

# **"Wave Linguistics"**

**The one true grammar is the  
grammar which cannot be known**



# **Overview:**

**The problem**

**Movie metaphor**

**Subjective truth - an explanation for the elusiveness of  
grammar**

**Examples**

**Everyone who has worked for a time in NLP knows eventually  
you can't add new rules without breaking old rules**

**This is a problem for statistical as well as symbolic models  
(statistical = overtraining)**

**Why have we not been able to find an  
exhaustive grammar for Natural Language?...**



## Paradigm Change



**What is the problem with CL?**

**Suggestion: A Paradigm Change**

**Forms Basic -> Collections of Examples Basic**

**(The structure you see is not the real structure.)**

## For instance the collection of examples

ADJ (foreign) N (exchange)

foreign exchange

... ..

... ..

foreign bonds

... ..

stock exchange

foreign currency

... ..

... ..

currency exchange

... ..

securities exchange

foreign bond

... ..

... ..

regional exchanges

regional exchange

... ..

Could be summarized as the rule:

**NP <- ADJ + N**

**But isn't that like building a puppet to mimic a flip-book?**

**How much more powerful is the flip-book.**

**In practice - do grammar by meshing vectors of examples rather than restricting yourself to the regularities expressed in traditional classes (N, V, PREP, NP etc.)**

**The basic process of language is not the expression of rules but the search for regularities.**

**Classes -> Vectors**

**Vectors - the power of many dimensions (of rules), extent, waves.**



E.g.  
NP <- ADJ + N  
NP <- NP + N

But nobody has found a set of classes and combinations which capture all the necessary distinctions in language.

Essentially I do the same, with the twist that I replace the classes with vectors or lists of similar words. This means, by virtue of one combination or another, I can find millions of virtual rules at run time. The exact set depending on the exact words in the sentence.

E.g.

NP (foreign exchange) <-	ADJ (foreign)	N (exchange)
foreign exchange	foreign	exchange
foreign bonds	...	...
stock exchange	...	...
the stock	foreign	bonds
foreign currency	...	...
the securities	stock	exchange
foreign languages	foreign	currency
currency	...	...
foreign ministry	...	...
discount	currency	exchange
foreign residents	...	...
equity	securities	exchange
the capital	foreign	bond
exchange	...	...
stock	...	...

everything is represented in terms of sub-strings of up to three words. This is purely for practical reasons. I have not explored in any principled way the limitations on accuracy of forcing even long sentences to have representations in terms of vectors of strings 1-3 words long. You must have strings of length at least two words, however, so that you can find entries for observed pairs when calculating a new vector category.

**Collective Meaning?** In this model every valid new sentence in a language corresponds to a valid rearrangement of previous examples of that language. It is interesting to consider a parallel definition of meaning. Define "meaning" to be an organization of experience. Now consider a set of previous experiences each of which is associated with a previous example of language use. New language will force a reorganization of the language examples, which will force a new organization of previous experiences, which will by definition determine new meaning, according to the assumptions of the model.

There is an analogy to what happens when you perform a Web search using an indexed search engine. The number of hits might be said to define the "grammaticality" of the search query, and the actual collection of hits its "meaning". A new meaning for each query.

**Uncertainty Principle?** Collections have interesting properties. We even get a kind of "linguistic uncertainty principle", where we see that it is impossible to order a collection of examples in two ways simultaneously. Generally speaking ordering in terms of one variable

**The extent to which the vectors mesh (the amount of red)  
...the number of examples which fit the pattern if you like  
...gives you a measure for the grammaticality of the  
combination.**

## **Example of Power:**

**Important property of systems based on collections of examples**

## **Subjective Truth**

**Because grammar is based on collections of examples they can be collected together in different ways.**

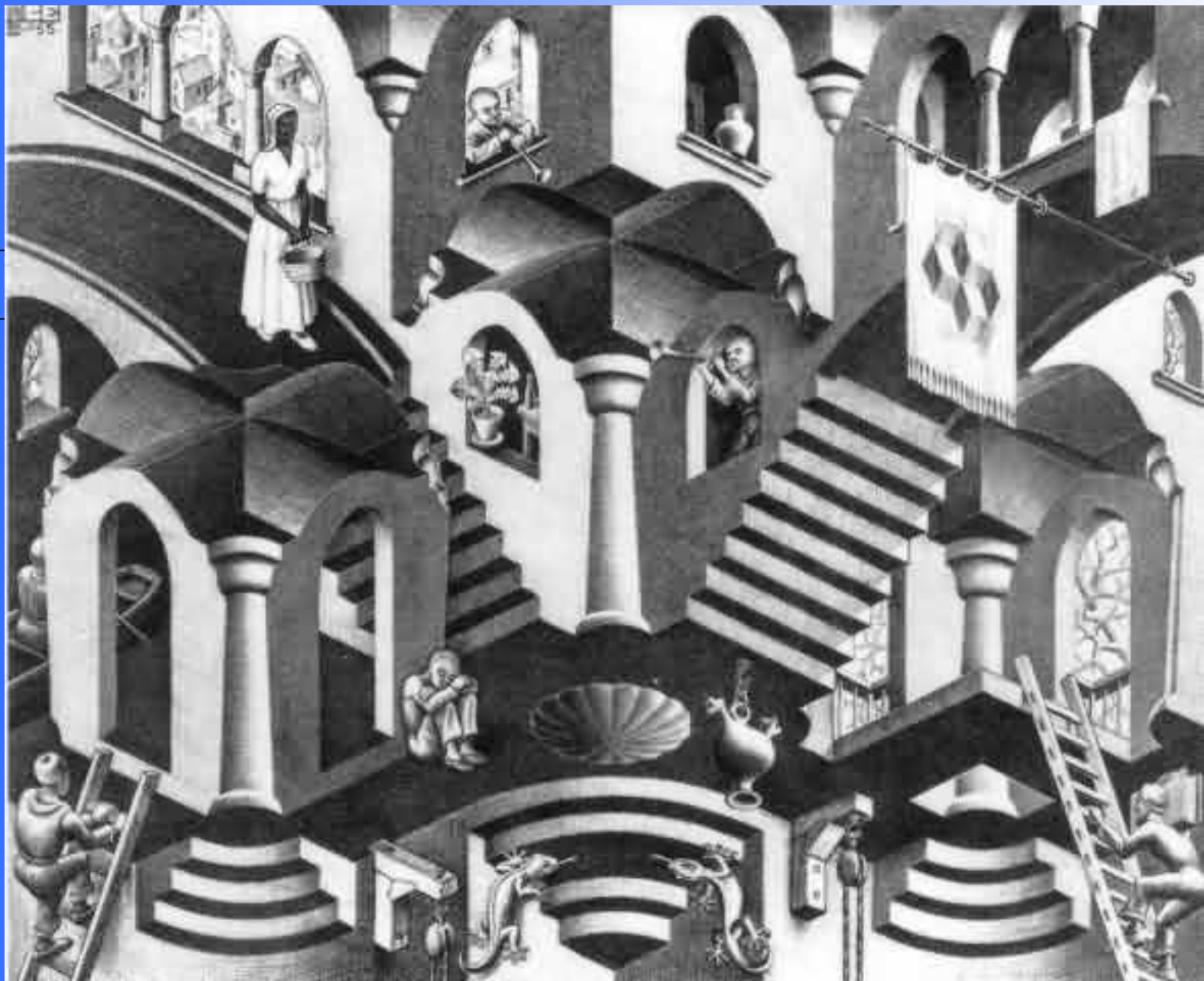
**Property of regularities over collections of elements:  
subjective truth - things that can't be true at the same time:**

**Letter sorted: A A A A F F F F**

**Colour sorted: A A F F A A F F**

**Example can be sorted according to colour or letter, but not  
both at the same time.**

**In this view the pattern (grammar) is rather like one or other perspective of an Escher drawing...**



**The picture can be seen as either convex or concave, but not both at the same time.**

**Example of subjective truth in grammatical categories:**

**Sentences where no one grammatical analysis seems obvious:**

**"He came only yesterday"**

**(pron + (v +(only yesterday)))**

**or**

**(pron + ((came only) + adv))**

**(c.f. "kino kita bakari"**

**"he only came yesterday"**

**"he came yesterday only")**

**You can make rules for both possibilities. But you can't capture the tension between possibilities. Except with collections of examples.**



**Example of subjective truth in grammatical categories:**

**A collocation is a rule reflecting a very specific sub-categorization of a word**

**But we can have a tension between different collocations.**

**E.g. "tea merchants" and "green tea"**

**"Tea merchants" is a term, and "green tea" is a term, but they can't both be terms in:**

**"The green tea merchants"**

**Rules don't handle this well. Collections of examples can be rearranged to sub-categorize smoothly and incompatibly in this way.**



## "Vector" Parser - Dictionary Lookup

Look up the "vector" class for another word:

### Vector class for "tea"

word	similarity score
tea	1
tea and	0.1
coffee	0.066
afternoon tea	0.064
supper	0.061
and tea	0.058
dinner	0.056
lunch	0.051
cake	0.049
iron	0.048
breakfast	0.047
minibar	0.047
wine	0.047
money	0.046
peace	0.045
death	0.045
friends	0.044
trouble	0.044
mini bar	0.044
a drink	0.043

## **Collection of examples defining global category for tea:**

coffee  
afternoon tea  
supper  
dinner  
lunch  
cake  
iron  
breakfast  
minibar  
wine  
money  
...  
a drink  
...

## Estimation of new "vector" category for: "(tea merchants)"

New vector - sum of components of  
vectors for observed pairs

Observed pairs between components of  
vectors for logical halves of:

word	similarity score
rich merchants	3.5
DIGIT1th century merchants	3.1
pottery lessons	0.62
work ethic subjects	0.4
rice fields	0.38
wheat fields	0.28
rice farmers	0.27
dancing lessons	0.24
shareholders equity	0.22
energy consumers	0.18
fruit trees	0.16
picture windows	0.16
french windows	0.16
kansas women	0.15
cooking lessons	0.15
king size	0.14
finance ministry	0.14
child care workers	0.14
copyright protection	0.13
the business men	0.13
mature trees	0.12
shipping documents	0.12
disclosure documents	0.12
queen size	0.11

←-- "(tea merchants)"

headword	tailword	pair frequency
rich	merchants	6
DIGIT1th century	merchants	10
pottery	lessons	11
work ethic	subjects	22
rice	fields	33
wheat	fields	10
rice	farmers	7
dancing	lessons	13
shareholders	equity	20
energy	consumers	6
picture	windows	117
french	windows	256
kansas	women	8
cooking	lessons	7
fruit	trees	129
child care	workers	16
copyright	protection	37
finance	ministry	75
the business	men	58
shipping	documents	6
king	size	350
disclosure	documents	30
risk	banks	82

## Collection of examples defining category of tea as modifier: (The combination of "vector" categories filter each other)

rich (merchants)  
pottery (lessons)  
work ethic (subjects)  
**rice** (fields)  
wheat (fields)  
dancing (lessons)  
share holders (equity)  
energy (...)  
picture  
french  
kansas  
cooking  
fruit  
...

Note how the countable concrete object quality of "merchants" picks out adjectival aspects of "tea": "rich", "french"... Qualities ("work ethic", "energy") or actions ("cooking", "dancing"), not objects.

## Estimation of new "vector" category for: "(green tea)"

New vector - sum of components of vectors  
for observed pairs

word	similarity score
green tea	17
green tees	13
afternoon tea	8.2
green lawns	7.3
green fields	4.2
green rice	3.3
green forests	3
green valleys	2.8
green lawn	2.3
green hills	2.2
umbrella pines	1.9
green sugar	1.8
beer house wine	1.7
green leaves	1.7
green mountains	1.6
green grass	1.6
bottle of wine	1.6
the beaches	1.6
soft drinks	1.5
green silk	1.5
wooded parkland	1.3
complimentary green tees	1.2
lush tropical gardens	1.2

Observed pairs between components of vectors for  
logical halves of:  
"(green tea)"

headword	tailword	pair frequency
green	tea	11
green	tees	276
afternoon	tea	726
green	lawns	41
green	fields	42
green	rice	15
green	forests	40
green	valleys	31
green	lawn	10
green	hills	55
umbrella	pin	25
green	sugar	10
beer	house wine	51
green	leaves	9
green	mountains	94
green	grass	6
green	silk	9
wooded	parkland	12
complimentary green	tees	12
lush	tropical gardens	135
terraced	tea	7
soft	drinks	594

## Collection of examples defining category of tea as modified:

(green) fees  
(green) lawns  
(green) fields  
(green) **rice**  
(green) forests  
(green) valleys  
(green) lawn  
(green) hills  
(umbrella) pines  
(...) sugar  
leaves  
mountains  
grass  
silk  
...

**Note how the modifier "green" has picked out mostly plurals, which characterize "categorical" nouns in English. Substantive, objects (a bit like "zi/jai" in Chinese c.f. "a chair", "the chair", "chairs")**

**A modifier picks out the substantive aspects "tea". A role as modifier picks out its qualitative aspects.**

**Both sub-categorizations of "tea" are possible, but not both at the same time.**

**In each case a sub-category is being crystallized from a broader of list of words in the general category of "tea" (which is in itself a rearrangement of examples characterizing the language.)**



**You can list rules and sub-categorizations for these two possibilities.**

**But in the limit you will need to list them for every such collocation and habitual usage.**

**In this sense a "knowable" grammar for a NL would be an exhaustive list of every possible production in the language!**

# Further examples of subjective grammatical categories:

Mozilla {Build ID: 2002051319}

File Edit View Search Go Bookmarks Tasks Help

← → ↶ ✕

http://localhost:9090/demo.html

Search


Google Search: comp.ai.n... Google Search: comp.ai.n... (Untitled) img17.png (PNG Image, 5...

# Chaotic Language

HOME PARSER DICTIONARY CONTACT

## Vector Parser

This site is currently a simple vehicle to demonstrate a parser (break a stream of language into meaningful parts) based on a novel perspective of language. In this view of language there are no grammar rules, no abstract "part-of-speech" classes. There are only "vectors" containing examples of language use. The parser works purely by rearranging the examples contained in these vectors. A valid rearrangement corresponds to a valid use of the language.



### Why "Collective"?

The name "Collective Language" recalls the key assumption of this approach, which is that collections of language examples are fundamental, and observed generalizations of these examples, like grammar rules and parts of speech, are secondary.

The basic idea is that it is cheaper to keep a set of examples of structure, and spin off different perspectives as necessary, than it is to list all possible orderings of a collection of examples beforehand. Cheaper, that is, if the examples really are fundamental. Which given the

## Existing Pa Technology

three basic approaches problem of Natural Lan structure:

- Syn rule (ru syn wh sim ma abs Bro get rule bro get cat
- Sta fin ma cat aut
- Dis me - fu

- [Experienced industry professionals](#)  
["Industry professional" identified as a term].
- [Securities industry professionals](#)  
["Securities" breaks "industry professional" as a term].
- [A chief executive decision](#)  
["Chief executive" identified as a term].
- [The chief executive officer](#)  
["Officer" breaks "chief executive" as a term].
- [Far East consulting](#)  
["Far East" identified as a term].
- [Far East Asia](#)  
["Asia" breaks "Far East" as a term].
- [Currency exchange rates](#)  
["Exchange rates" identified as a term].
- [Foreign exchange rates](#)  
["Foreign" breaks "exchange rates" as a term].
- [London currency exchange](#)  
["Currency exchange" identified as a term].
- [Foreign currency exchange](#)  
["Foreign" breaks "currency exchange" as a term].

## Vector" Parser - Example

Estimated Pars

(a ((chief executive) decision))

/ \

a ((chief executive) decision)

/ \

**Note: you do not get this indeterminate and combinatorially idiosyncratic (infinitely sub-categorizable) quality in a system based on rules (e.g. a computer language)**

**There it is always possible to find a finite and unambiguous set of rules and classes which completely describe all the productions.**

**But only because all productions are exactly produced by those rules!**

**It will ALWAYS happen in a system based on examples. You will always have these combinatorial possibilities**

**The essential point is there is no single label which can summarize the several potentialities (collocational sub-categorizations) of the collection of examples defining "tea" or any other word.**

**This is an example of the fundamental incompatibility between arrangements of examples mentioned before:**

**i.e.**

**Letter sorted: A A A A F F F F**

**Colour sorted: A A F F A A F F**

**In the same way you can't find a single labelling over these two sets which reflects these two patterns, so no absolute labelling over language is possible which captures grammar perfectly.**

**We NEED to base language on collections of examples - the movie metaphor - to get this behaviour.**

## **Appendix I**

### **Allusions to unknowability in existing work**

## **Hopper - Emergent Grammar:**

<http://eserver.org/home/hopper/emergence.html>

**"The notion of emergence is a pregnant one. It ... takes the adjective emergent seriously as a continual movement towards structure, a postponement or 'deferral' of structure, a view of structure as always provisional, always negotiable, and in fact as epiphenomenal, that is at least as much an effect as a cause."**

**"Structure, then, in this view is not an overarching set of abstract principles, but more a question of a spreading of systematicity from individual words, phrases, and small sets."**

**"Because grammar is always emergent but never present, it could be said that it never exists as such, but is always coming into being. There is, in other words, no 'grammar' but only 'grammaticization'"**



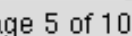
**Kenneth Pike ('50s Structuralist):**

**The structure of language is like that of a population of people...**

## **Appendix II**

### **Existing vector models of grammaticality**

**c.f. The grammar of a word expressed as its position in a word association space (Hinrich Schuetze):**



## **Similarity modelling - Dagan, Marcus, Markovitch '93:**

**Estimates probabilities from ad-hoc collections of examples**

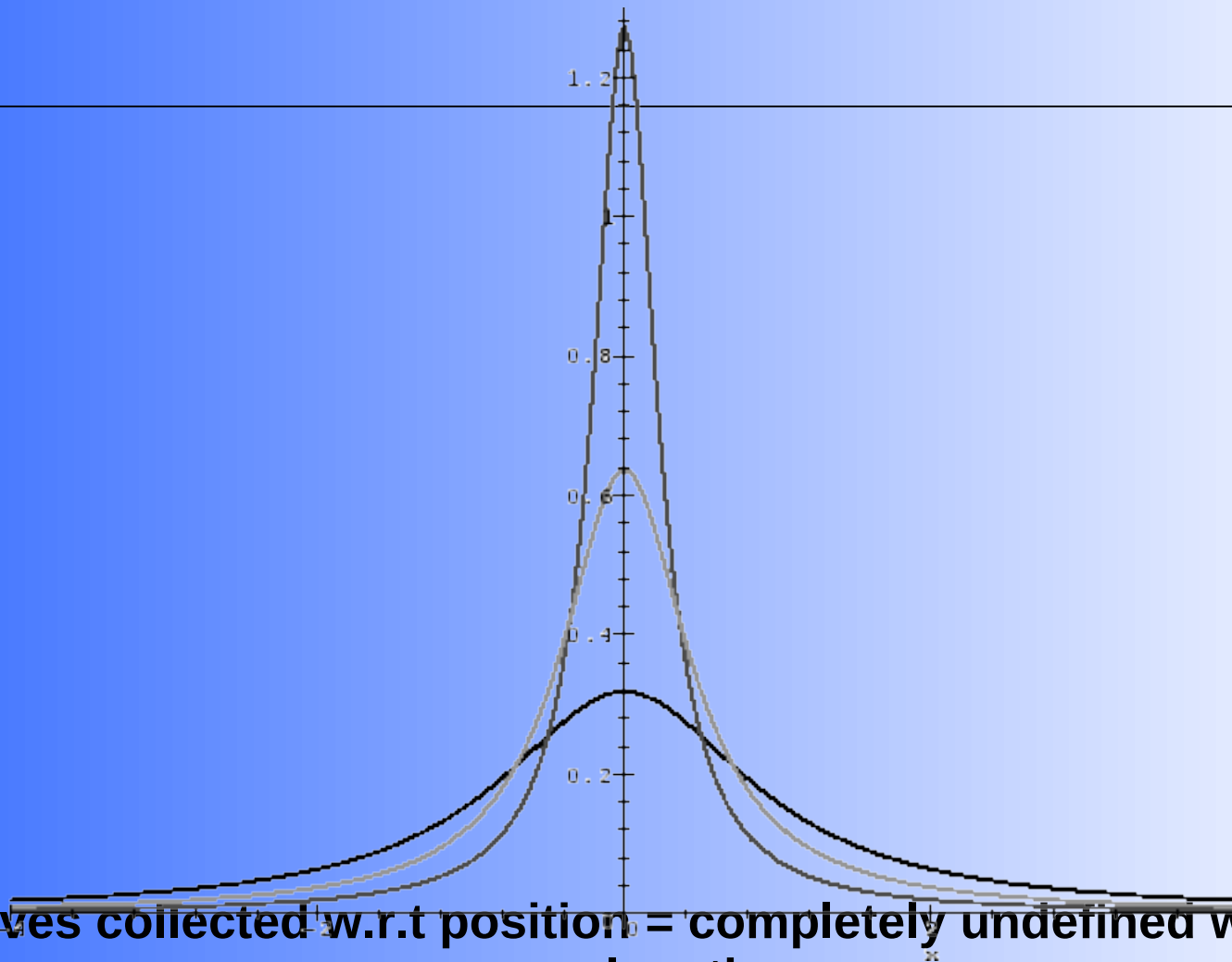
**"While traditional approaches, especially for semantic classification, have the view that information should be captured by the maximal possible generalizations, our method assumes that generalizations should be minimized. Information is thus kept at a maximal level of detail, and missing information is deduced by the most specific analogies, which are carried out whenever needed."**

## **Appendix III**

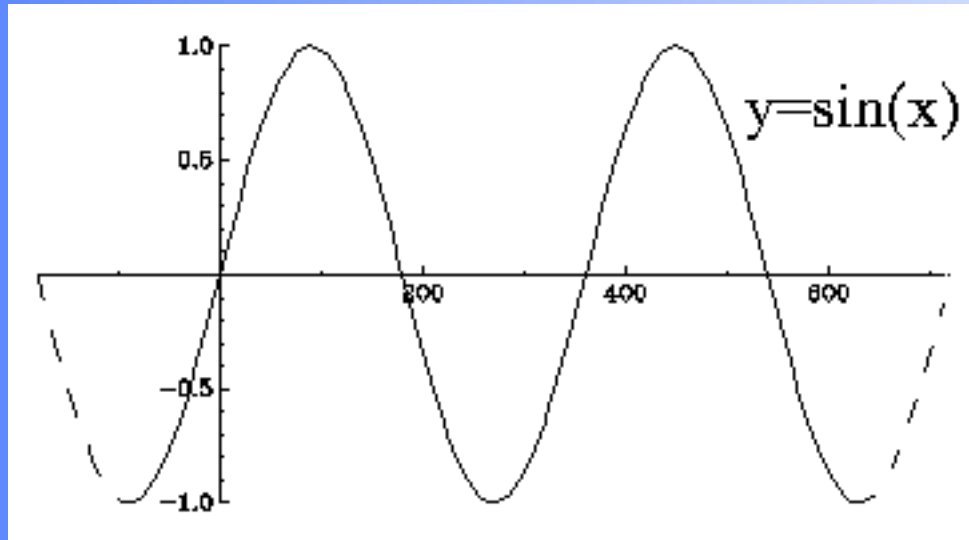
### **Other examples of "unknowability"**

#### **1) Sub-atomic Physics...**

**You can collect waves according to wavelength or position, but not both at the same time. Not beyond a certain constant.**



**Waves collected w.r.t position = completely undefined w.r.t. wavelength**



**Waves selected w.r.t. wavelength = completely undefined w.r.t position**

**This is the basis of the famous uncertainty principle of Quantum or Wave Mechanics in Physics - It is impossible to simultaneously know the momentum and position of a particle.**

**Hence my whimsical characterization of this approach as "wave linguistics"**

**(The parallel is not coincidental - they both result from degrees of freedom in ways of organizing collections.)**



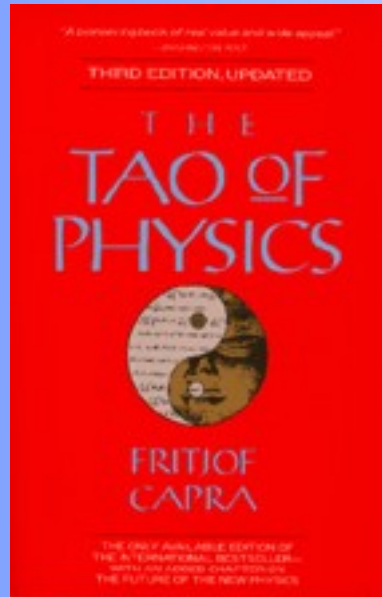
## **Other examples of "unknowability"**

### **2) Eastern Philosophy**

**Emphasis of practice over principle**

**"The one true Dao is the Dao which cannot be known"**

**There is a long observed parallel between this "unknowability" of modern Physics and Daoist philosophy. The same tension between two aspects of reality.**



**Could we now be seeing the same parallel between Daoist philosophy and our processes of thought expressed as language.**

**Collections of examples predict language will behave in this way.**