraction or correction, the sites which have commented on the original seldom update their own text to reflect the changes, leaving the incorrect information in place.

- ***Hidden networks*** – Sources may have undisclosed ties to one. Individual sources (nodes) of a network may be connected on the basis of similarities (same location or similar attributes), social relations (friendship) or interactions between nodes, not all of which may be known to or easily discoverable by intelligence. Regardless of their nature, these linkages may result in undetected connections of information pieces provided by different sources. As a result, such information pieces will be analyzed under the false assumption of there being several independent sources, with impact on the accuracy of the outcome. One of the methods used by intelligence professionals to assess the veracity of information has to do with its having been independently reported by multiple sources. Thus, the existence of a hidden network may degrade information credibility, as what appears to be independent reporting may actually be hearsay.

While vital to both the military and the civilian intelligence communities, the human as a sensor, as we have seen, is not unproblematic. Using human as information sources requires acknowledgement of the psychological, social and emotional aspects of human-derived information referred to above. Without careful consideration and focused approaches to dealing with these aspect, the outcome of soft fusion data would be unreliable intelligence products which are at a minimum not useful to daily operations, but at the extreme costly and dangerous.

Even when we have concrete information about the level to which we can trust our source, there are yet other elements in the soft data that signal uncertainty. We will discuss these in the following sections.

4.0 OVERVIEW OF UNCERTAINTY IN SOFT DATA

> *"…natural language sentences will very often be neither true, nor false, nor nonsensical but rather true to a certain extent and false to a certain extent, true in certain respects and false in other respects"* George Lakoff [7]

When humans communicate, we do more than convey facts. We express thoughts, hopes and wishes, we speculate about the future, we pass on information that others have communicated to us; we tell lies and half-truths to elicit cooperation, to be accepted as part of a group, to win approval from others or to evade censure. Even when we are in fact passing on concrete information, we don't necessarily deliver it in clear, concise and precise wording. We may use words that have multiple meanings or formulate our sentences so that they are ambiguous. There are a variety of ways in which soft data can be uncertain which we will examine in this section.

Liddy et al. [8] have developed a model which touches on some of the aspects – hearsay (which they refer to as "point of view"), opinion (under "focus") – which we discussed in the previous section which have to do with the use of humans as sources of information. They introduce a significant dimension in soft data, namely ***time***, that is a dimension significantly different to that of hard data. Sensors report only historical data, the events which are recorded have in fact taken place. Algorithms which operate on that data may project into the future. In the case of soft data, the sensor itself reports (possible) future events, for example, to plot a possible path for incoming threats. Since much intelligence work is focus on prevention of future threats, the use of speculative information about future events can be both useful and desirable. However, information concerning possible future events is of its nature uncertain: the event may not happen, the meeting may not take place, the person in question may not act as speculated. The use of prediction of future events must be handled with care: at the point in time where the expected event has or has not taken place, a decision must be made to keep, revise or jettison that data.

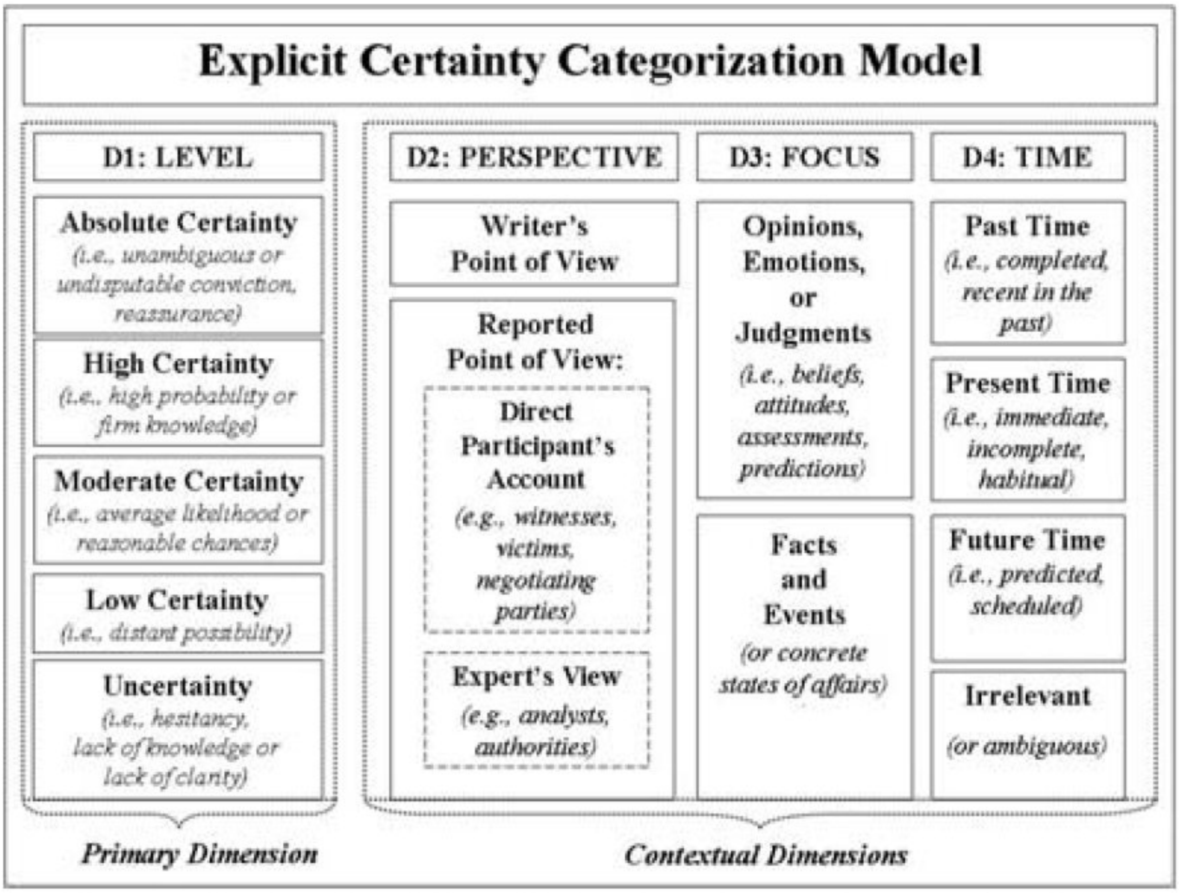

Explicit Certainty Categorization Model
D1: LEVEL
Absolute Certainty (i.e., unambiguous or undisputable conviction, reassurance)
High Certainty (i.e., high probability or firm knowledge)
Moderate Certainty (i.e., average likelihood or reasonable chances)
Low Certainty (i.e., distant possibility)
Uncertainty (i.e., hesitancy, lack of knowledge or lack of clarity)
D2: PERSPECTIVE
Writer's Point of View
Reported Point of View:
Direct Participant's Account (e.g., witnesses, victims, negotiating parties)
Expert's View (e.g., analysts, authorities)
D3: FOCUS
Opinions, Emotions, or Judgments (i.e., beliefs, attitudes, assessments, predictions)
Facts and Events (or concrete states of affairs)
D4: TIME
Past Time (i.e., completed, recent in the past)
Present Time (i.e., immediate, incomplete, habitual)
Future Time (i.e., predicted, scheduled)
Irrelevant (or ambiguous)
Primary Dimension
Contextual Dimensions

Figure 1: Categorization Model of Liddy et al. [8] showing different dimensions of uncertainty in text information.

In their 2008 article [9], Auger and Roy discuss uncertainty within the framework of soft data. They divide uncertainty in soft data, shown in Figure 2, into two main categories: linguistic ambiguities and referential ambiguities.

Under “referential ambiguities” we can recognize a few aspects (e.g., intent, cultural influences) which were discussed in the preceding section concerning the human as a source. Here one should note the sub-branch “contextual elements” includes a leaf which is specific to spoken as opposed to written natural language communication, namely “body language.” This plays a role when information is conveyed in spoken rather than in written form; body language often provides information that helps us to assess the reliability of the information being transmitted: body language that conveys unease may be a sign of subterfuge. Additionally, when we are dealing with written text, we may not have any idea as to context elements such as weather, mood, situation, etc., that the writer is experiencing at the time of reporting, although there is a possibility that these elements may affect the data we receive. For example, exhaustion or illness may reduce the observational abilities of a human reporter, and therefore the data reported is less reliable and possibly error-prone. Unlike environmental factors such as temperature or weather, we may not be able to detect such contextual elements which affect the sensor.

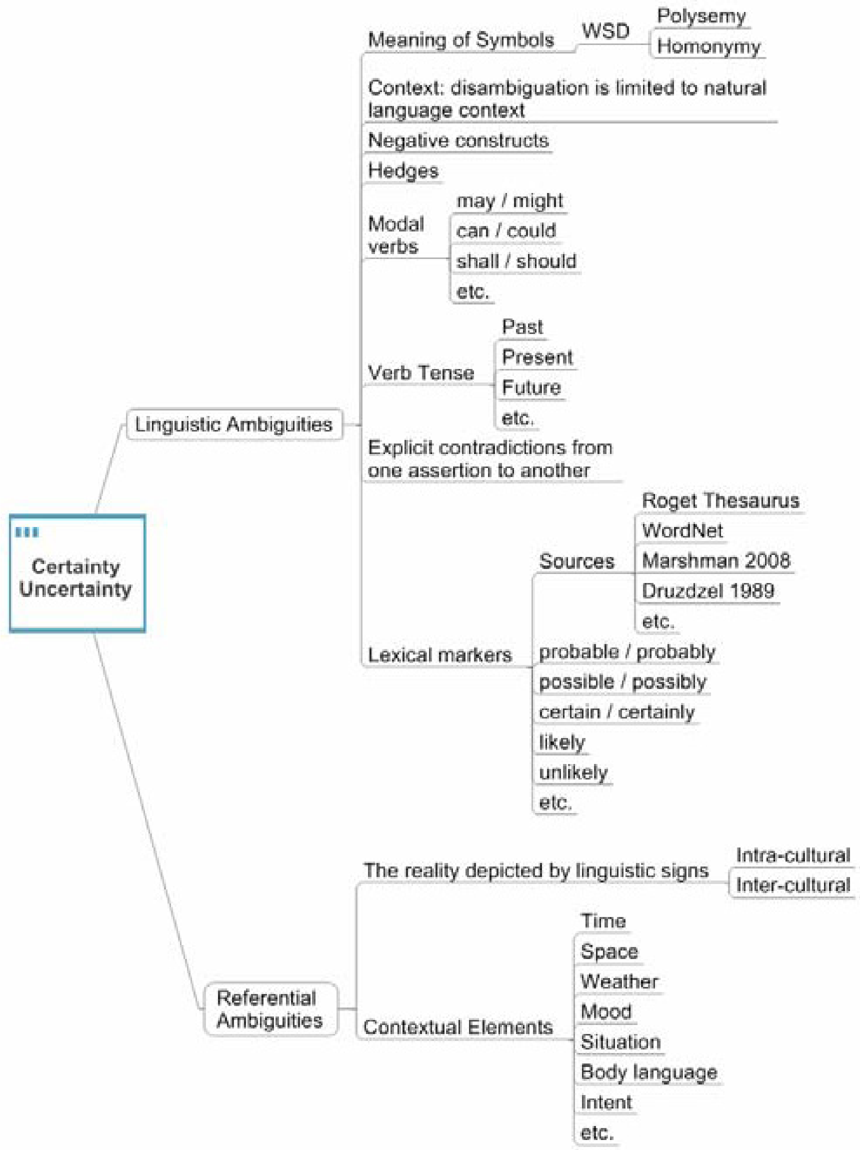


Figure 2: Auger and Roy [9] divide uncertainty in linguistic data into two broad categories: linguistic ambiguities and referential ambiguities.

As shown in the figure linguistic ambiguity is tied to the meaning of the words (symbols) used in a given language, while referential refers to cultural or other contextual elements. A number of the elements listed under the category of linguistic ambiguity have to do with things that, while important, may be difficult to resolve. For example, "explicit contradictions from one assertion to another" may ultimately only be resolved too late to be of benefit.

However, a number of the other elements are things which can be easily identified and exploited by computer to provide an initial assessment of the certainty of the information which we receive from human sources. For example, as previously discussed, we can identify the verb tense to differentiate between events or states which have occurred in the past as opposed to events or states which may or may not take place at some point in the future. Other words or phrases also give us clear signals: modal verbs ("might", "could", etc.), lexical markers such as modal adjectives ("possibly", "likely", etc.), or indicators of hearsay or opinion ("Sources said", "I believe", etc.). All of these should be taken into consideration when determining the quality of the data pulled from text.

To summarize, there are two basic categories of detectable uncertainty which appear at the sentence level within written text or in speech:

- Uncertainty ***within*** the content, including:
    - Imprecision;
    - Vagueness; and
    - Ambiguity and polysemy.
- Uncertainty ***about*** the content, including:
    - Modal verbs;
    - Modal adverbs (including "words of estimative probability");
    - Hearsay markers;
    - "Mindsay" markers → belief, inference, assumption, etc.; and
    - Passive voice.

Referring again back to the five types of uncertainty discussed in Section 2, the uncertainty ***about*** the data is content uncertainty (data level),while uncertainty ***within*** the content comes into play at the correlation of various discrete elements of data (fusion level).

In the following sections we will examine these in more depth.

5.0 UNCERTAINTY WITHIN THE CONTENT

When humans communicate, we do more than convey facts. We express thoughts, hopes and wishes, we speculate about the future, we pass on information that others have communicated to us; we tell lies and half-truths to elicit cooperation, to be accepted as part of a group, to win approval from others or to evade censure. Even when we are in fact passing on concrete information, we don't necessarily deliver it in clear, concise and precise wording. We may use words that have multiple meanings or formulate our sentences so that they are ambiguous. There are a variety of ways in which soft data can be uncertain which we will examine in this section.

5.1 Imprecision and Vagueness

Text from which soft data is often formulated in ways that obscure, however unintentionally, details that may be useful in the fusion process.

Take, for example the following sentence:

1) *There were some animals in the road.*

"*Some*" is an imprecise number. The reader might possibly make some judgments on the range of numbers represented by "*some*," although this may be bounded by other equally imprecise values. For example, the reader might assume the speaker would have selected "*a couple*" as a descriptor if there were only two or three animals, that "a bunch" would have used for a dozen or so, and that "*many*" would have been preferred if there were a noticeably larger quantity, such as a twenty or fifty. "*Several*" would be more than "*a couple*" – perhaps five or six – but generally would not be considered to be "*many*." However, "*some*" simply implies "*multiple*" without any further hint as to how many, so we can only guess.

Additionally, it is not clear what sort of animals these may be: cats? dogs? cows? elephants? Or possibly it was a mixture of different types (a sheepdog and five sheep, for example). One type of animal would most likely not be considered in the count here: a human. In that case, "*someone*" or "*some people*" would have been used in place of "*animals*." However, statement 1) does not necessarily leave out the presence of a human: for example, when a herd of sheep are in the road, the shepherd is usually somehwere in the vicinity as well, but the animals, not the shepherd, would likely be considered as some sort of anomaly worth mentioning. Similarly, a high level of precise detail may also be inaccurate: even if we were told that there were six brown Jersey cows in the road, it may well be that the observer neglected to let us know that there were two black herding dogs as well, or failed to detected that one brown cow was in fact a Hereford and not a Jersey.

Correlating information about a shepherd and his dogs moving his cattle to new pasturage with "some animals in the road" requires understanding of a variety of things, including that dogs and cattle are animals, that "some" means "multiple" and, perhaps, figuring out if the road in both cases was the same, or at least in the same general vicinity.

To complicate things further, many such vague or imprecise formulations may be context or domain dependent. For example, while it is generally understood that "*large*" is an adjective related to size of an object, its exact (quantifiable) meaning is extremely domain dependent. There are many orders of magnitude difference in the numerical values of indicated by "*large*" between a *large city*, a *large ship*, a *large dog* and a *large molecule*.

Furthermore, even within the same domain, there may be variations due to other factors such as context information. For example, the phrase "a lot of people" will generate a different numerical range depending on expectation or physical factors such as facility size. If a smallish meeting room is filled to standing room only, it will be reported that the 50 persons attending the event were "a lot of people." However, those same 50 persons would not be classified as "a lot of people" if they are sitting in a 500-seat auditorium, and would be completely insignificant within the context of a 30,000-seat sport stadium. Therefore, the decision about the numerical range represented relies on what we know about the location. Gross et al. [ref] have at some length about such problems and their resolution.

5.2 Ambiguity and Polysemy

Statements may be ambiguous, i.e., they may be open to more than one interpretation or have more than one possible meaning:

2) *Students hate annoying professors*.

3) *Sally gave Mary her book*.

In 2) either the students strongly dislike professors who irritate them, or whether students try to avoid making their professors angry, perhaps in the hope of receiving a better grade in the course. In 3) we do not know whether the book that Mary received from Sally was Sally's book, or whether Sally was returning Mary's own book to her – or even if there may yet another female involved. For example, it is possible that in a preceding sentence, the writer informs us that Jane is a well-known author and a friend of Sally's, and thus 3) tell us that Sally presented Mary with a copy of Jane's latest hit novel.

Another example of ambiguity is shown in 4):

4) *I saw her duck.*

In 4) it is unclear whether the female person referred to has lowered her head to avoid, say, a low-hanging branch, or whether she keeps a waterfowl as a pet. This ambiguity stems from what is called ***polysemy***, the fact that the same word (label) can refer to two or more separate concepts. Other examples of polysemy are words such as “bank” which could be a financial institution, the side of a river, or a motion executed by a flying aircraft, to name just a few of the many meanings. Determining which meaning is meant may be determined by analysis of the surrounding text, but, as in the case of 4), this is not always possible.

Regardless of the lack of detail in all of the examples above, there is nothing in any of these sentences to make us believe that the sentences are not true. However, very often sentences contain clues which the writer uses to signal to us there may be reason to doubt the veracity of the *content* contained in the sentence. These we will discuss in the following section.

6.0 UNCERTAINTY ABOUT THE CONTENT

In the preceding section we used the following sentence as an example of ambiguity:

2) *Sally gave Mary her book.*

While our previous confusion about the book continue, we have no reason to believe that this event did not take place. However, suppose the sentence read as follows:

5) *It is possible that Sally gave Mary her book.*

The confusion as to the owner of the book continues to exist, but now we are no longer certain as to whether indeed the event of Sally giving Mary a book exists. Perhaps it was Georgina who gave Mary the book, or perhaps it was Mary who gave Sally the book. The presence of “it is possible” changes the credibility of the event significantly. Natural languages are filled with a variety of different mechanisms which inject some uncertainty into the soft data they convey; analysis of these mechanisms will support fusion of soft data in that we may better assess the quality of the data which we are using.

6.1 “Words of Estimative Probability”

In a 1964 article, Sherman Kent of the United States Central Intelligence Agency relates the following anecdote about an intelligence report concerning the possibility of a Soviet invasion of Yugoslavia:

> A few days after the estimate [“NIE 29-51, "Probability of an Invasion of Yugoslavia in 1951”] appeared, I was in informal conversation with the Policy Planning Staff's chairman. We spoke of Yugoslavia and the estimate. Suddenly he said, "By the way, what did you people mean by the expression `serious possibility'? What kind of odds did you have in mind?" I told him that my personal estimate was on the dark side, namely, that the odds were around 65 to 35 in favor of an attack. He was somewhat jolted by this; he and his colleagues had read "serious possibility" to mean odds very considerably lower. Understandably troubled by this want of communication, I began asking my own colleagues on the Board of National Estimates what odds they had had in mind when they agreed to that wording. It was another jolt to find that each Board member had had somewhat different odds in mind and the low man was thinking of about 20 to 80, the high of 80 to 20. The rest ranged in between. [10].

What makes this anecdote of particular interest is that the various individuals with whom Kent spoke were all intelligence analysts, that is, people who were working in the same domain (intelligence), who most likely had similar educational backgrounds and, presumably, who also had had similar training for their jobs. This anecdote is significant in that it shows us that, in spite of similarities in backgrounds, the understanding of such terms can be quite diverse.

Intrigued by Kent's observation, another CIA analyst Heuer [11] ran an informal study, asking a number of his CIA colleagues to assign a single probability to a number of common expressions used by the analysts. Figure 3 shows the mapping of the various probabilities assigned to each expression.

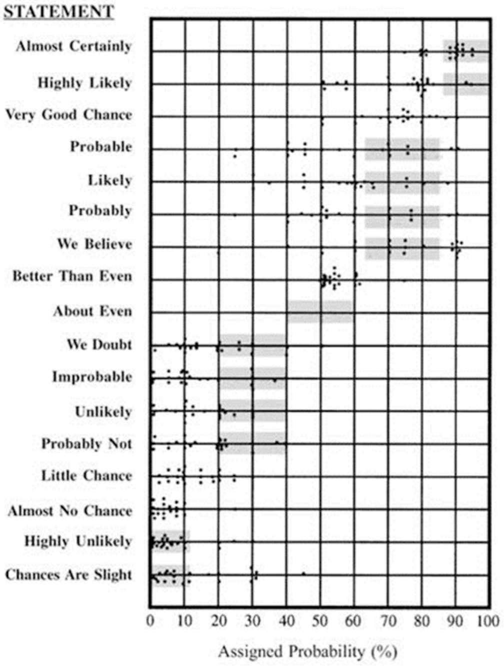


Figure 3: Results of Heuer's informal study mapping probabilities onto various expressions of uncertainty [11].

The probabilities assigned to a number of the terms of uncertainty were clustered very closely (*better than even*, *about even*, *highly unlikely*). A number varied quite dramatically: *highly likely* ranged more than 40 percentage points, as did *improbable*, *probably not* and *chances are slight*, while the range for *probable* was from 25% to just over 90%.

A few years later, and still within the analyst realm, Rieber [12] requested analysts training at the Kent School (named after Sherman Kent) to assign ranges of percentages instead of specific values to a number of hedges. The results are shown in Figure 4 below.

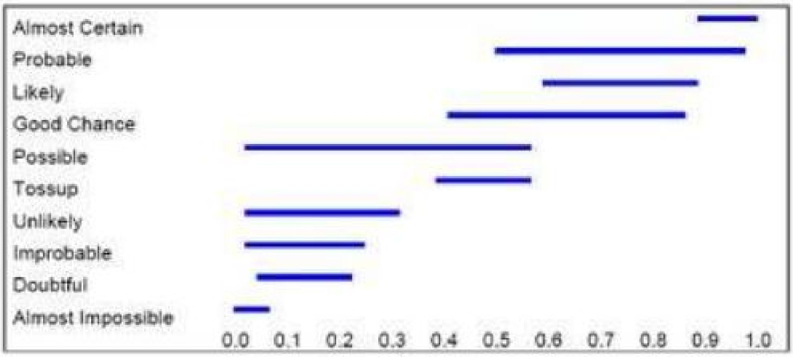


Figure 4: Ranges of percentages assigned to hedges by analysts in training [12].

Again one can see that the ranges of percentages range from quite narrow to relatively large, but the ranges are not necessarily identical to those in the first chart, even for identical hedges (compare *probable* in both). One can almost assume that giving the task of assigning probabilities for hedges to any random group of English-speakers will result in somewhat different numerical ranges.

In the decades since Kent's initial work, the US intelligence community continued to struggle to standardize the terminology which they used to assess situations, in order to reach a common understanding of the meaning of those terms.

Ultimately, the intelligence communication settled on a standard spectrum of WEPs as shown in Figure 5 below.

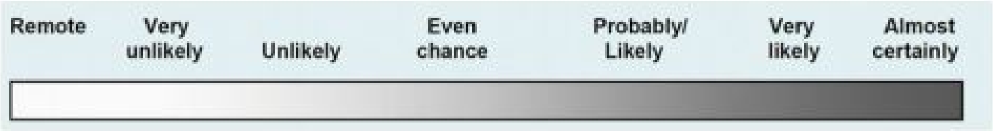


Figure 5: Words of Estimative Probability, as appeared in the 2007 National Intelligence Estimate, Iran Nuclear Intentions and Capabilities as well as in the front matter of several other recent intelligence products (via Friedmann/Zeckhauser [13]).

The discussion above involves examples of uncertainty which are straightforward and obvious. However, there are other, less obvious forms of uncertainty which we will look at in the following sections.

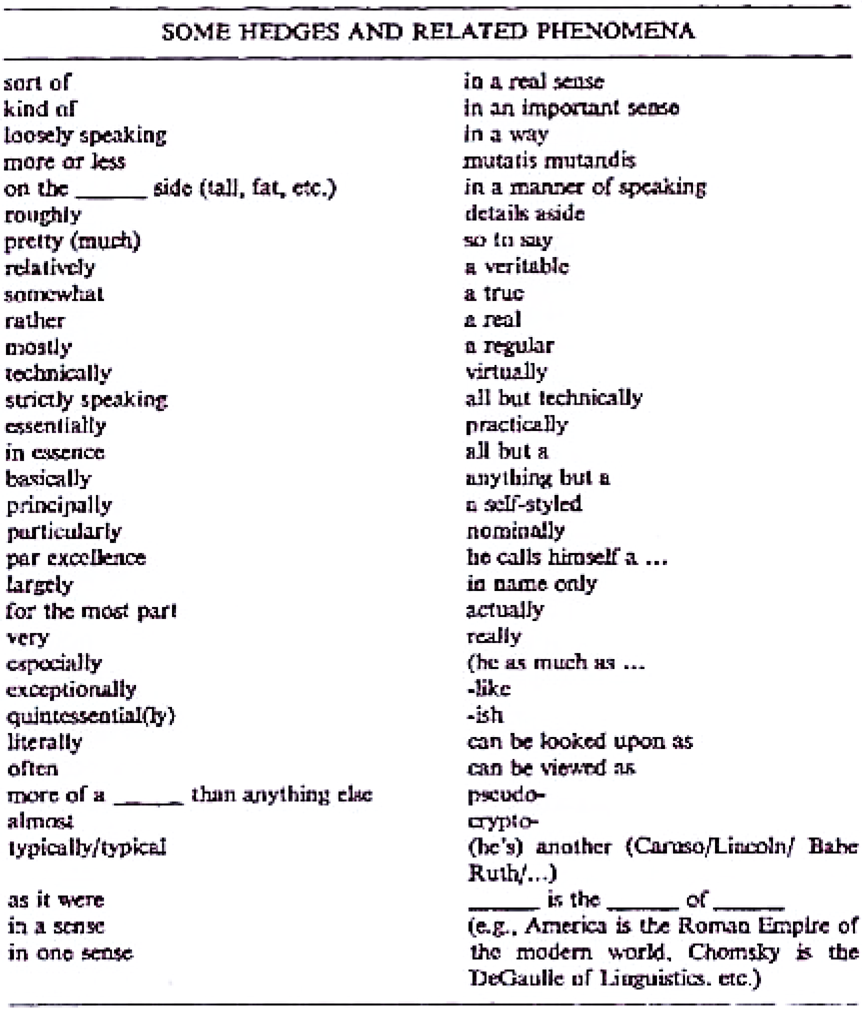
6.2 Hedges and Evidential Markers

In general, when one is asked to consider markers of uncertainty in natural language, the first group that comes to mind is modal adverbs as we have seen above: *possibly*, *probably*, *likely*, etc. The next categories are often modal verbs: *might*, *could*, *may*, etc., followed by nouns such as *likelihood*, *possibility*, *probability*, and so on. Lexical verbs like *suggest*, *assume*, *seem*, *guess*, etc., likewise convey uncertainty, as do adjectives such as *possible*, *probable*, *doubtful*, etc. For many researchers, all of the elements on above manifestations of uncertainty are included in a category called ***hedges***.

The term "hedges" is attributed to Lakoff [7], who used this term to mean any lexical or grammatical form which indicates "fuzziness" in natural language. Fascinated by the mathematical theories of Zadeh, he defines a broad spectrum of lexical and grammatical elements in natural language which indicate any softening of the formulation of propositions, which express vagueness or imprecision:

> For me, some of the most interesting questions are raised by the study of words whose meaning implicitly involves fuzziness – words whose job is to make things fuzzier or less fuzzy. I will refer to such words as 'hedges' [7].

Figure 6 below shows a list of some elements that he considers hedges; it should be noted that since his focus in this initial paper was on category membership (prototype theory), "related phenomena" appear to be such elements as "a real" or "a regular" or "____is the ____ of ___" which are represent, for example, analogies.



SOME HEDGES AND RELATED PHENOMENA

sort of
kind of
loosely speaking
more or less
on the ______ side (tall, fat, etc.)
roughly
pretty (much)
relatively
somewhat
rather
mostly
technically
strictly speaking
essentially
in essence
basically
principally
particularly
par excellence
largely
for the most part
very
especially
exceptionally
quintessential(ly)
literally
often
more of a ______ than anything else
almost
typically/typical
as it were
in a sense
in one sense

in a real sense
in an important sense
in a way
mutatis mutandis
in a manner of speaking
details aside
so to say
a veritable
a true
a real
a regular
virtually
all but technically
practically
all but a
anything but a
a self-styled
nominally
he calls himself a ...
in name only
actually
really
(he as much as ...
-like
-ish
can be looked upon as
can be viewed as
pseudo-
crypto-
(he's) another (Caruso/Lincoln/ Babe Ruth/...)
______ is the ______ of ______
(e.g., America is the Roman Empire of the modern world, Chomsky is the DeGaulle of Linguistics, etc.)

Figure 6: Hedges and related phenomena [7].

Since Lakoff's first article, the definition of hedging has shifted to focus more narrowly on expressions of uncertainty or commitment on the part of the speakers. Some researchers consider modals verbs (*could*, *should*, *might*, etc.) to be hedges, others not (in Figure 2 above, Auger and Roy list them as separate categories).

But hedges (used in the broadest sense) are not the only elements which signal uncertainty in text information. There are also markers that indicate that where the information contained in the sentence comes from. These are called "evidential markers" which can be divided into two broad categories: *hearsay* and what Bednarek [14] refers to as "*mindsay*."

Hearsay, that is, information which the writer has gotten from another source (not himself) is uncertain by nature in that we can never be certain that the writer has correctly and fully understood what the original source said, and therefore we cannot be certain that the information that has been passed on is reliable.

Mindsay is information which comes from the original source but is based upon belief, speculation, assumption – in other words, not on fact or observation, but as a product of some process in the source's mind.

Take, for example, the following sentences:

6) John is a terrorist.

7) The CIA have reported that John is a terrorist.

8) I believe that John is a terrorist.

9) Mary thinks that John is a terrorist.

In each of these sentences, the relationship ("fact") pattern of the sentence might produce the relation *John IS-A terrorist*. In 6) there are no lexical clues as to what the writer believes nor where the information comes from, so the basis of our judgment as to whether John is, in fact, a terrorist must come from, say, previous knowledge.

However, there is other information contained in nearly all of these sentences which gives us a reason to doubt the veracity of the "fact" of John being a terrorist -- the lexical clues surrounding this "fact" weaken the belief in its veracity. In 7) and 9) there are clear indicators of third party information, i.e., *hearsay*, which may or may not have been repeated accurately by the writer.

In 8) and 9) indicate belief, i.e., *mindsay*, rather than knowledge on the part of the sources, the first one is a first-person reporting of that belief, whereas the second is either third person reporting ("Mary told me"), or an assumption on the part of the writer that Mary believes this.

Sentence 9) is interesting not just because it is, in fact, ambiguous. In one interpretation, one could say that it contains both hearsay and mindsay: the writer informs us about something another person (Mary) has told him about her thoughts regarding John. A second interpretation could be that the writer expresses his belief (mindsay) about what Mary thinks (also mindsay).

A single sentence may contain multiple clues as the veracity of the main proposition of the statement. For example, consider the variations on sentence 7):

10) The CIA have concluded that John is probably a terrorist.

11) The CIA have concluded that John is most probably a terrorist.

In 10) adding the adverb "probably" to 7) weakens the "fact" of John being a terrorist, whereas in11) adding "most" before "probably" in 10) strengthens the assertion from 7), but it still remains weaker than in 7). If requested, based upon similar clues, an English speaker would be able to identify and rank assertions from strongest to weakest according to the clues the writer has left in the sentence.

Other factors that may be considered in the assessment of the strength or weakness of an uncertain proposition include such factors as to whether , for example, in hearsay an original source is named ("the CIA") as opposed to an unnamed source ("rumor has it") or general knowledge ("it is widely accepted"). The strength of mindsay may lie with the verb used, e.g.,"inferred" is stronger than "guessed".

Since most, if not all, decision-making models using information will use some sort of mathematical weighting system based upon the perceived certainty or doubt about the veracity of the data which populates the model. Frajzyngier [15] comments, “the different manners of acquiring knowledge correspond to different degrees of certainty about the truth of the proposition.” Models designed for device-based information such as sensors, cameras, radar, etc., may use factors based upon testing, calibration, the influence of environmental factors such as light, heat or humidity to adjust of reliability of readings or to fine-tune results. The human-generated information, in contrast, which comes in as text or speech is often assessed by a human, who uses her understanding of various factors including the background knowledge domain of the information, heuristic or scientific models, or even just a “gut feeling” to evaluate the information and assign it some sort of credibility weight. In lieu of other information, lexical markers may also play a role in the assessment; the analyst may well assign a lower “truth value” to information tagged by the informant through hedges to be “doubtful” than to information considered as “highly likely”.

Whereas many hedges such as the “words of estimative probability” discussed in the preceding subsection tend to be relatively easy for humans to weight with a numerical value, hearsay and mindsay markers are less evident. One could easily argue that weighting of the information from different types of sources in such a hierarchy implicit. For example, while it is generally acknowledged that direct perception (e.g., “I saw”) is more reliable than conveying hearsay (“he told me”), there has been no attempt to associate the difference mathematically such that it could factor into the association of a reliability value for the information contained in the sentence. i.e., downgrade its reliability to reflect somewhat more doubt.

6.3 Passive Voice, Depersonalization, Time and Other Subtle Forms

Hedges and evidential markers are relatively obvious indicators of uncertainty, even to non-linguists. However, there are some more subtle ways in which uncertainty may appear.

In his discussion of hedging, Hyland [16] includes several other phenomena such as passive voice, conditionals (if-clauses), question forms, impersonal phrasing and time reference. Particularly in scientific writing, the use of passive voice and impersonal phrasing are widely, almost universally, used, conveying an undertone of “but I might be wrong or have overlooked something.” With regard to impersonal phrasing, Hyland writes:

> …the writer inevitably uses a wide range of depersonalized forms which shift responsibility for the validity of what is asserted from the writer to those whose views are being reported. Verb forms such as *argue*, *claim*, *contend*, *estimate*, *maintain* and *suggest* occurring with third person subjects are typical examples of forms functioning in the way, as are adverbials like *allegedly*, *reportedly*, *supposedly* and *presumably*.

Passive voice, however, can also be used to express politeness, rather than uncertainty, which can only be determined by knowing some information about the context of the statement. Likewise, passive voice is also sometimes used in the case of differences in social ranking or power, in order not to offend.

While time might not immediately spring to mind when considering expressions of uncertainty, it nevertheless plays a significant role, and should therefore be discussed.

Any sentence which is formulated in the future tense is inherently uncertain, simply because the event or state which is described has not happened yet:

12) Mary will be at the meeting next week.

That is, of course, unless Mary decides not to go, her plane flight is cancelled due to snowfall or she gets sick and lands in the hospital, or worse.

That being said, some future things are more certain than others:

13) The next presidential election in the US will take place in November 2016.

Clearly, unless something unbelievably catastrophic happens, there is virtually no chance that the elections will not take place, so this may be treated as a fact, rather than a possibility.

For intelligence purposes, information based upon future actions often plays a very significant role, but should nearly always be considered uncertain, until the expected date of that action has passed (and it has or has not occurred).

6.4 Which Language?

Last, but not least, one of the most obvious problems is quite straightforward, indeed almost trivial: there are a multitude of spoken and written natural languages. Even a "single" language such as English has a variety of regional variants: the Irish playwright George Bernard Shaw is credited with commenting that "England and America are two countries separated by a common language." This is not simply a matter of pronunciation or even spelling -- it is also a matter of vocabulary. The "biscuit" of a Brit is an American's "cookie" – and an American's "biscuit" more akin to an unsweetened British "scone." Here we have an instance of the same word being associated with two different concepts based upon the variant of English being used – an example of the polysemy discussed in Subsection 4.2. There are also some lexical differences: if you ask an American about a "lorry" you will get a blank look – for her, the object referred to is a "truck" and the word "lorry" does not exist in that variant. Here we have two separate words for the same object – in essence, synonyms, but only in a cross-variant sense. Therefore, it is essential to know which variant of English is being used. And, just to complicate things, there are nearly 7000 distinct according to the Linguistic Society of America, of which "only" 230 are spoken in Europe.[17] And, of course, the last complication has to do with irregularities that have to do with non-native speakers of a given language, which can also cause misunderstanding and communication problems.

Furthermore, within a given language there may be examples of polysemy that are domain-specific, that is, they may be very specific to certain subgroups of the native speakers of that language. For example, "POV" within the US military is generally understood to be "privately owned vehicle" (i.e., a soldier's own car), but within, say, the fiction community, writers often use "POV" to mean "point of view." Therefore, the meaning attached to the acronym will vary according to the context in which it appears.

Uncertainty in the form of misunderstanding or miscommunication is not an insignificant problem and particularly in multi-cultural or multi-lingual endeavors it may play a role.

7.0 CONCLUSION

As we have seen, there is a multitude of ways in which soft data is uncertain. Effective use of soft data in the fusion process is dependent upon understanding the types of uncertainty in that information and how uncertainty affects the fusion process. Textual markers of uncertainty exist in the documents, websites and other natural language sources which are being used for intelligence purposes, and can be exploited to ensure a realistic assessment of information quality. Algorithms which are designed to extract information from natural language text should be taking advantage of these markers to give at least an initial assessment of the credibility of information by exploiting this knowledge.

8.0 REFERENCES

[1] http://www.cse.buffalo.edu/~shapiro/Papers/fusion2012.pdf.

[2] http://dictionary.cambridge.org/dictionary/english/soft-data (downloaded Sept 2015).

[3] Kruger, K., Schade, U. and Ziegler, J. (2008) Uncertainty in the fusion of information from multiple diverse sources for situation awareness. *Proceedings of the 11th International Conference on Information Fusion*, July 2008, Cologne.

[4] Richards J. Heuer, Jr, *Limits of Intelligence Analysis*, Seton Hall Journal of Diplomacy and International Relations, Foreign Policy Research Center of the University of Pennsylvania, Elsevier Ltd, Winter 2005.

[5] Geoff A. Gross, Rakesh Nagi, Kedar Sambhoos, Daniel R. Schlegel, Stuart C. Shapiro, Gregory Tauer, Towards Hard+Soft Data Fusion: Processing Architecture and Implementation for the Joint Fusion and Analysis of Hard and Soft Intelligence Data.

[6] Dragos, V & Rein K. (2014) Integration of soft data for information fusion: pitfalls, challenges and trends. *Proceedings of the 17th International Conference on Information Fusion*, Salamanca, Spain, , July 2014.

[7] Lakoff, George (1973), Hedges: A study in Meaning Criteria and the Logic of Fuzzy Concepts, *Journal of Philosophical Logic*, Volume 2, pages 459-508. D. Reidel Publishing Co., Dordrecht, Holland, 1973.

[8] Elizabeth D. Liddy and Noriko Kando and Victoria L. Rubin (2004). Certainty Categorization Model, *The AAAI Symposium on Exploring Attitude and Affect in* Text AAAI-EAAT 2004, American Association for Artificial Intelligence, Stanford, CA, 2004.

[9] Alain Auger and Jean Roy (2008), Expression of Uncertainty in Linguistic Data, Proceedings of FUSION 2008, Cologne, Germany.

[10] Sherman Kent (1964), Words of Estimative Probability, https://www.cia.gov/library/center-for-the-study-of-intelligence/csi-publications/books-and-monographs/sherman-kent-and-the-board-of-national-estimates-collected-essays/6words.html, Central Intelligence Agency.

[11] Richards J. Heuer, Jr (1999), Psychology of Intelligence Analysis, Center for the Study of Intelligence, Central Intelligence Agency.

[12] Steven Rieber (2006). Communicating Uncertainty in Intelligence Analysis, http://citation.allacademic.com/meta/p100689_index.htmlf.

[13] Jeffrey A. Friedman and Richard Zeckhauser (2015). Handling and Mishandling Estimative Probability: Likelihood, Confidence, and the Search for Bin Laden}"Intelligence and National Security},vol.30, nr. 1, pp. 77-99.

[14] Monika Bednarek (2006). Evaluation in Media Discourse: Analysis of a Newspaper Corpus, Continuum, London.

[15] Frajzyngier, Zygmunt. 1985. Truth and the indicative sentence. *Studies in Language* 9.2.243-254.

[16] Ken Hyland (1998). Hedging in Scientific Research Articles, John Benjamins, Amsterdam/ Philadelphia.

[17] http://www.linguisticsociety.org/content/how-many-languages-are-there-world.