Lecture Series IST-155,Fall 2016

Aspects of uncertainty in natural language information

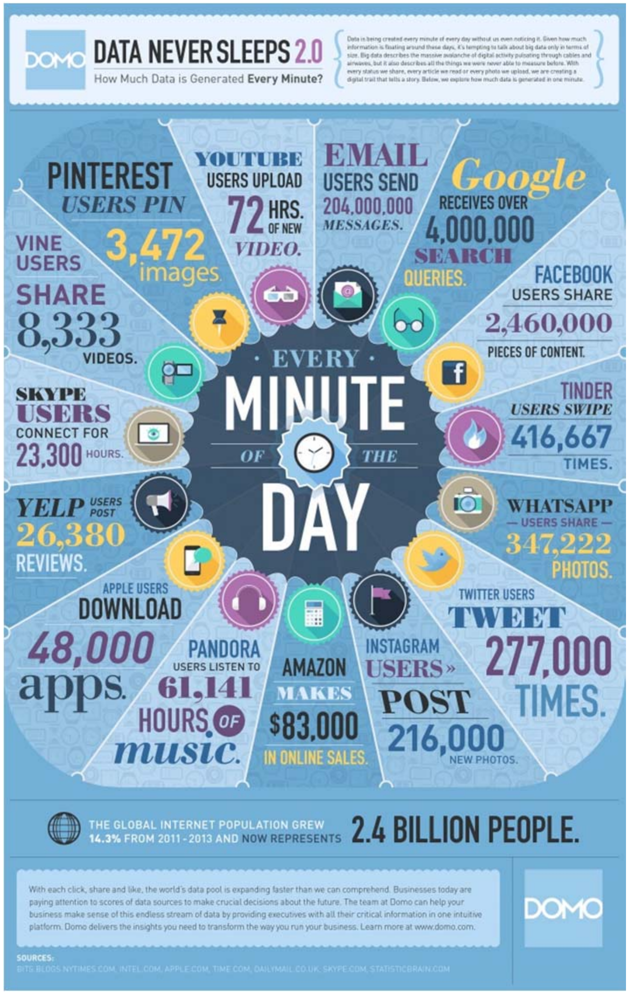
Kellyn Rein
Fraunhofer FKIE
Command and Control Systems
Information Analysis

Fraunhoferstr. 20, 53343 Wachtberg
GERMANY

kellyn.rein@fkie.fraunhofer.de

We are drowning in information but starved for knowledge.

John Naisbitt



Data is being created every minute of every day without us even noticing it. Given how much information is floating around these days, it's tempting to talk about big data only in terms of size. Big data describes the massive avalanche of digital activity pulsating through cables and airwaves, but it also describes all the things we were never able to measure before. With every status we share, every article we read or every photo we upload, we are creating a digital trail that tells a story. Below, we explore how much data is generated in one minute.

PINTEREST USERS PIN 3,472 images.
YOUTUBE USERS UPLOAD 72 HRS. OF NEW VIDEO.
EMAIL USERS SEND 204,000,000 MESSAGES.
Google RECEIVES OVER 4,000,000 SEARCH QUERIES.
VINE USERS SHARE 8,333 VIDEOS.
FACEBOOK USERS SHARE 2,460,000 PIECES OF CONTENT.
· EVERY · MINUTE OF THE DAY
SKYPE USERS CONNECT FOR 23,300 HOURS.
TINDER USERS SWIPE 416,667 TIMES.
YELP USERS POST 26,380 REVIEWS.
WHATSAPP — USERS SHARE — 347,222 PHOTOS.
APPLE USERS DOWNLOAD 48,000 apps.
TWITTER USERS TWEET 277,000 TIMES.
PANDORA USERS LISTEN TO 61,141 HOURS OF music.
AMAZON MAKES $83,000 IN ONLINE SALES.
INSTAGRAM USERS» POST 216,000 NEW PHOTOS.
THE GLOBAL INTERNET POPULATION GREW 14.3% FROM 2011 - 2013 AND NOW REPRESENTS 2.4 BILLION PEOPLE.
DOMO

With each click, share and like, the world's data pool is expanding faster than we can comprehend. Businesses today are paying attention to scores of data sources to make crucial decisions about the future. The team at Domo can help your business make sense of this endless stream of data by providing executives with all their critical information in one intuitive platform. Domo delivers the insights you need to transform the way you run your business. Learn more at www.domo.com.

SOURCES:
BITS.BLOGS.NYTIMES.COM, INTEL.COM, APPLE.COM, TIME.COM, DAILYMAIL.CO.UK, SKYPE.COM, STATISTICBRAIN.COM

"According to computer giant IBM, 2.5 exabytes - that's 2.5 billion gigabytes (GB) - of data was generated every day in 2012."

"Think of it this way—five exabytes of content were created between the birth of the world and 2003. ***In 2013, 5 exabytes of content were created each day.***" (italics added)

Volume of data produced every ***minute*** of every day (by early 2014!)

What is “soft data”?

- “...information about things that are difficult to measure such as people’s opinions or feelings.”

  The Cambridge Business English Dictionary

- “information that is susceptible to interpretation and opinion.” Objectivity

- 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).

”…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

Uncertainty in the fusion process

There are several types of uncertainty :

1. Source uncertainty (how reliable is the source?)
2. Content uncertainty (how reliable is the content?)
3. Correlation uncertainty (how certain is it that various reports are related?)
4. Evidential uncertainty (how strongly is our information indicative of a specific threat?)
5. Model uncertainty (even with all factors present, how certain are we that the model mirrors reality?)

Uncertainty in the fusion process

There are several types of uncertainty :

data level

1. Source uncertainty (how reliable is the source?)
2. Content uncertainty (how reliable is the content?)
3. Correlation uncertainty (how certain is it that various reports are related?)
4. Evidential uncertainty (how strongly is our information indicative of a specific threat?)
5. Model uncertainty (even with all factors present, how certain are we that the model mirrors reality?)

Uncertainty in the fusion process

There are several types of uncertainty :

1. Source uncertainty (how reliable is the source?)
2. Content uncertainty (how reliable is the content?)

data level

3. Correlation uncertainty (how certain is it that various reports are related?)
4. Evidential uncertainty (how strongly is our information indicative of a specific threat?)

fusion level

5. Model uncertainty (even with all factors present, how certain are we that the model mirrors reality?)

Uncertainty in the fusion process

There are several types of uncertainty :

1. Source uncertainty (how reliable is the source?)
2. Content uncertainty (how reliable is the content?)

data level

3. Correlation uncertainty (how certain is it that various reports are related?)
4. Evidential uncertainty (how strongly is our information indicative of a specific threat?)

fusion level

5. Model uncertainty (even with all factors present, how certain are we that the model mirrors reality?)

model level

Source uncertainty

"On the Internet, nobody knows you're a dog."

The Human as Sensor

- Humans excel at many things including:
  - complex pattern recognition and
  - to examine information and arrive at a conclusion
- However!
  - They cannot be tested and calibrated
  - They (intentionally and unintentionally) self-filter the information which they pass on
  - The information which is passed on is itself problematic for a number of reasons

The Human as Sensor - issues

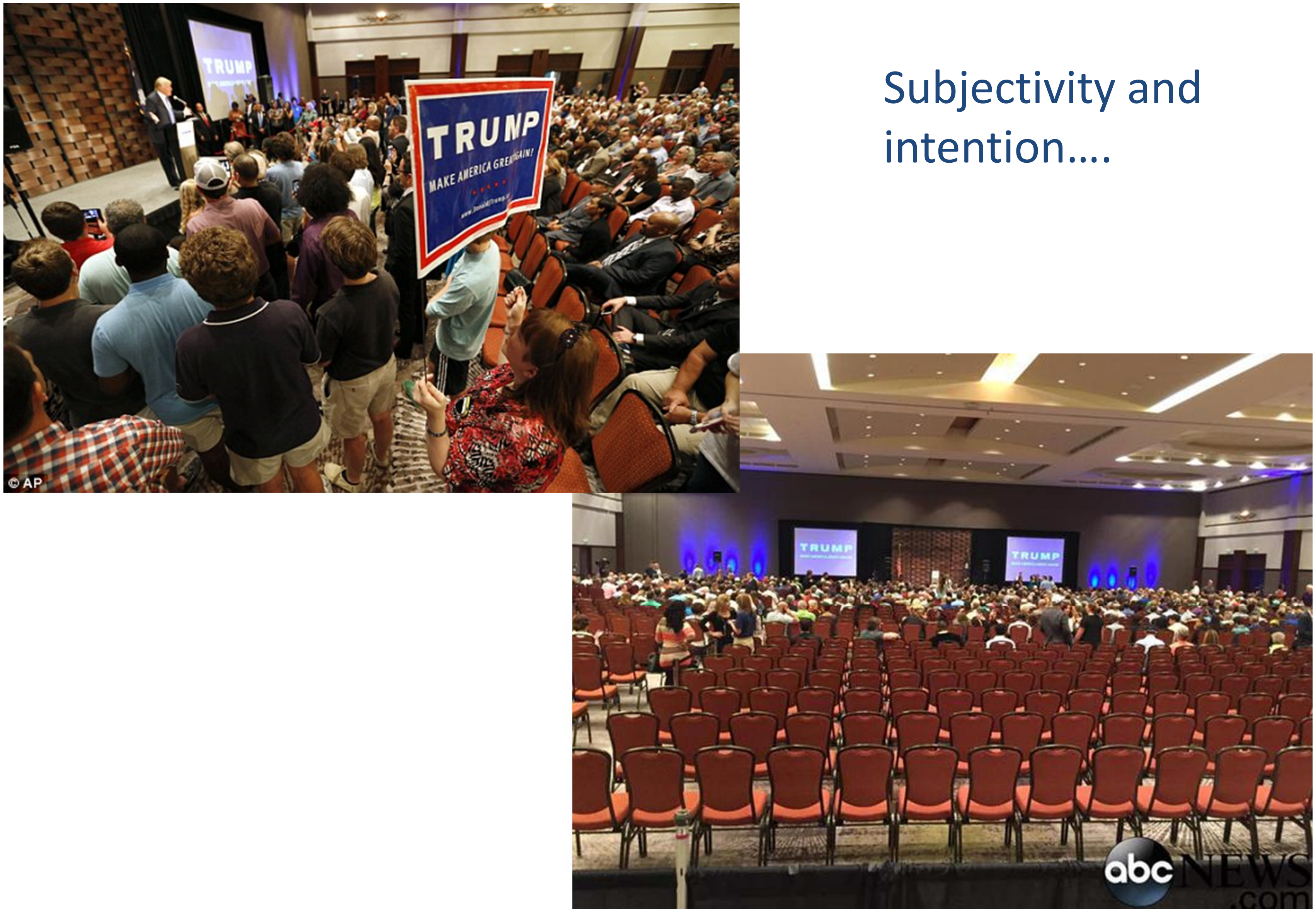
- **Subjectivity** – seldom will two individuals give exactly the same account of an event
  - interpretation of an event may be based upon perceptive and cognitive filters, experience, skills, background knowledge, etc.

The Human as Sensor - issues

- **Subjectivity** – seldom will two individuals give exactly the same account of an event
  - interpretation of an event may be based upon perceptive and cognitive filters, experience, skills, background knowledge, etc. (“eyewitness” unreliability)
- **Intention** – humans may deliberately alter information consciously to deceive or distort
  - Through omission of details (“cherry-picking”)
  - Mixing truth and lies (disinformation)
  - Outright lies (misinformation)
  - Saying what the speaker believes the hearer wants to hear (self-preservation, to win favor, “15 minutes of fame”)

Subjectivity and intention….

Subjectivity and intention….



The Human as Sensor - issues

- **Opinion** – offers judgment, assumptions, belief or opinions concerning events
  - An element of subjectivity, but the added dimension of adding interpretation or rationalization

The Human as Sensor - issues

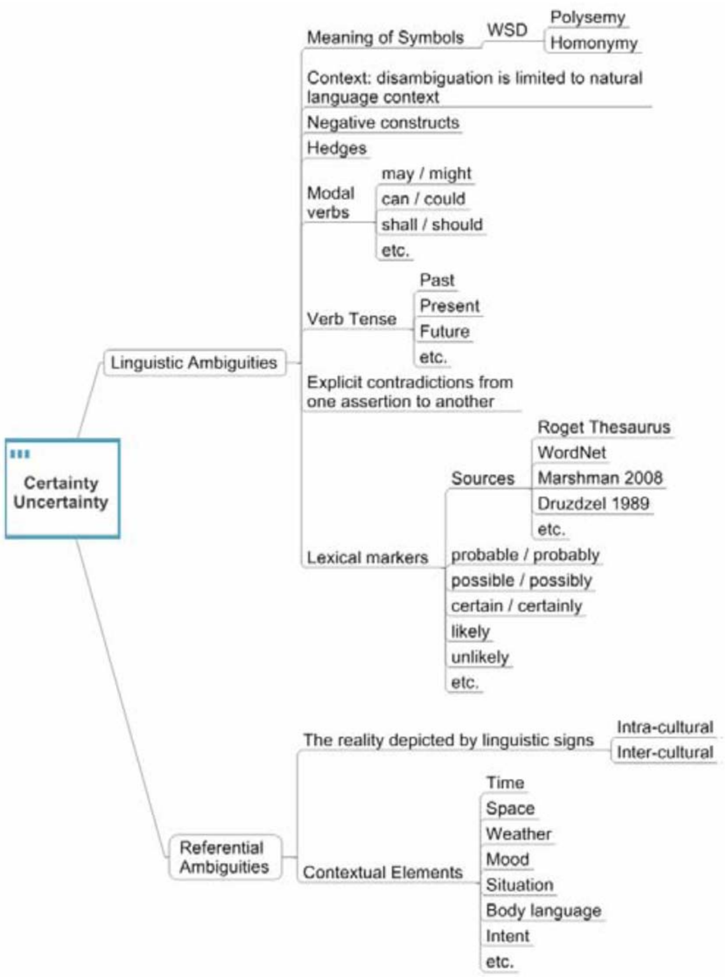
- **Opinion** – offers judgment, assumptions, belief or opinions concerning events
  - An element of subjectivity, but the added dimension of adding interpretation or rationalization
- **Hearsay** – humans pass on information derived from other sources
  - Who is the original source?
  - “Chinese whispers” effect (unintentional distortion via re-telling)

The Human as Sensor - issues

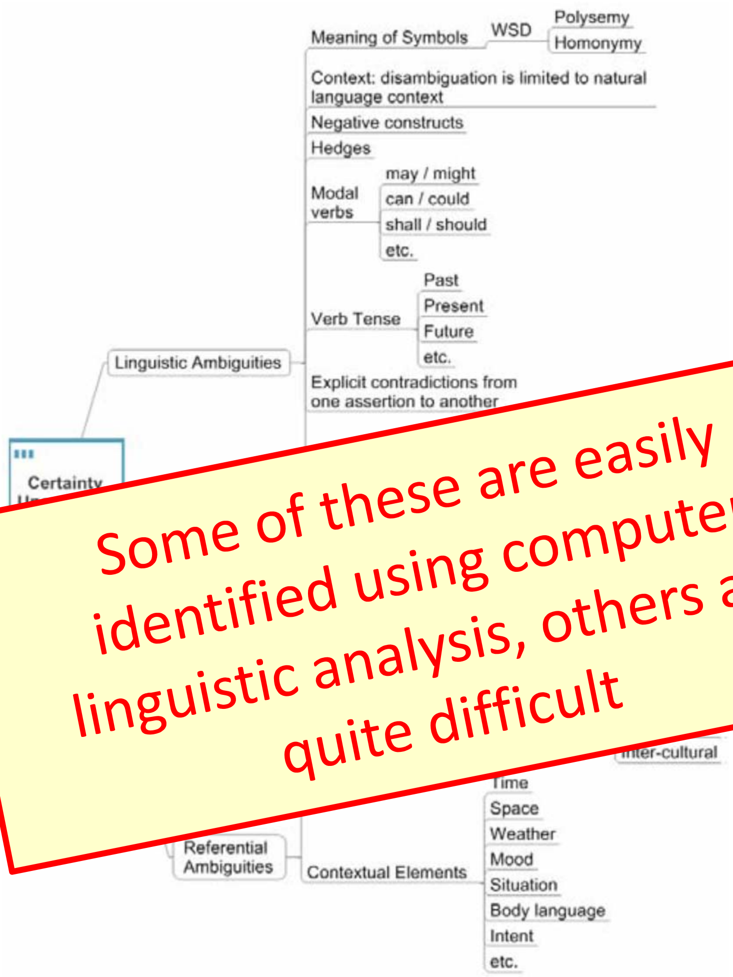
- **Hidden networks** – ties between sources which are unknown affecting our ability to judge information independence
  - Intelligence analysts often give more credence to information they believe to be derived from multiple “independent” sources

The Human as Sensor - issues

- **Hidden networks** – ties between sources which are unknown affecting our ability to judge information independence
  - Intelligence analysts often give more credence to information they believe to be derived from multiple "independent" sources
- **Time and tense** – whereas sensors record "historical" data, i.e., data about the past, soft data contains past, future, conditional and habitual information



Auger and Roy divide uncertainty in linguistic data into two broad categories: linguistic ambiguities and referential ambiguities



Some of these are easily identified using computer linguistic analysis, others are quite difficult

Auger and Roy divide uncertainty in linguistic data into two broad categories: linguistic ambiguities and referential ambiguities

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
- Ambiguity and polysemy (multiple meanings)

Uncertainty ***about*** the content, including:

- Modal verbs
- Modal adverbs (including "words of estimative probability")
- Hearsay markers
- "Mindsay" markers → belief, inference, assumption, etc.
- Passive voice

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).

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
- Ambiguity and polysemy (multiple meanings)

Uncertainty ***about*** the content, including:

- Modal verbs
- Modal adverbs (including "words of estimative probability")
- Hearsay markers
- "Mindsay" markers → belief, inference, assumption, etc.
- Passive voice

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).

Uncertainty ***within*** the content: Imprecision and vagueness

There were some animals in the road.

"*Some*" is imprecise -- we can only guess how many.

Some alternative formulations and dimensions:

-- "*a couple*" might have been as a descriptor if there were only two or three animals,

-- "a bunch" might have used for a dozen or so

-- "*many*" or "*a lot*" would have been preferred if there were a noticeably larger quantity, such as a twenty or fifty.

Uncertainty ***within*** the content:

Imprecision and vagueness

There were some animals in the road.

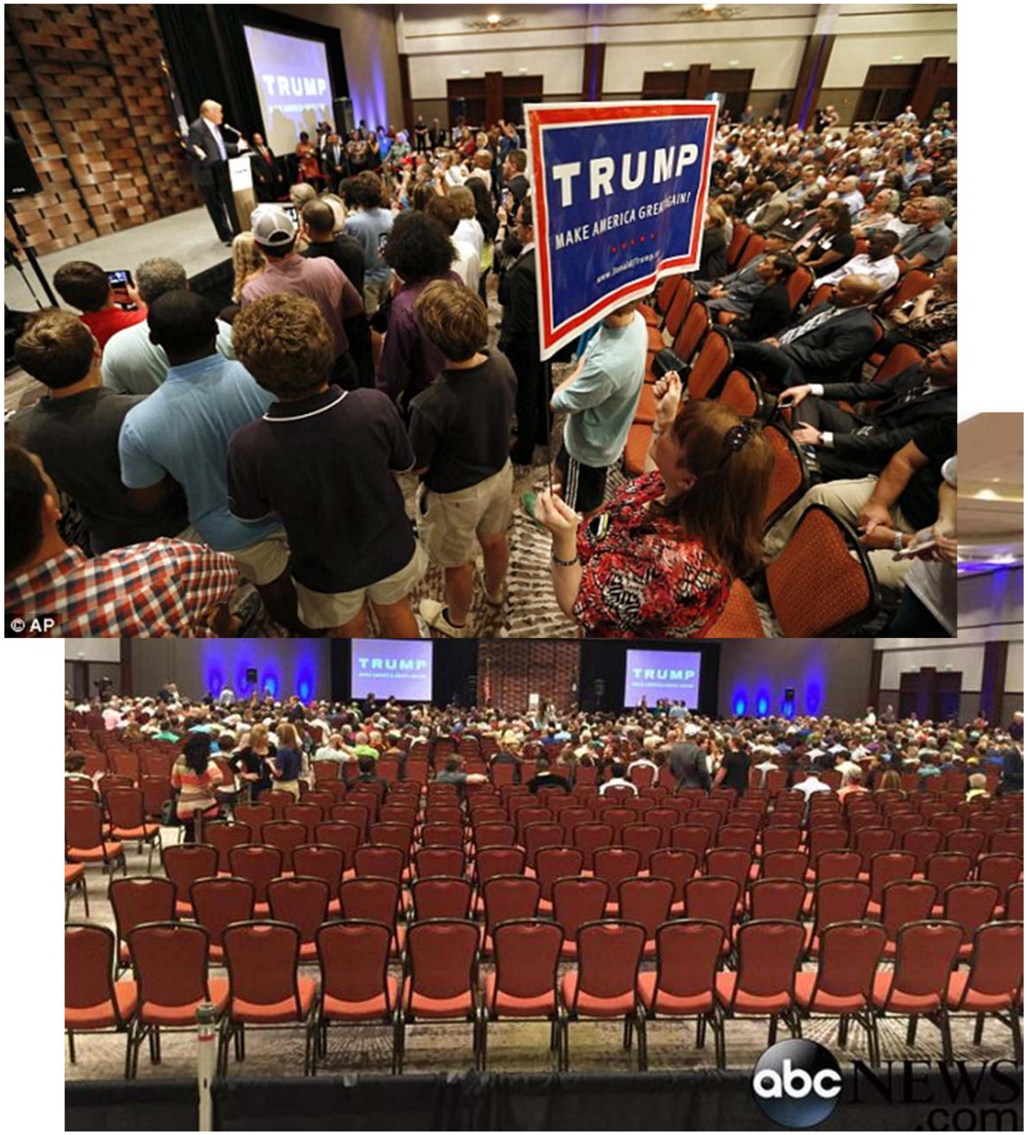
"*Some*" is imprecise -- we can only guess how many.

Uncertainty ***within*** the content:

Imprecision and vagueness

There were some animals in the road.

"*Some*" is imprecise -- we can only guess how many.



"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.

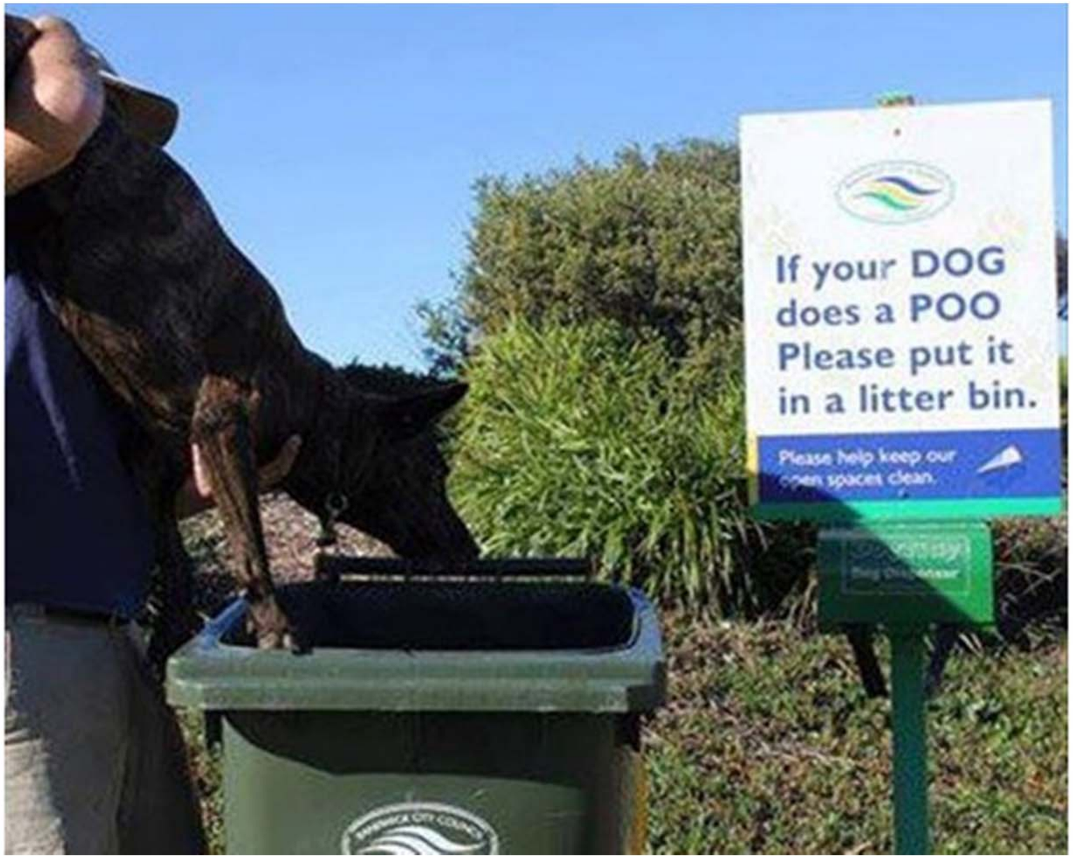
Uncertainty ***within*** the content:

Imprecision and vagueness

There were some animals in the road.

Uncertainty ***within*** the content:

Ambiguity and polysemy


If your DOG does a POO Please put it in a litter bin.
Please help keep our open spaces clean.

Uncertainty ***within*** the content:

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:

Students hate annoying professors.

(Who is annoying?)

Uncertainty ***within*** the content:
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:

I saw her duck

Uncertainty ***within*** the content:

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:

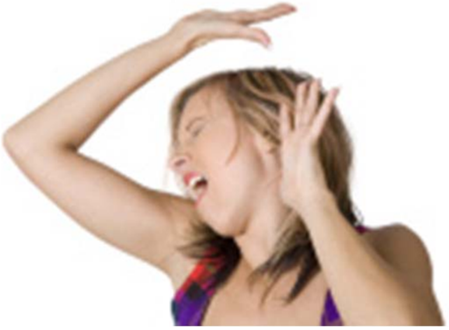
I saw her duck

Uncertainty ***within*** the content: 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:

I saw her duck

(movement or waterfowl ?)



Uncertainty ***within*** the content: 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:

Sally gave Mary her book.

While the other statements each had two possible resolutions, this statement has at least four…!

Uncertainty ***within*** the content:

Synonymy

In contrast to ambiguity, in which a single word or phrase multiple may have multiple interpretations, synonymy is where the same state or event is described in multiple ways…

Uncertainty ***within*** the content: Synonymy

In contrast to ambiguity, in which a single word or phrase multiple may have multiple interpretations, synonymy is where the same state or event is described in multiple ways…

Terrorists leave chlorine gas containers in a car and set the car up to explode, releasing the gas.

Uncertainty ***within*** the content:
Synonymy

After the explosion, calls start arriving at the emergency call center.

There‘s a car burning across the street from my office. It sounded like it exploded.

I had just come out of my office building when a car parked opposite the building exploded.

I would like to report a burning car. It doesn‘t look like there was an accident. Also, I heard what sounded like an explosion a minute ago.

Hey, man, there‘s a car on fire down the road. I heard a noise like a bomb went off and came around the corner.

I heard a really loud boom, like a bomb went off, and now there‘s a car on fire down the road.

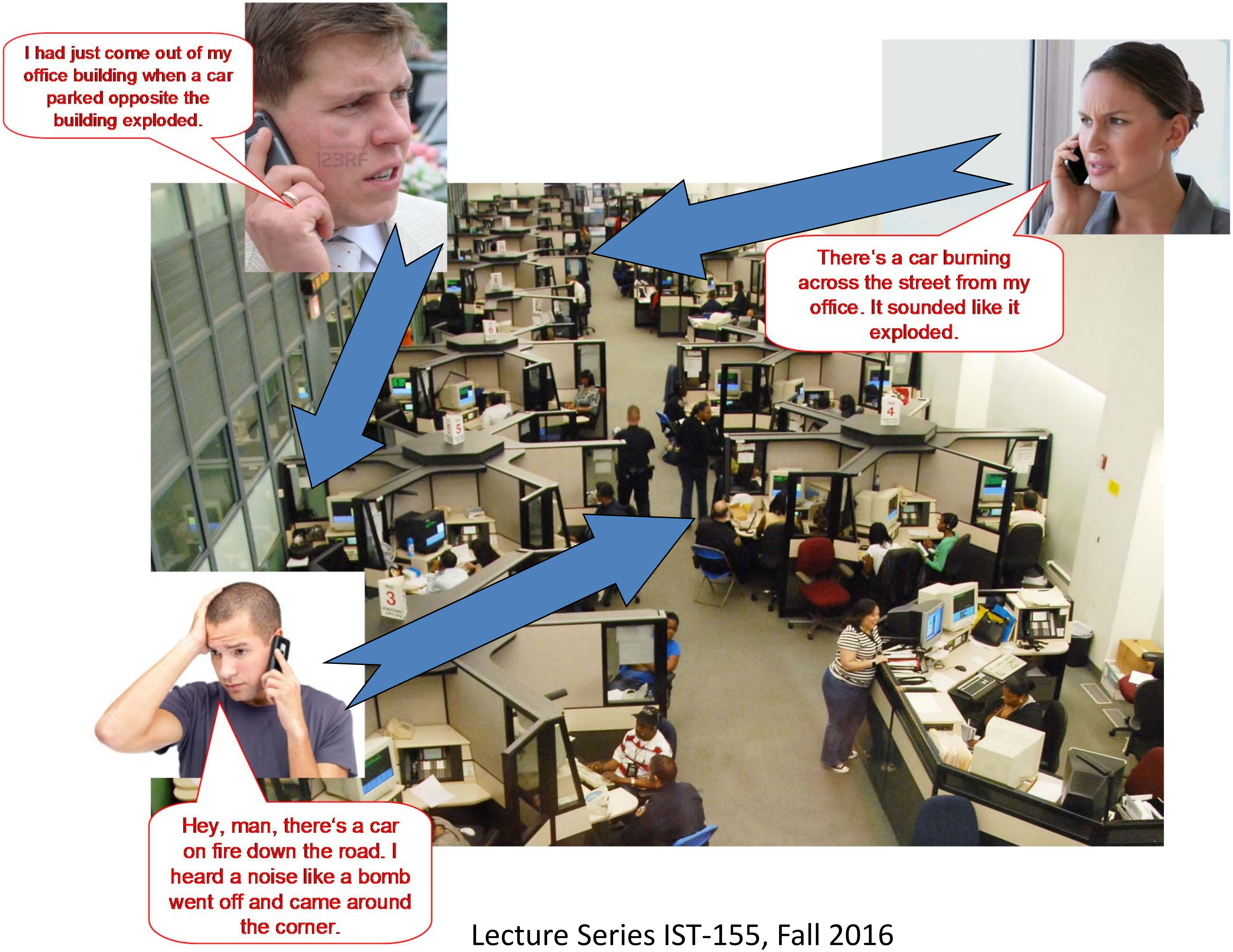
Uncertainty ***within*** the content:
Synonymy

“Burning…exploded…blew up…on fire…
explosion…sounded like a bomb“

These are all synonyms relating to the incident which has just occurred.

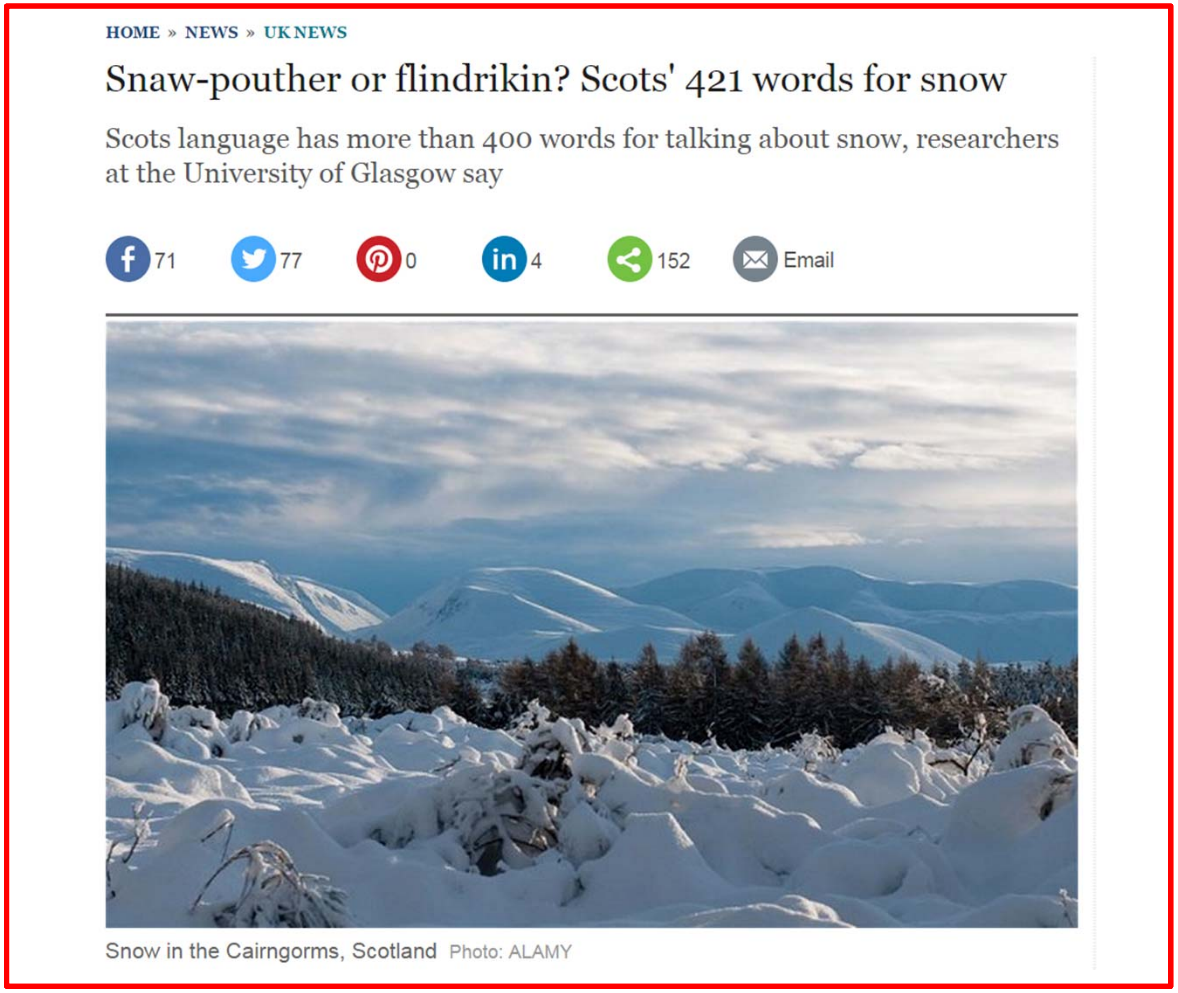
As humans and speakers of English we understand that these various words and phrases (may) belong together and may point to the same incident.

A computer system doesn‘t recognize this unless we tell it how.


I had just come out of my office building when a car parked opposite the building exploded.
There's a car burning across the street from my office. It sounded like it exploded.
Hey, man, there's a car on fire down the road. I heard a noise like a bomb went off and came around the corner.

Uncertainty *within* the content:

Synonymy



Snaw-pouther or flindrikin? Scots' 421 words for snow

Scots language has more than 400 words for talking about snow, researchers at the University of Glasgow say

Snow in the Cairngorms, Scotland Photo: ALAMY

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
- Ambiguity and polysemy (multiple meanings)

Uncertainty ***about*** the content, including:

- Modal verbs
- Modal adverbs (including "words of estimative probability")
- Hearsay markers
- "Mindsay" markers → belief, inference, assumption, etc.
- Passive voice

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).

Uncertainty ***about*** the content:

Linguistic markers of uncertainty

Specific content within a given statement is often packed with lexical elements that indicate in some manner the uncertainty of the content itself or that indicate the original source of information.

- ***John is a terrorist.***
- ***The CIA have concluded that John is a terrorist.***
- ***I believe that John is a terrorist.***
- ***My neighbor thinks John is a terrorist.***
- ***It has been definitely disproved that John is a terrorist.***
- ***Unless things change, John will be a terrorist one day.***
- ***The CIA have concluded that John is probably a terrorist.***

Uncertainty ***about*** the content:
Linguistic markers of uncertainty

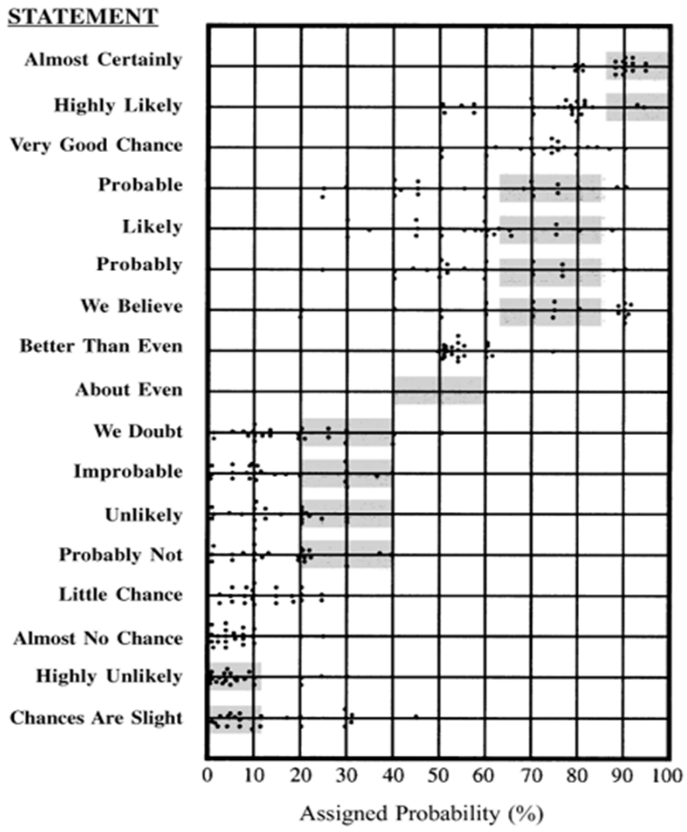
“A few days after the estimate [“NIE 29-51, "Probability of an Invasion of Yugoslavia in 1951”] appeared, I was in informal conversation with the Policy Planning Staff's chairman. We spoke of Yugoslavia and the estimate. Suddenly he said, "By the way, what did you people mean by the expression **`serious possibility'**? …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.**

“[it turned out 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.

Uncertainty ***about*** the content:

Linguistic markers of uncertainty

Figure 18: Measuring Perceptions of Uncertainty

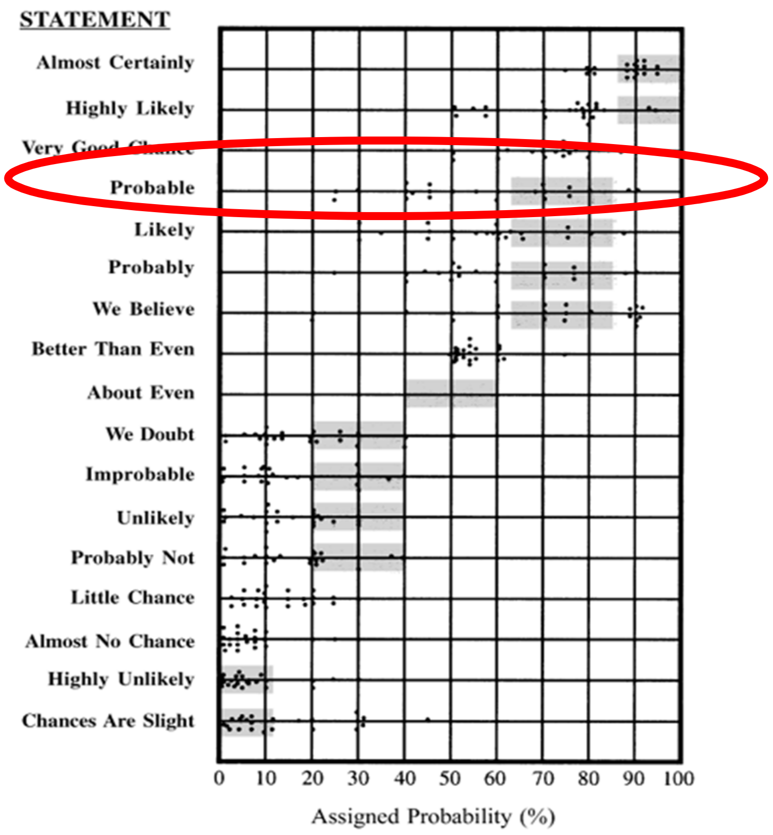


Results of CIA analyst Sherman Kent‘s informal study of weighting by colleagues

Uncertainty ***about*** the content:

Linguistic markers of uncertainty

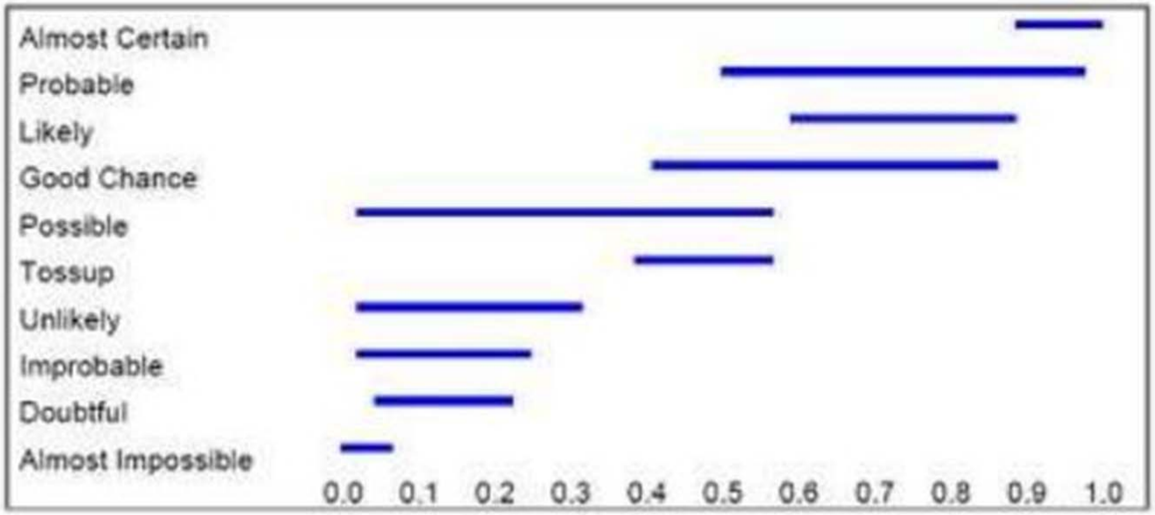
Figure 18: Measuring Perceptions of Uncertainty



Results of CIA analyst Sherman Kent‘s informal study of weighting by colleagues

Uncertainty ***about*** the content:

Linguistic markers of uncertainty

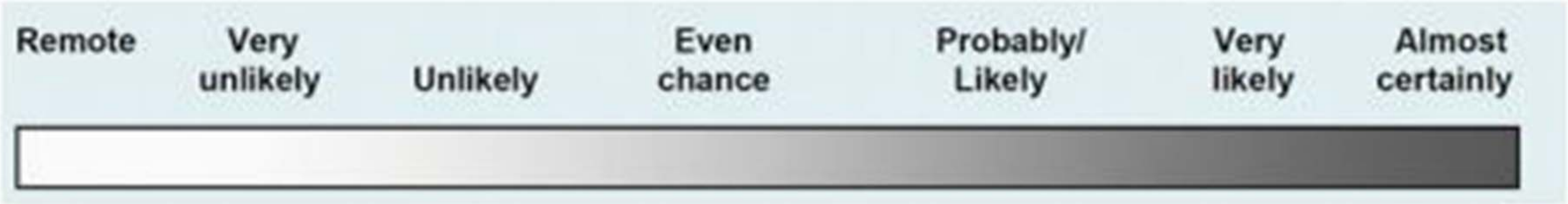


Ranges of percentages assigned to hedges by analysts in training.

Uncertainty ***about*** the content:

Linguistic markers of uncertainty

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



Uncertainty ***about*** the content:

Linguistic markers of uncertainty

Specific content within a given statement is often packed with lexical elements that indicate in some manner the uncertainty of the content itself or that indicate the original source of information.

- ***John is a terrorist.***
- ***The CIA have concluded that John is a terrorist.***
- ***I believe that John is a terrorist.***
- ***My neighbor thinks John is a terrorist.***
- ***It has been definitely disproved that John is a terrorist.***
- ***Unless things change, John will be a terrorist one day.***
- ***The CIA have concluded that John is probably a terrorist.***

Uncertainty ***about*** the content:
Linguistic markers of uncertainty

Specific content within a given statement is often packed with lexical elements that indicate in some manner the uncertainty of the content itself or that indicate the original source of information.

- ***John is a terrorist.***
- ***The CIA have concluded that John is a terrorist.***
- ***I believe that John is a terrorist.***
- ***My neighbor thinks John is a terrorist.***
- ***It has been definitely disproved that John is a terrorist.***
- ***Unless things change, John will be a terrorist one day.***
- ***The CIA have concluded that John is probably a terrorist.***

Uncertainty ***about*** the content:

Linguistic markers of uncertainty

Specific content within a given statement is often packed with lexical elements that indicate in some manner the uncertainty of the content itself or that indicate the original source of information.

- ***John is a terrorist.***
- ***The CIA have concluded that John is a terrorist.***
- ***I believe that John is a terrorist.***
- ***My neighbor thinks John is a terrorist.***
- ***It has been definitely disproved that John is a terrorist.***
- ***Unless things change, John will be a terrorist one day.***
- ***The CIA have concluded that John is probably a terrorist.***

No linguistic markers
Truth determined by other factors such as source or corroboration with other sources

Uncertainty ***about*** the content:

Linguistic markers of uncertainty

Specific content within a given statement is often packed with lexical elements that indicate in some manner the uncertainty of the content itself or that indicate the original source of information.

Not original source (citation)
and also opinion/inference

- ***John is a terrorist.***
- ***The CIA have concluded that John is a terrorist.***
- ***I believe that John is a terrorist.***
- ***My neighbor thinks John is a terrorist.***
- ***It has been definitely disproved that John is a terrorist.***
- ***Unless things change, John will be a terrorist one day.***
- ***The CIA have concluded that John is probably a terrorist.***

Uncertainty ***about*** the content:
Linguistic markers of uncertainty

Specific content within a given statement is often packed with lexical elements that indicate in some manner the uncertainty of the content itself or that indicate the original source of information.

Belief (mindsay)

- ***John is a terrorist.***
- ***The CIA have concluded that John is a terrorist.***
- ***I believe that John is a terrorist.***
- ***My neighbor thinks John is a terrorist.***
- ***It has been definitely disproved that John is a terrorist.***
- ***Unless things change, John will be a terrorist one day.***
- ***The CIA have concluded that John is probably a terrorist.***

Uncertainty ***about*** the content:

Linguistic markers of uncertainty

Specific content within a given statement is often packed with lexical elements that indicate in some manner the uncertainty of the content itself or that indicate the original source of information.

Hearsay plus mindsay

- ***John is a terrorist.***
- ***The CIA have concluded that John is a terrorist.***
- ***I believe that John is a terrorist.***
- ***My neighbor thinks John is a terrorist.***
- ***It has been definitely disproved that John is a terrorist.***
- ***Unless things change, John will be a terrorist one day.***
- ***The CIA have concluded that John is probably a terrorist.***

Uncertainty ***about*** the content:

Linguistic markers of uncertainty

Specific content within a given statement is often packed with lexical elements that indicate in some manner the uncertainty of the content itself or that indicate the original source of information.

Negative assertion, vagueness as to original source (possible mindsay plus possible hearsay)

- ***John is a terrorist.***
- ***The CIA have concluded that John is a terrorist.***
- ***I believe that John is a terrorist.***
- ***My neighbor thinks John is a terrorist.***
- ***It has been definitely disproved that John is a terrorist.***
- ***Unless things change, John will be a terrorist one day.***
- ***The CIA have concluded that John is probably a terrorist.***

Uncertainty ***about*** the content:
Linguistic markers of uncertainty

Specific content within a given statement is often packed with lexical elements that indicate in some manner the uncertainty of the content itself or that indicate the original source of information.

Conjecture/conditional of possible future state

- ***John is a terrorist.***
- ***The CIA have concluded that John is a terrorist.***
- ***I believe that John is a terrorist.***
- ***My neighbor thinks John is a terrorist.***
- ***It has been definitely disproved that John is a terrorist.***
- ***Unless things change, John will be a terrorist one day.***
- ***The CIA have concluded that John is probably a terrorist.***

Uncertainty ***about*** the content:

Linguistic markers of uncertainty

Specific content within a given statement is often packed with lexical elements that indicate in some manner the uncertainty of the content itself or that indicate the original source of information.

Mindsay plus estimative probability

- ***John is a terrorist.***
- ***The CIA have concluded that John is a terrorist.***
- ***I believe that John is a terrorist.***
- ***My neighbor thinks John is a terrorist.***
- ***It has been definitely disproved that John is a terrorist.***
- ***Unless things change, John will be a terrorist one day.***
- ***The CIA have concluded that John is probably a terrorist***

Uncertainty ***about*** the content:

Linguistic markers of uncertainty

Hedges and evidential (how the information was acquired) markers are relatively obvious indicators of uncertainty, even to non-linguists. However, there are some more subtle ways in which uncertainty may appear.

> …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***.

Uncertainty ***about*** the content:

Linguistic markers of uncertainty

Passive voice:

- Particularly in scientific writing, the use of passive voice and impersonal phrasing are widely,
  “I might be wrong or have overlooked something.”
- 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.
- Also sometimes used in the case of differences in social ranking or power, in order not to offend – again more information needed

Uncertainty ***about*** the content:

Linguistic markers of uncertainty

Verb tenses / moods / temporal expressions

- Future is inherently uncertain because ***it may not happen***

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

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

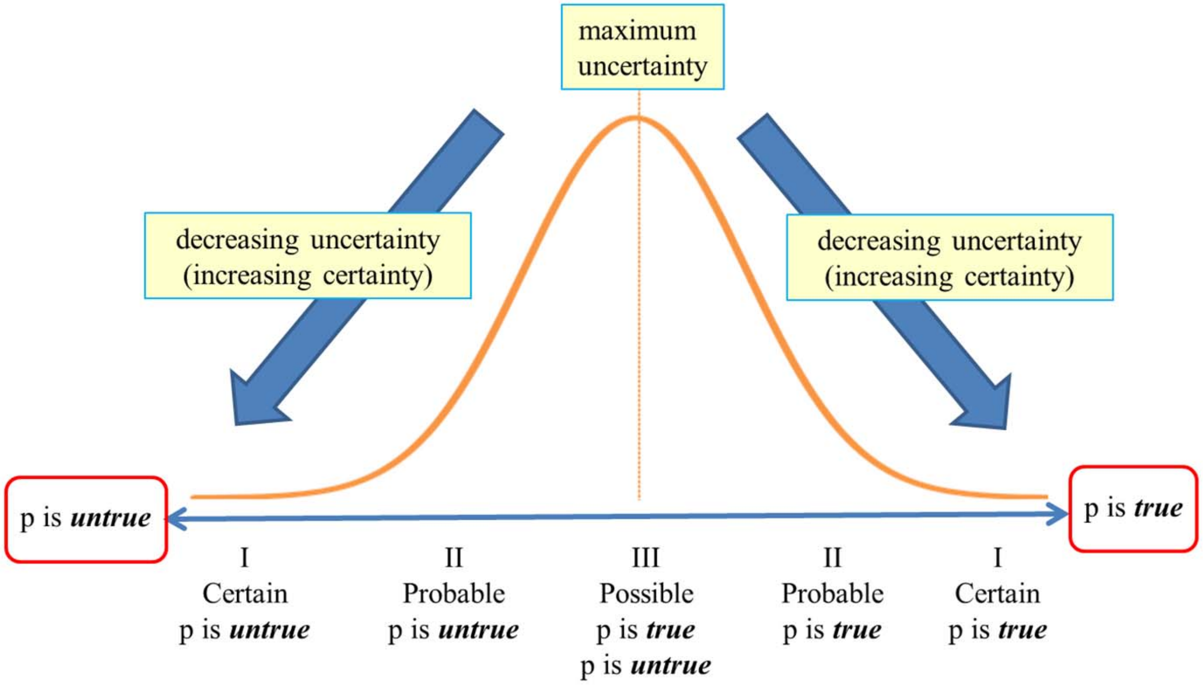
- Expressions of routine activity: ***The group meets every Monday at 10 a.m.***

  ("Well, not next week because it's a holiday…")

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).

Uncertainty ***about*** the content:

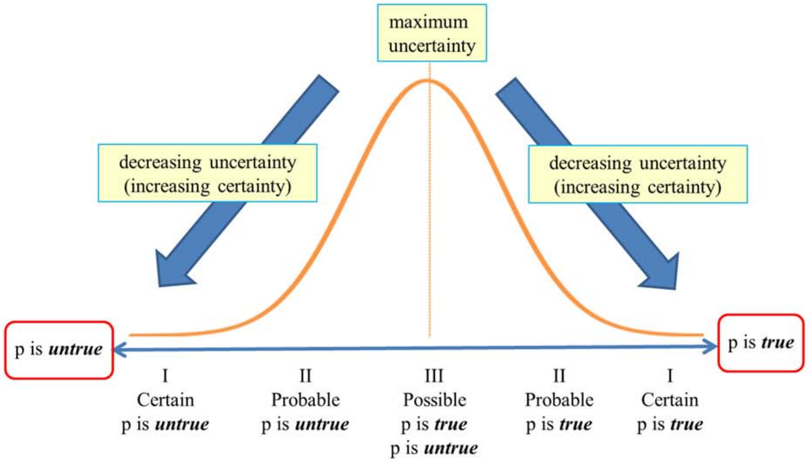
Linguistic markers of uncertainty

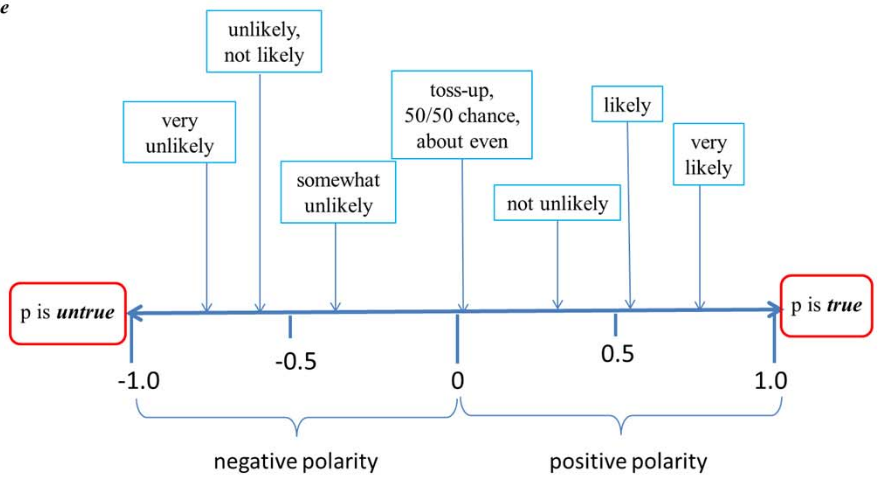


Lecture Series IST-155, Fall 2016

Uncertainty ***about*** the content:

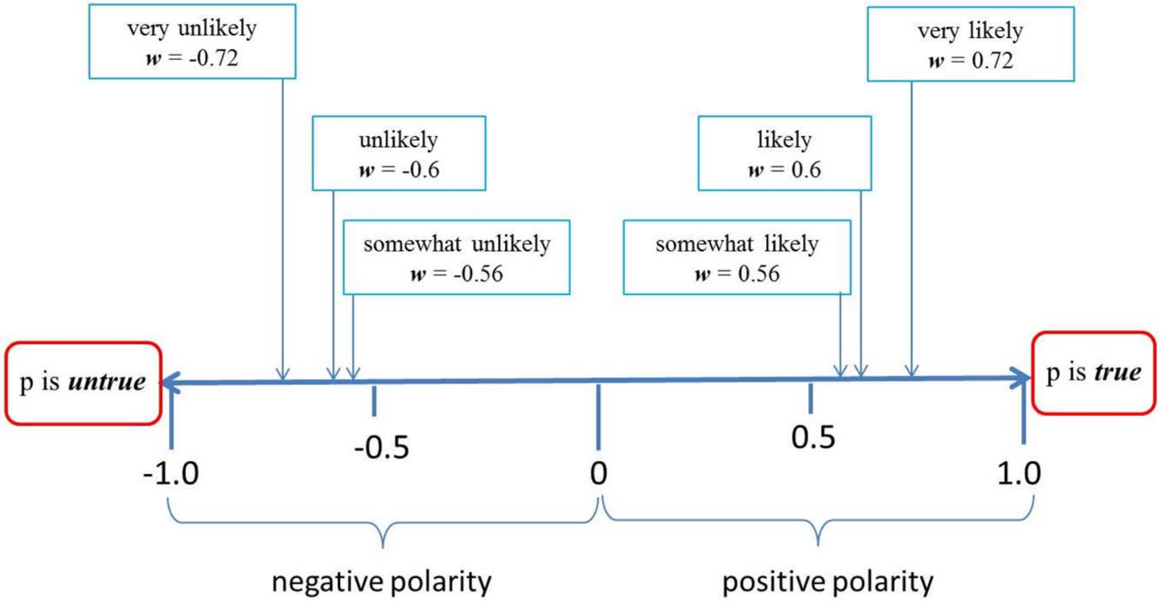
Linguistic markers of uncertainty





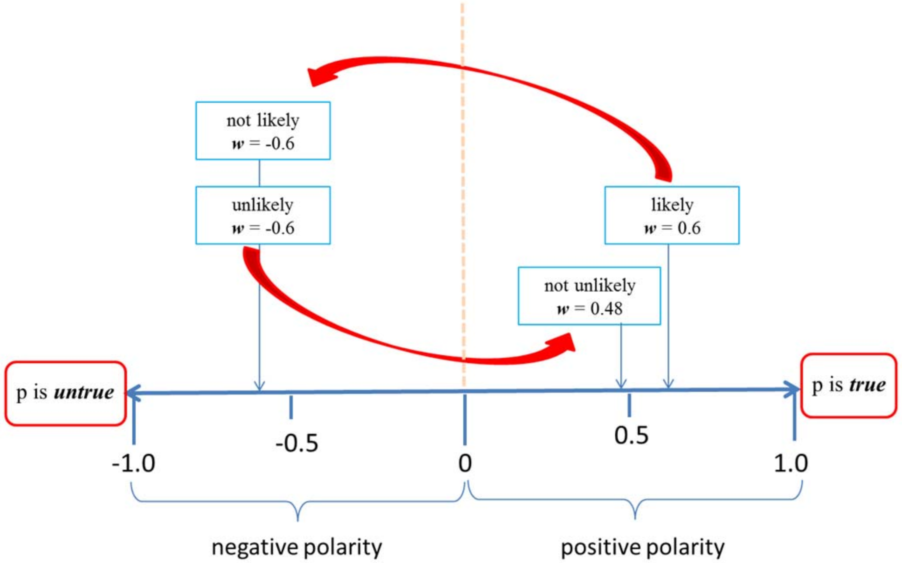
Uncertainty ***about*** the content:

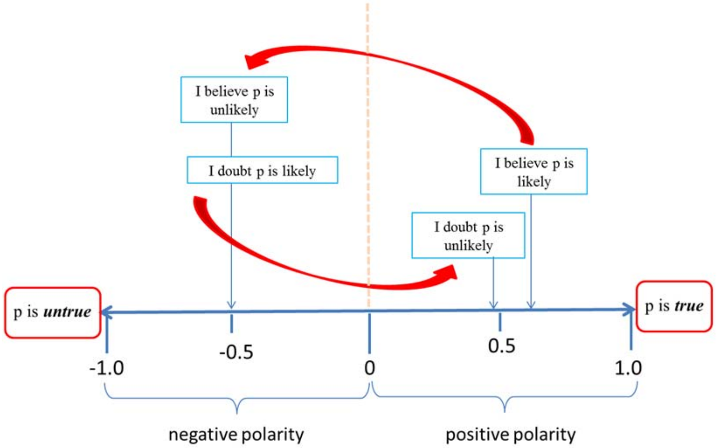
Linguistic markers of uncertainty



Uncertainty ***about*** the content:

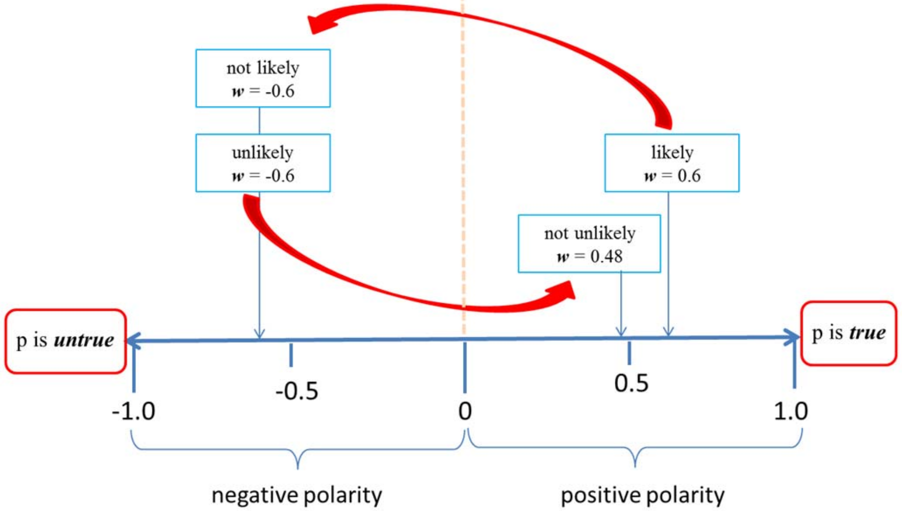
Linguistic markers of uncertainty

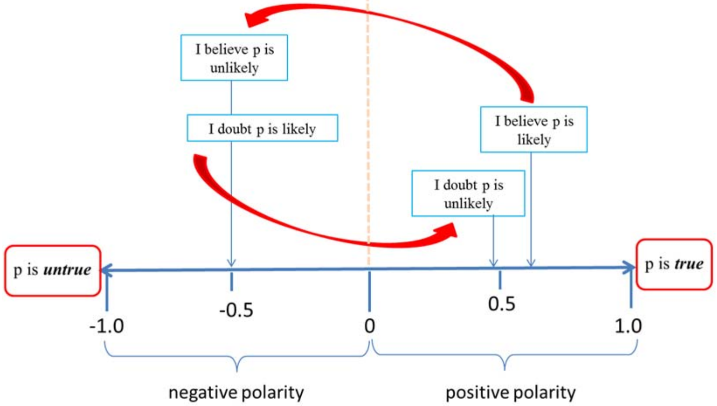




Uncertainty ***about*** the content:

Linguistic markers of uncertainty

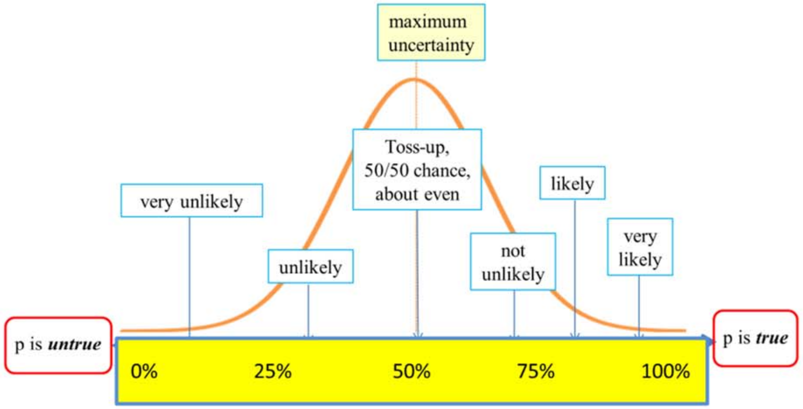


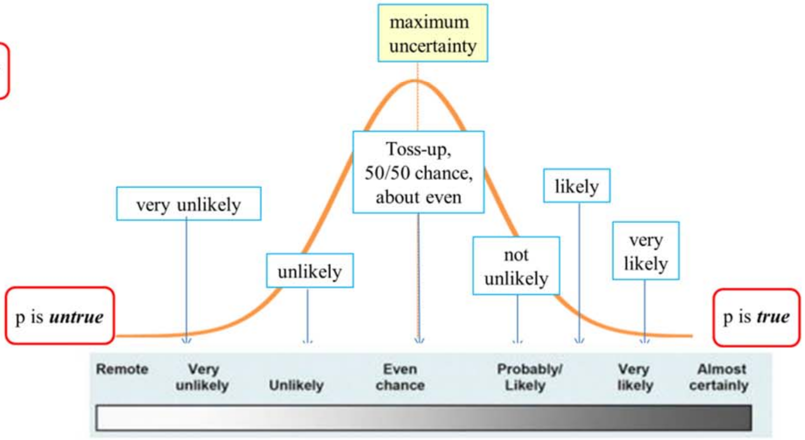


$$e = \prod_{i=1}^{n} w_{hedge_i} * \prod_{j=1}^{m} w_{hearsay/mindsay_j}$$

Uncertainty ***about*** the content:

Linguistic markers of uncertainty





Last but not least….

Which language??

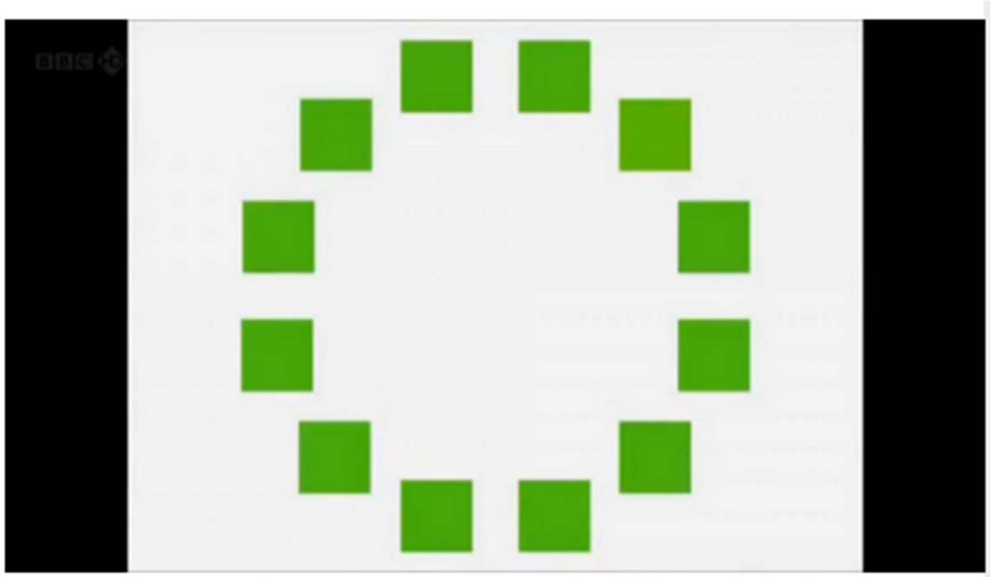
- Nearly 7000 distinct languages in the world
  - Of which “only“ 230 are spoken in Europe
- Within languages there are regional and domain-specific differences which cause confusion and misunderstanding:
  - The “biscuit” of a Brit is an American’s “cookie” – and an American’s “biscuit” more akin to an unsweetened British “scone.”
  - “ POV” within the US military is generally understood to be “privately owned vehicle” (i.e., a soldier’s own car), but within the writers’ community means “point of view.”
  - SME – subject matter expert? Small to medium enterprise?? Other??
- Social media, texting and other new forms of communication result in constantly changing forms of expression (emoji, lol, r u , etc.)

The OvaHimba use four color names:

1. *zuzu* stands for dark shades of blue, red, green and purple;
2. *vapa* is white and some shades of yellow;
3. *buru* is some shades of green and blue; and
4. *dambu* is some other shades of green, red and brown.

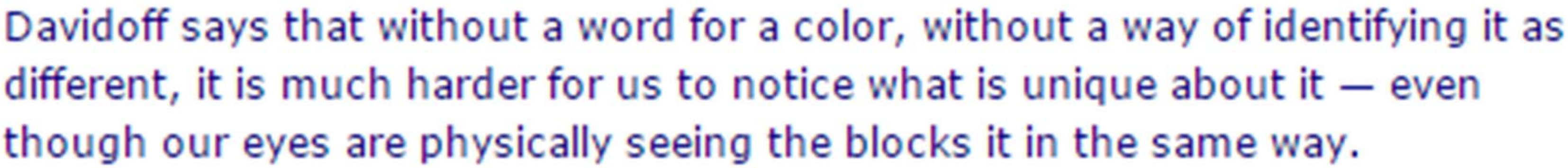
The OvaHimba use four color names:

1. *zuzu* stands for dark shades of blue, red, green and purple;
2. *vapa* is white and some shades of yellow;
3. *buru* is some shades of green and blue; and
4. *dambu* is some other shades of green, red and brown.

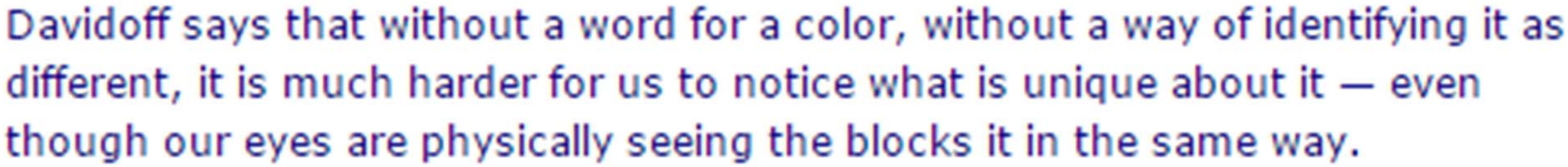


The OvaHimba use four color names:

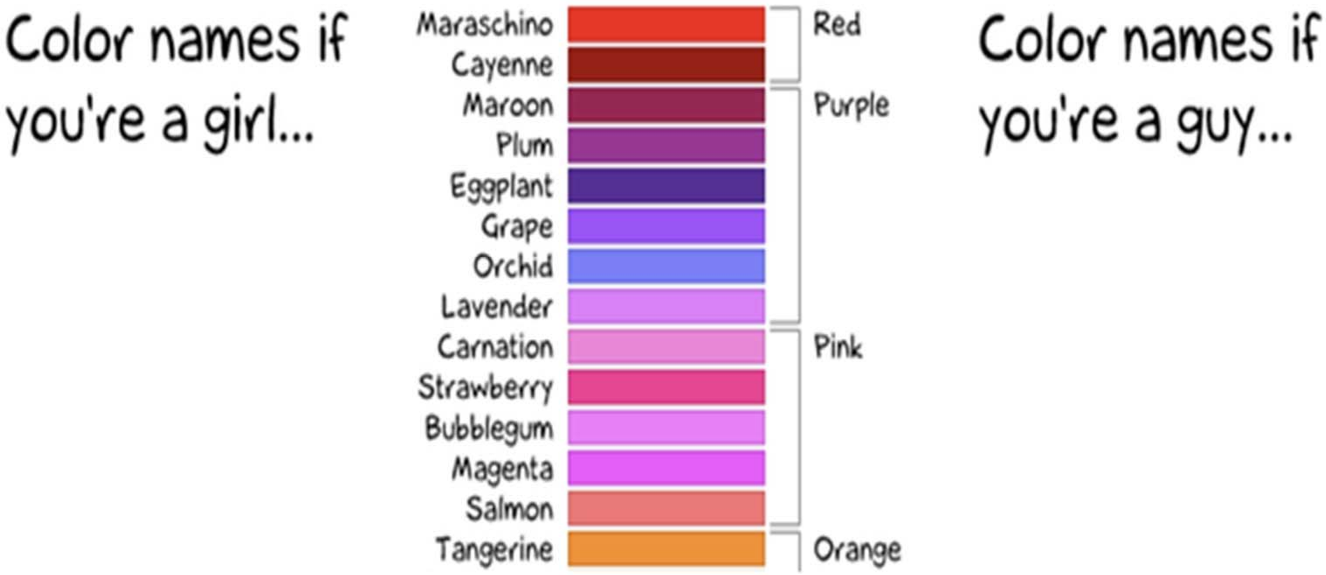
1. *zuzu* stands for dark shades of blue, red, green and purple;
2. *vapa* is white and some shades of yellow;
3. *buru* is some shades of green and blue; and
4. *dambu* is some other shades of green, red and brown.



Davidoff says that without a word for a color, without a way of identifying it as different, it is much harder for us to notice what is unique about it — even though our eyes are physically seeing the blocks it in the same way.

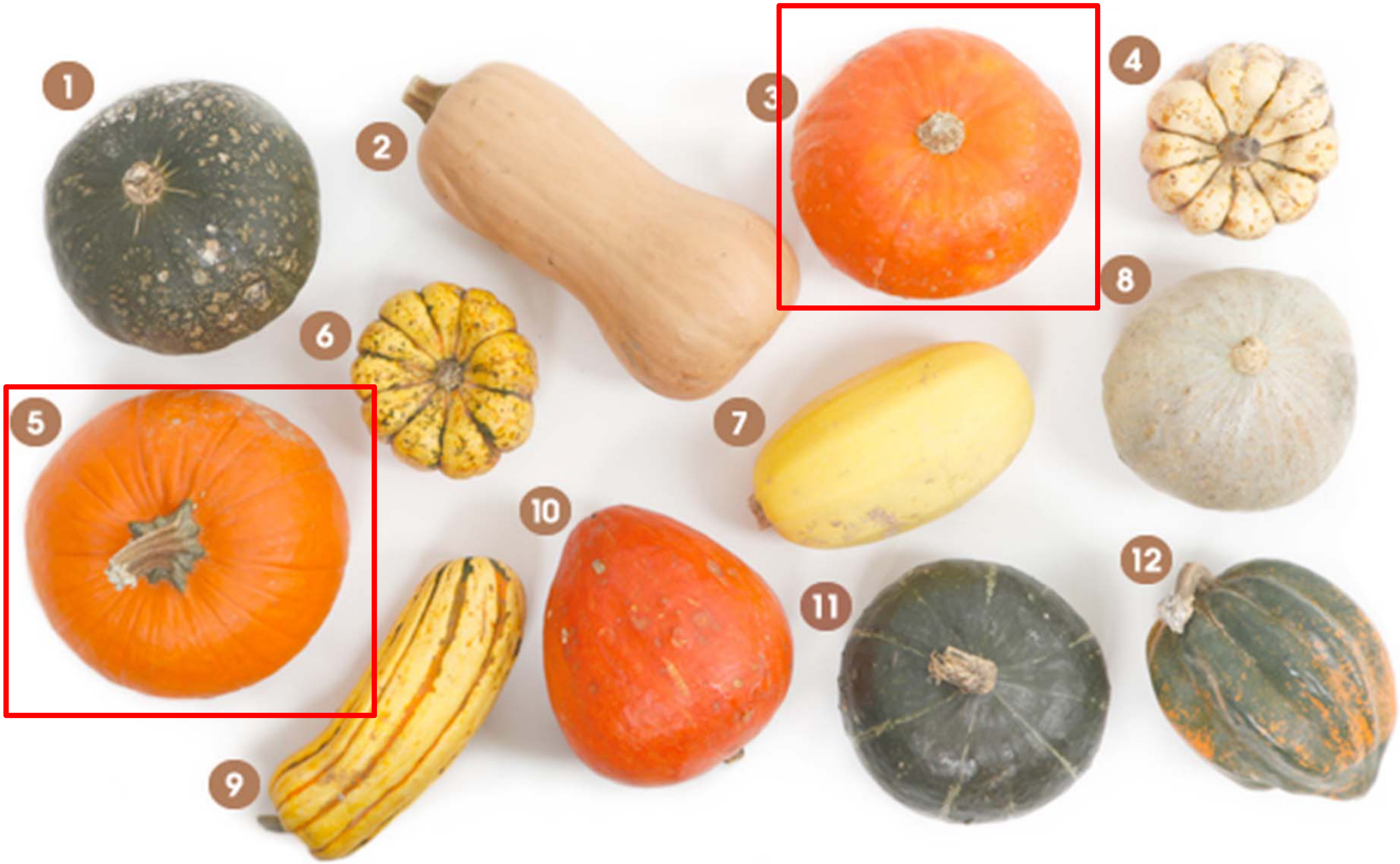


Davidoff says that without a word for a color, without a way of identifying it as different, it is much harder for us to notice what is unique about it — even though our eyes are physically seeing the blocks it in the same way.



Pumpkins

Pumpkins



ns

Squash (red kabocha)

1
2
6
5
9

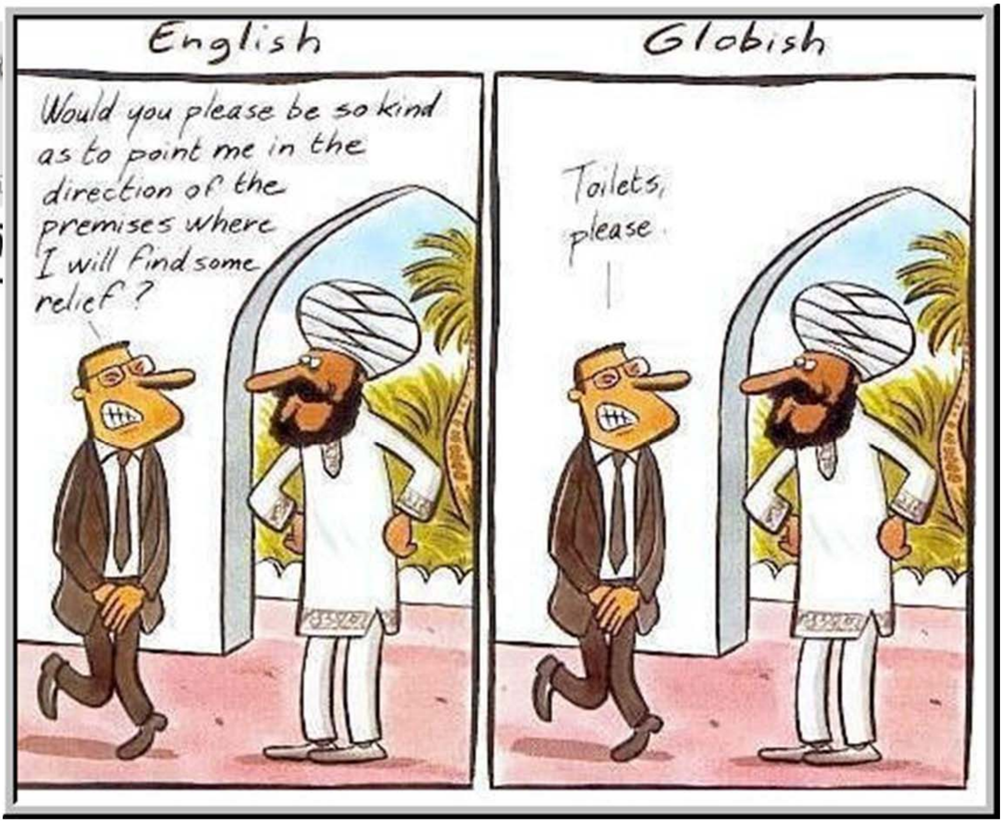
5

(sugar) pumpkin

Many thanks for your attention!

Questions???

PLEASE, MATE, BUNG US SOME DOSH FOR THE DOG AND BONE!
I DON'T UNDERSTAND. I ONLY SPEAK ENGLISH!


English
Globish
Would you please be so kind as to point me in the direction of the premises where I will find some relief?
Toilets, please.

> We are drowning in information but starved for knowledge.
>
> John Naisbitt

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.

The sheer volume of new natural language information being generated daily means that automatic (initial) processing of text-based information is increasingly vital.

However, 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.