

**In my beginning is my end:
reflections on 45 years of NLP and
corpora.**

Yorick Wilks
Oxford Internet Institute
and
Computer Science Department
University of Sheffield
www.dcs.shef.ac.uk/~yorick
LREC-08 Marrakech

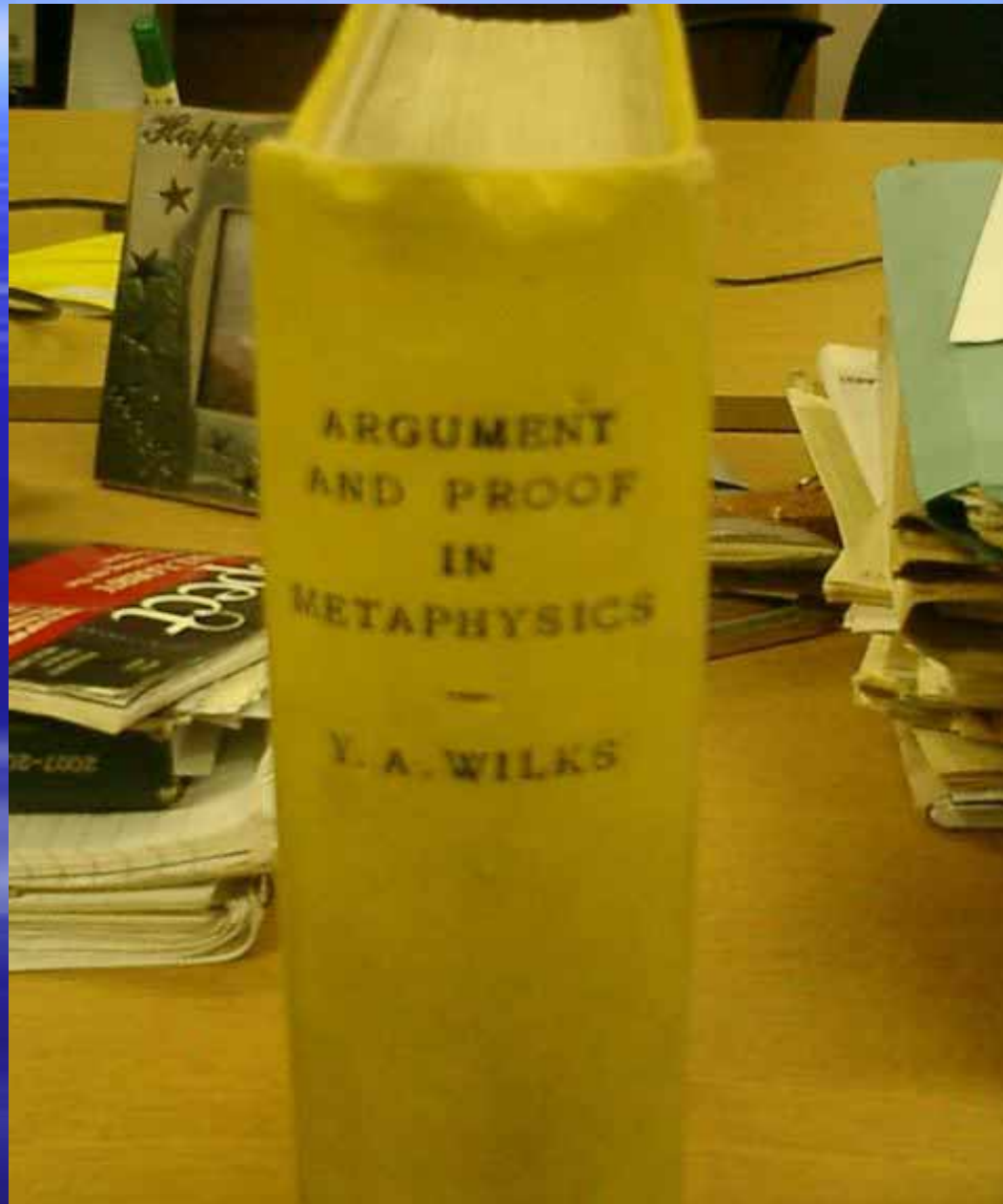
This talk is about:

- Starting with philosophical corpora, with (toy) newspaper control texts.
- Out of this came Preference Semantics, a quantitative, non-statistical, non-logical, non-syntactic, language representation.
- WSD, and novel sense as the continuing task.
- Adaptation to large scale methods and structure building: IE and the work of some students.
- Future hopes: big models and the Semantic Web
- Current work: dialogue corpora
- Return to the Holy Grail search: novel sense?

Zampolli and philosophical origins

- AZ began assisting the Padre Busa with the punch-carding of all St. Thomas Aquinas' works (over 2m words?);
- I also began in philosophy: my Cambridge thesis was called "Argument and proof in metaphysics, from an empirical point of view"
- It had an appendix with masses of LISP code

CAMBRIDGE PREFERRED ITS THESE BOUND IN BLUE OR BLACK



It had two sources:

- 1) Carnap's claim that meaningfulness in text could be determined by "logical syntax"--rules of formation and transformation (sound familiar?).
- Der Logische Syntax der Sprache (1936)
- My claim was that this was a bad demarcation and a better criterion of meaningfulness would have one interpretation rather than many i.e. that WSD was possible for a given text.
- Hence the "meaningless" text had too many interpretations rather than none (or one). A word in isolation is thus often meaningless.
- Preference Semantics was a WSD program to do just that, and provide a sense where WSD failed.

And.....

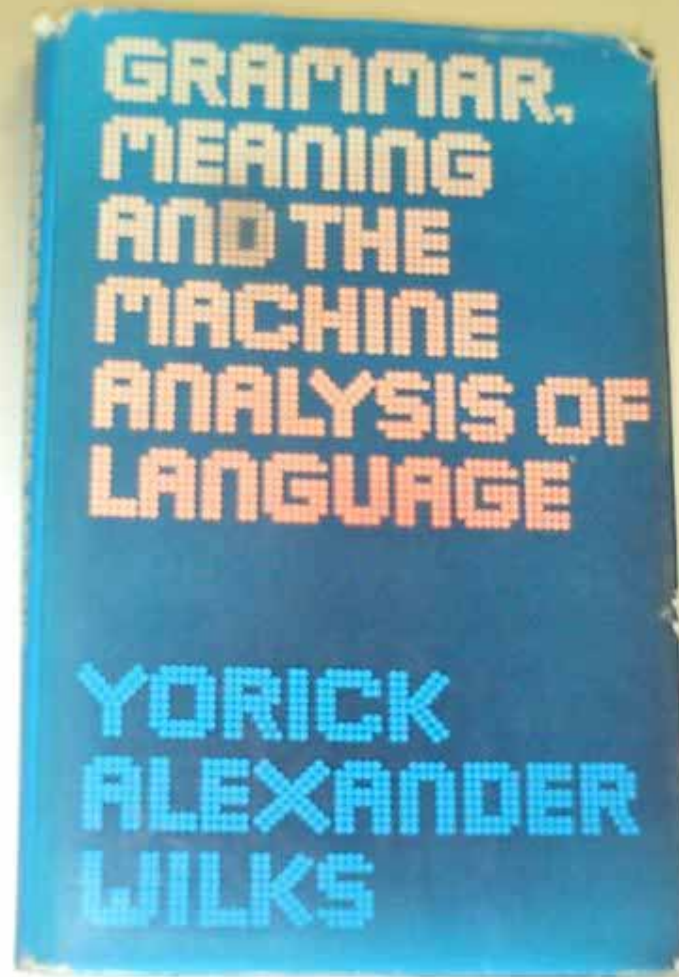
- 2) Bosanquet on the nature of metaphysical discourse.
- R. Bosanquet: Some Remarks on Spinoza's Ethics, MIND 1937.
- He argued that Spinoza's logical arguments are all false
- BUT what S. is doing is rhetorical: imposing a new sense on the reader ("Nature" for "God" for example.
- The thesis system tried to do WSD and detect a new sense, if WSD failed, by "sense construction"

MARGARET MASTERMAN (1910-1986)



“AI IS THE PURSUIT OF METAPHYSICS BY
OTHER MEANS” -----Longuet-Higgins

THE PREFERENCE SEMANTICS WITHOUT THE
METAPHYSICS: 1972



DICTIONARY
ENTRIES AS
PRIMITIVE
“FORMULAE”
FOR THE
WORDS OF
THE FIVE
TIMES
EDITORIAL
PARAGRAPHS-
--TOTAL OF
TEXTS ONLY
ABOUT 500
WORDS
(TYPES) BUT
THIS WAS
1966!

(FIGURE (((((WHERE SPREAD) (LINE GRAIN))
(FIGURE AS THE FORM OF SOMETHING IN SPACE))
(((MAN FOR) ((THINK THING) USE)) (FIGURE AS THINK))
(((COUNT SIGN) (LINE SIGN)) (FIGURE AS NUMERICAL CHARACTER))))
(FIND (((((MAN FOR) (((MAN FROM) THING) SENSE))
(FIND AS DETECT))))
(FIRST (((((COUNT SIGN) KIND) (FIRST AS INITIAL))
((UP KIND) (FIRST AS IMPORTANT))))
(FISHERMEN (((((BEAST USE) FOLK)
(FISHERMEN AS THOSE WHO CATCH FISH))))
(FISHERY (((((BEAST USE) KIND) (FISHERY AS TO DO WITH FISH))))
(FIVE (((((COUNT KIND) (FIVE AS HAVING IN NUMBER))
((COUNT SIGN) (FIVE AS NUMBER))))
(FIXED (((((NOTCHANGE KIND) (FIXED AS NOT ALTERING))
(((MAN FOR) (((((WHERE POINT) NOTCHANGE) THING) (BE CAUSE))
(FIXED AS PLACED IN A GIVEN POSITION))))
(FLEXIBLE ((((((WHERE SPREAD) CHANGE) CAN) KIND)
(FLEXIBLE AS ABLE TO HAVE ITS SHAPE CHANGED))))
(FORMED (((((BE CAUSE) LET) KIND)
(FORMED AS MADE THE WAY IT IS))))
(FORM (((BE CAUSE)
(FORM AS CAUSING SOMETHING TO BE THE WAY IT IS))
(((THINK STUFF) WANT) FOLK) (FORM AS A CLASS OF SCHOLARS))
(((THIS HOW) BE) CAUSE) SIGN)
(FORM AS WHAT CAUSES SOMETHING TO BE THE WAY IT IS))
((LINE GRAIN) (FORM AS OUTLINE))))
(FOR (((FOR BE) (FOR AS TO BE ASSIGNED TO))))
(FORWARD ((((((THIS WHERE) FROM) (WHERE CHANGE)) HOW)
(FORWARD AS IN A MANNER AWAY FROM SPEAKER))
(((THIS WHERE) FROM) (WHERE CHANGE)) KIND)
(FORWARD AS IN A DIRECTION AWAY FROM THE SPEAKER))))
(FREQUENTLY (((((MUCH WHEN) HOW) (FREQUENTLY AS OFTEN))))
(FROM (((FROM DO) (FROM AS AWAY))))
(FURTHERMORE ((NIL NIL))
(GENERAL (((WHOLE KIND) (GENERAL AS NOT SPECIFIC))
(((MUCH UP) MAN) (GENERAL AS HIGH RANKING OFFICER))))
(GOD (((((MUCH UP) MAN) (GOD AS A SUPERHUMAN PERSON))
(WHOLE WORLD) (GOD AS EVERYTHING THAT THERE IS))))
(GO (((WHERE CHANGE) (GO AS TO MOVE IN SPACE))))
(GREATER (((MORE KIND) (GREATER AS LARGER))))

SOMEWHAT
LATER
BOGURAEV
DECLARED
THE
AVERAGE
NLP LEXICON
WAS 36
WORDS!!

164 *Practical semantic experiments*

parsing pattern for the paragraph so that an operator of the system can see which 'semantic compatibilities' between fragments have given rise to any particular resolution. If it can find no resolutions at all, for any frame, PARSPARA prints at the typewriter (NO RESOLUTION ALL PATHS BLOCKED).

Below is a sample output from the overall resolution of the first paragraph, printed out at the teletypewriter.

*Sample output from the semantic parser
(first seven fragments of the first paragraph)*

```
((BRITAINS TRANSPORT SYSTEM ARE CHANGING)
((WORDS RESOLVED IN FRAGMENT)
((TRANSPORT AS PERTAINING TO MOVING THINGS ABOUT)
(BRITAINS AS HAVING THE CHARACTERISTIC OF A PARTICULAR PART
OF THE WORLD)
(SYSTEM AS AN ORGANIZATION)
(ARE AS HAVE THE PROPERTY) (CHANGING AS ALTERING)))
((WORDS NOT RESOLVED IN FRAGMENT) NIL))
((WITH IT THE TRAVELLING PUBLICS HABITS)
((WORDS RESOLVED IN FRAGMENT)
((TRAVELLING AS MOVING FROM PLACE TO PLACE)
(IT AS INANIMATE PRONOUN)
(PUBLICS AS CONNECTED WITH THE WHOLE PEOPLE)
(HABITS AS REPEATED ACTIVITES)))
((WORDS NOT RESOLVED IN FRAGMENT) NIL))
((IT IS THE OLD PERMANENT WAY)
((WORDS RESOLVED IN FRAGMENT)
((IT AS INANIMATE PRONOUN)
(IS AS HAS THE PROPERTY)
(OLD AS HAVING BEEN THROUGH A LOT OF TIME)
(PERMANENT AS UNCHANGING) (WAY AS PATH OR ROUTE)))
((WORDS NOT RESOLVED IN FRAGMENT) NIL))
((WHICH ONCEMORE IS EMERGING)
((WORDS RESOLVED IN FRAGMENT)
((ONCEMORE AS ONE MORE TIME)
(IS AS HAS THE PROPERTY)
(EMERGING AS MOVING OUT OF SOMETHING)))
((WORDS NOT RESOLVED IN FRAGMENT) NIL))
((AS THE PACEMAKER)
((WORDS RESOLVED IN FRAGMENT)
((AS AS CORRESPONDING TO)
```

6. METAPHYSICAL AND PHILOSOPHICAL ARGUMENTS

"What we do is to bring words back from their metaphysical to their everyday usage".

Wittgenstein, Investigations (116).

6.1 In this section I discuss the results of the computer application of the system of section 4E to five metaphysical paragraphs. The main point of applying the system to them is to apply the associated criterion of meaningfulness: a criterion proposed in section 2H, discussed in some detail in section 3E, and stated precisely by reference to a system of analysis in section 4E.

The system was applied to the metaphysical texts in exactly the way in which it was applied to the five editorial texts described in section 4E. Everything said here rests on the fact that the system applied was found to work satisfactorily with texts in 'ordinary English', and it is in that sense that I intend the application of Wittgenstein's description of his own activity to the present activity; "bringing words back from their metaphysical to their everyday usage", by displaying a procedural continuity from one to the other.

The general result is briefly described: the system worked perfectly well for three of the texts and resolved



FORMULAS WERE TREES OF PRIMITIVES THAT REPRESENTED WORDS SENSES ---MANY PER WORD FOR MOST WORDS-----AND THE FORMULA HEADS HAD TO FIT TEMPLATES OF TRIPLES LIKE

MAN -HAVE -THING

Y.Wilks, Text Searching with Templates, Cambridge Language Research Unit, ML 154, 1964.

A (PARTIAL) INFERENCE RULE CONNECTING TEMPLATES OF (TREE) FORMULAS

I 1. $(((*ANI\ 1)(SELF\ IN)\ (MOVE\ CAUSE))\ (*REAL\ 2)) \rightarrow (((*JUDG)\ 2)$

or, in 'semi-English',

$\{animate-1\ cause-to-move-in-self\ real-object-2\} \rightarrow \{1\ * judges\ 2\}$

I 2. $(1\ BE\ (GOOD\ KIND)) \rightarrow ((*ANI\ 2)\ WANT\ 1)$

or, again,

$\{1\ is\ good\} \rightarrow \{animate-2\ wants\ 1\}$

Principles of Preference Semantics

- Process (vs. Schank and diagrams)
- Affinity and repulsion of senses (cf. Waltz+Pollack's WSD connectionism)
- "Best fit" interpretation--one with most satisfied preferences
- Least informative/effort interpretation
- No explicit syntax, only segmentation and order
- Meaningfulness as enabling interpretation choice
- Meaning as other words
- Gists or templates as underlying entities
- No correct interpretation or set of primitive concepts (vs. Schank)

Another way of putting these is:

- Semantics is not necessarily deep but also superficial (cf: results on WSD and POS)
- Quantitative phenomena are unavoidable (cf. McCarthy on AI)
- Reference structures (like lexicons) are only temporary snapshots of a language
- What is important is to locate the update mechanism of language, the creation of new word senses, which is NOT Chomsky's sense of the creativity of language.

Preference affinities eventually became quantitative and empirically based

- (Briefly) with connectionism:
 - Waltz said “things are going your way”
 - Connectionist sense graphs of affinity and repulsion of senses, but no data
 - Cf. D. Beeferman, A. Berger, and J. Lafferty, A Model of Lexical Attraction and Repulsion, Proceedings of the ACL-EACL'97 Joint Conference, Madrid Spain, 1997
- Later, with corpora, and the statistical derivation of preferences:
 - Resnik
 - Grishman
 - Lehnert

A long career break away from issues of corpora

- Working with MT systems (EUROTRA, ULTRA, XTRA, PANGLOSS)
- Belief systems (VIEWGEN, with Janusz Bien, Afzal Ballim and others) based on principles cognate with PS:
- That a machine belief had to be something that implied choice (like PS): that a belief could in principle be seen to be different from one held by another entity
- Thus an ATM does NOT have any beliefs about what is in my account because it has no alternatives to choose among (e.g. my beliefs).

Like many I was slow to see the relevance of Leech's work on POS-tagging

- Which we now all take as a basic process
- The PANGLOSS MT system was funded by DARPA in direct competition with Jelinek's CANDIDE system
- I came to see how important what he was doing was--as we (nearly!) all did in time.

Extending the Jelinek revolution to the highest linguistics levels from corpora

- Building STRUCTURE empirically----not present in CANDIDE
- Continuity with work with Guthrie and others in 1980s--NSF program on the construction of an ontology empirically from LDOCE--to empirical work on ontologies from texts

Corpus-related work of some students I must mention

- (with Mark Stevenson) multi-engine WSD produced best figures over all text words ; the engines did WSD with different types of semantics, and showed that the types of semantic information were at least partially independent, since their sum (under ML) was better than their parts; original test text. (Comp Ling. 2000)
- (With Nick Webb) High levels of Dialogue Act tagging with a very simple algorithm. (LREC-08)
- (with Christopher Brewster) Abraxas which learns proto-ontologies from corpora. Justifying the semantic terms of an ontology as non-a priori. (LREC-04)

A diversion from Wittgenstein to experiment:

- Wittgenstein said ask for the use not the meaning..
- Where would be better to look than the Web, as its usage is now so much larger than any human's (though we then give up any pretence of cognitive explanation).
- Roger Moore on a hundred year's of human training to get a modern speech recognition system.

Computation over the whole web- as-corpus can make Wittgenstein's "usage" a real concept:

- Jelinek on language as a "system of very rare events" = data sparsity
- BUT how much language to give a full model of a language?
- Trigram model may require a trillion words of training (=60,000 reading years)--loss of any connection to human language activity but still useful: see next figure-->



31/05/05

REVEAL

But does the 1.5 billion word corpus
(at 70%+ trigram coverage) show
things aren't so bad?

- By extrapolation it would need 75×10^{10} words to give 100% trigram coverage
- Our 1.5bn corpus at 74% was 15×10^8 , and Greffentette in 2003 calculated there were over 10^{11} words on English on the web then (i.e. about 12 times what Google indexed). The (English) web must be big enough by now.
- Since the whole web is hard to get at, could we go another way?
- Work with Louise Guthrie and other colleagues on Skipgrams

Skip-Gram Example

Chelsea celebrate Premiership success.

tri-grams:

Chelsea celebrate Premiership
celebrate Premiership success

one-skip tri-grams:

Chelsea celebrate success

Chelsea Premiership success

The moral here is that recent work shows data sparsity for training may not be quite as bad as we thought:

- Trigram models give smooth natural output which rule grammars never can
- Trigrams are very good models for speech, not so good for meaning (but translations?)
- Skipgram models may compensate for size at the risk of nonsense--but that isnt so---see below:

Skip-grams prove their worth

- **Using skip-grams can be more effective than increasing the corpus size!**
- In the case of a 50 million word corpus, similar results are achieved using skip-grams as by quadrupling corpus size.
- This illustrates a serious possible use of skip-grams to expand contextual information

Some talk of synthesis of statistical and symbolic methods

- But which symbolic methods would they want to participate in that synthesis?
 - See Clarke and Pulman, Combining symbolic and distributional models of meaning. AAI spring symposium, Stanford 2007
 - A reductio ad absurdum? Use skipgrams to map proto-facts? Cf. Lafferty, J, Sleator, B. and Temperley, D. Grammatical trigrams, AAI Fall Symposium 1992
- I would claim there is still a proper tradition of non-logical non-statistical semantic representation (Schank and I presented in the 1960s) and that is the only form of representational semantics--at the moment---showing the ability to move to phenomena of scale.

Superficial semantics re-surfaced as Information Extraction (>1990) and now flexing its muscles as the RDF format of the semantic web.

- IE has been thoroughly despised by both formal logicians (as not a “deep” representation”) and
- by statisticians--till it took on ML
- and by IR---as an attempt to smuggle NLP back
- But it has survived

IE and automated annotation factories at Sheffield (>1993)

The GATE NLP platform and tools project
(EPSRC, DARPA, the EU Semantic Content programme)

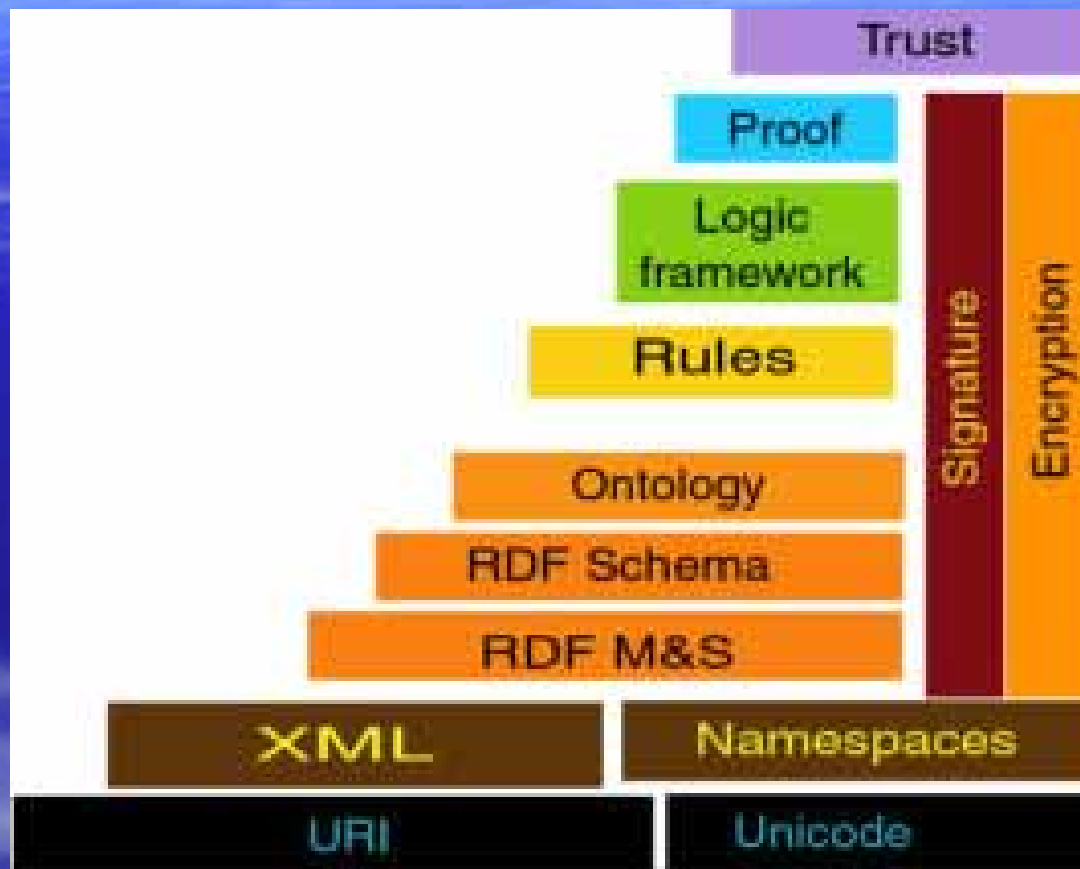
Gaizauskas

Cunningham

Ciravegna

The Semantic Web project IN EUROPE has grown from IE and annotation, and has roots going back into TEI and SGML.

What is the relationship of the
Semantic Web to NLP?



The Semantic Web and AI/NLP

- Some look at the top of the SW-pyramid and say “it’s just GOFAI isnt it”
- My case will be that if you look at the bottom of the pyramid, the SW rests much more on NLP than is usually realised.
- Of course, NLP people would say that, wouldn’t they, but they may be right!

The Semantic web

(Berners-Lee, Hendler, and Lassila, Scientific American
2001)

- A vision of making the Internet as readable by computers (agents) as it is by us.
- A similar notion to the “ascent” in the semantic web pyramid---meaning/interpretation somehow trickling UP it from the bottom (cf. Braithwaite’s view of scientific theories--neutrinos linked to experiment)
- Is this last what SW people mean/want, or do they assume that the higher-level structures are self-interpreting?

Forms of knowledge in the SW

- 1) Universal Resource Indicators (URIs)
- 2) Resource Description Format
- RDF triples---putting all facts in the form:
 - John-LOVES-Mary
 - Not quite logic yet! Basically IE output; explicitly called « subject » « object »!
- 3) Ontologies---trees of concepts in hierarchical and functional relations: again like
 - Canary-ISA-Bird
- 4) DAML/OIL reasoning languages

The bottom levels of the SW
always sit on NLP annotations of
objects and actions (i.e. IE)

- Classic IE detects named entities--populates SW's "namespace"
- Does semantic type annotation
- Detects actions
- All recent IE works with an ontology
- It is also, of course, a SW annotator and RDF finder

Semantic Web getting real??- Slashdot 11/2/08!

- Reuters just [1]opened access to their[2]corporate semantic technology [3]crown jewels. For free. For anyone. Their [4]Calais API lets you turn unstructured text into a formal RDF graph in about one second. I ran about 5,000 documents through it and played with a subset of them in [5]RDF-Gravity. The results were impressive overall. Is this the start of the [6]semantic web getting real? When big names and big money start to act, [7]not just talk, it may be time to pay attention. Semantic applications anyone?

The philosophical problems may or may not just vanish as the SW pushes ahead with annotations!

- David Lewis and « markerese »: his 1970s critique of Fodor and Katz, and of any non-formal semantics (such as semantic type annotation)
- The Semantic Web takes this head on and carries on, hoping URIs and « popping out of the virtual world » (e.g. giving the web your phone number!) will solve semantic problems.
- Can all you want to know be put in RDF triples, and do the reasoning with them?
- But agents so-based do seem to work--*Eppur si Muove!*

The problem of Recoding Content

- Another version of the Lewis argument against Fodor/Katz semantics:
- Sparck Jones said there are only words and content cannot be recoded
- (KSJ auction catalog example) "A Charles II parcel-gilt cagework cup, circa 1670"
- What, she asks, can be recoded beyond {object type: CUP}?
- The rest, she says, is the English.
- BUT THERE ARE DICTIONARIES-----she cannot be right

Taking stock here: three views of what the SW is:

- 1) An updating of the old AI dream of representing everything in logic for reasoning over the world (GOFAI); actually the SW is much less sophisticated than that--it has traded representation power for tractability.
- 2) An apotheosis of annotation in IE, attempting to build up to concepts in ontologies for e.g. scientific knowledge by very large shallow computations over texts: problem of grounding the terms other than in texts, and tying the general concepts plausibly to the distributions of usage in text.
- On this view the SW is the WW of text **plus meanings**.
- 3) A system of trusted data bases that ground meanings in something close to objects (TB-Ls own view?)
- This is close to Putnam's view that scientists are Guardians of Meaning.

AI was always right that language interpretation was driven by, responsible to, a knowledge base

- But what they failed to see was that the KB is far closer to the language than they realised.
- “Ontological predicates” are not safe from change in a modern KB
- The Semantic Web adventure makes it more likely that this will become clear than the old AI paradigm--this one really has lots of language processing and knowledge in it.

What is the way out of wanting data/usage and concepts?

- Maybe “words as they stand” (Sparck Jones) but perhaps not all words are equal
- Some words may be aristocrats, not democrats
- Perhaps “semantic primitives” are just words but also special words: forming a special language of translation, that is not pure but ambiguous, like all language.
- If that is so, perhaps we can have explanations, innateness (even definitions) on top of an empiricism of use.
- The consequence (for logic , ontology and the SW) of annotations also being words

Empirical semantics and facts

Imagine extracting from a vast corpus all forms of Agent-Action-Object triples (i.e. all examples of who does what to whom etc.)--sites like USC/ISI (Pantel) are doing this on a vast scale with shallow parsing.

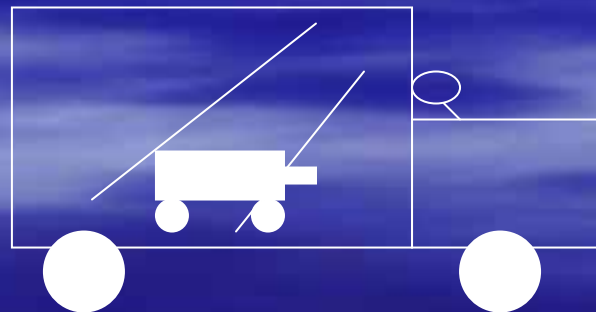
- Use these to resolve ambiguity and interpretation problems of the kind that obsess people who are into concepts like 'coercion' 'projection', 'metonymy' etc. in lexical semantics.
- E.g. if in doubt what 'my car drinks gasoline' means, look at the stored triples about what cars do with gasoline and take a guess.
- As we noted, skip-grams may be an even simpler way of reaching them.

The man drove down the road in a car

((The man)(drove (down the road)(in a car))))



((The man)(drove(down the road(in a car))))



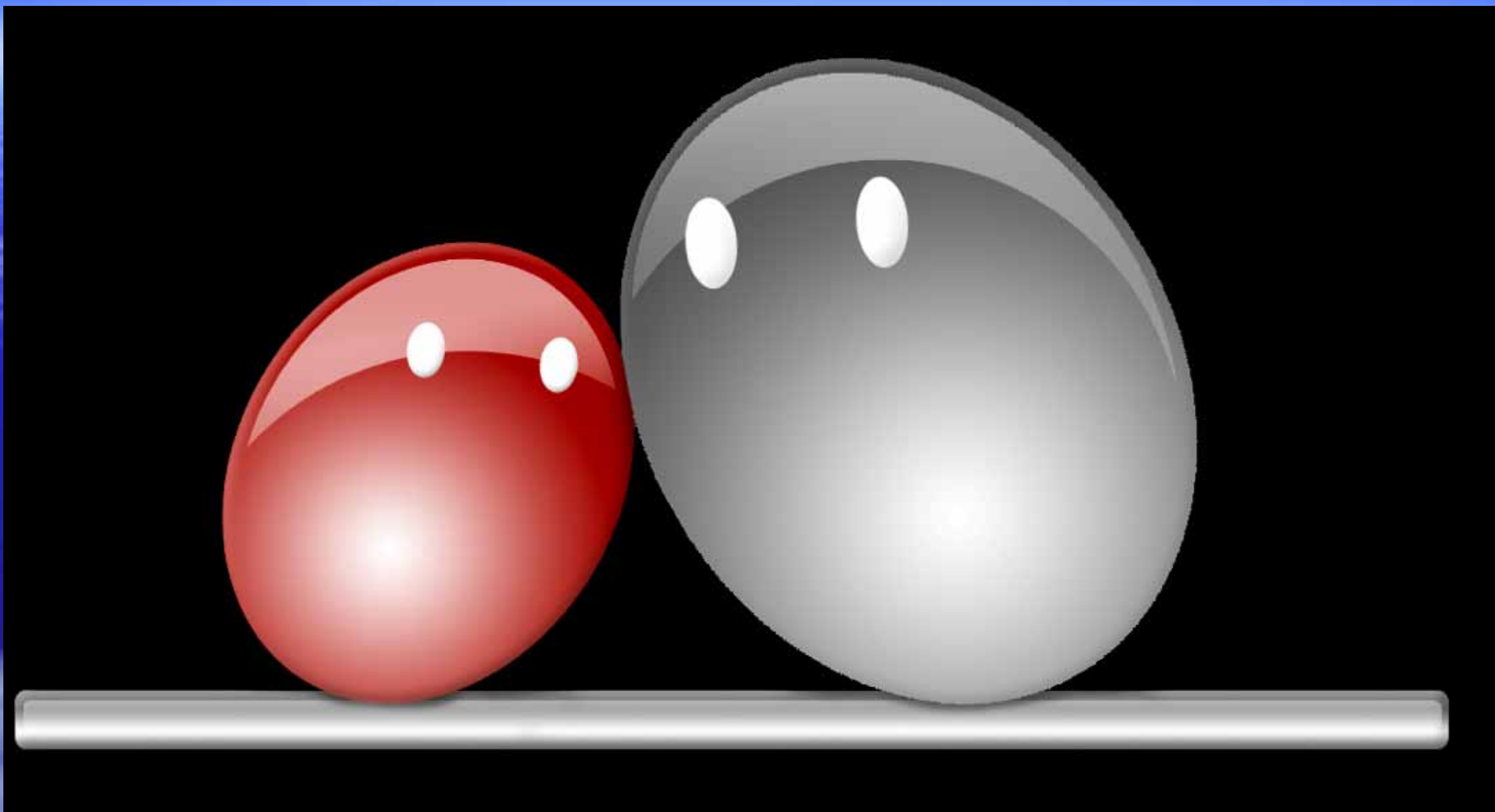
Forms-of-facts as a reality?

- This isn't a very good algorithm, but it should stir memories of Bar Hillel's (1959) argument against the very possibility of MT, namely that you couldn't store all the facts in the world you would need to interpret sentences
- AI has always believed you could
- Modern empirical corpus linguistics suggests you can find vast numbers of facts and use them.
- Is this also a way of doing classic AI "knowledge-based understanding" on the cheap? CYC without People...!
- This is one view of SW's RDF (Wikipedia as RDF!)

- Early Preference Semantics (1967), as we saw, did interpretation by means of a list of all possible interlingual Agent-Action-Object triples! (only I made the list up!)
- These were intended then as Wittgensteinian forms-of-facts; but now they can be extracted automatically, by Information Extraction technology
- and are also very close to the RDF triples (Subject-Relation-Object) underlying the second generation Berners-Lee “Semantic Web”.

All such methods in NLP are “Big Data Small Program”

- Which is another term for corpus-orientated linguistics
- Connectionism
- Grefenstette’s “vast lexicon”
- The Semantic Web
 - Etc. Etc.
- Recall Sparck Jones’ observation on lexical entries in Montague Grammar
 - Meaning of Life is Life’
 - Meaning of The is ENORMOUS

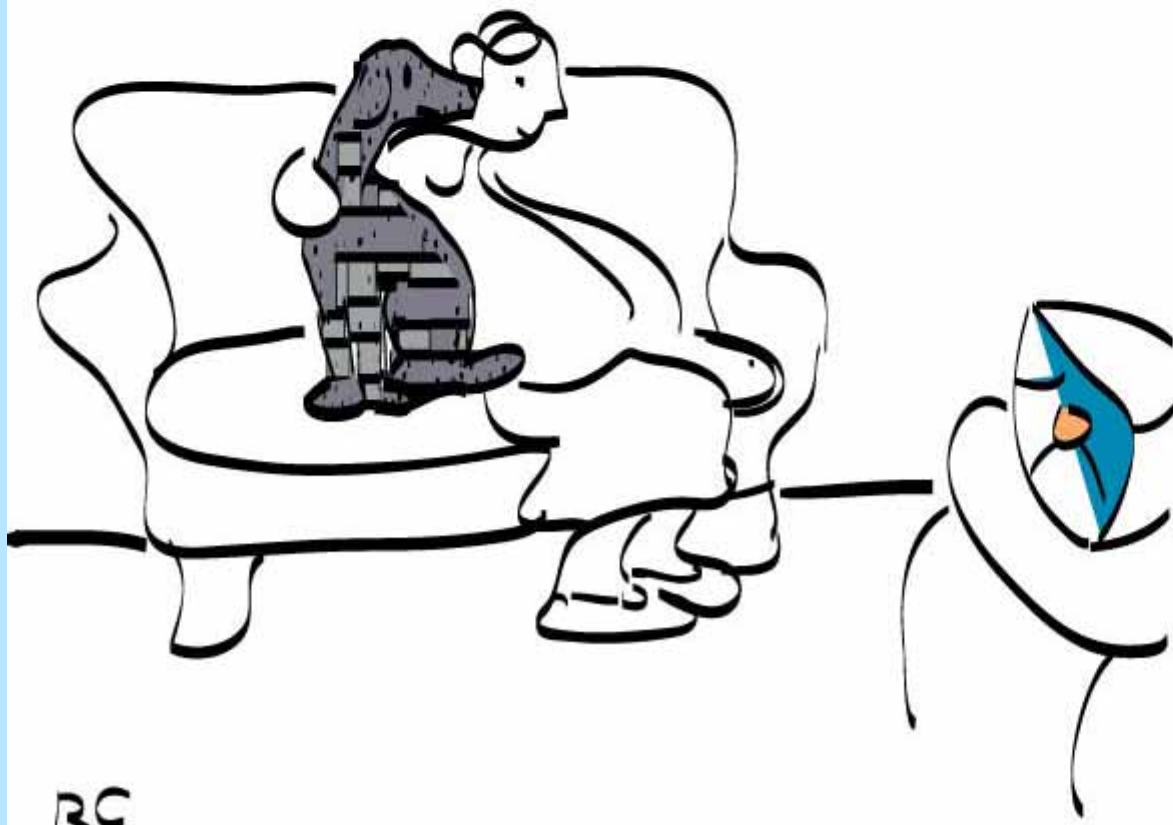


COMPANIONS

A series of intelligent and sociable COMPANIONS

- The SeniorCompanion

- The EU will have more and more old people who find technological life hard to handle, but will have access to funds
- The SC could sit beside you on the sofa but be easy to carry about--like a furry handbag--not a robot
- It will explain the plots of TV programs and help choose them for you
- It will know you and what you like and don't
- It will send your messages, make calls and summon emergency help
- It will read you the news or tell a joke when you're bored



RC

Early evidence of the acceptability of this kind of companion

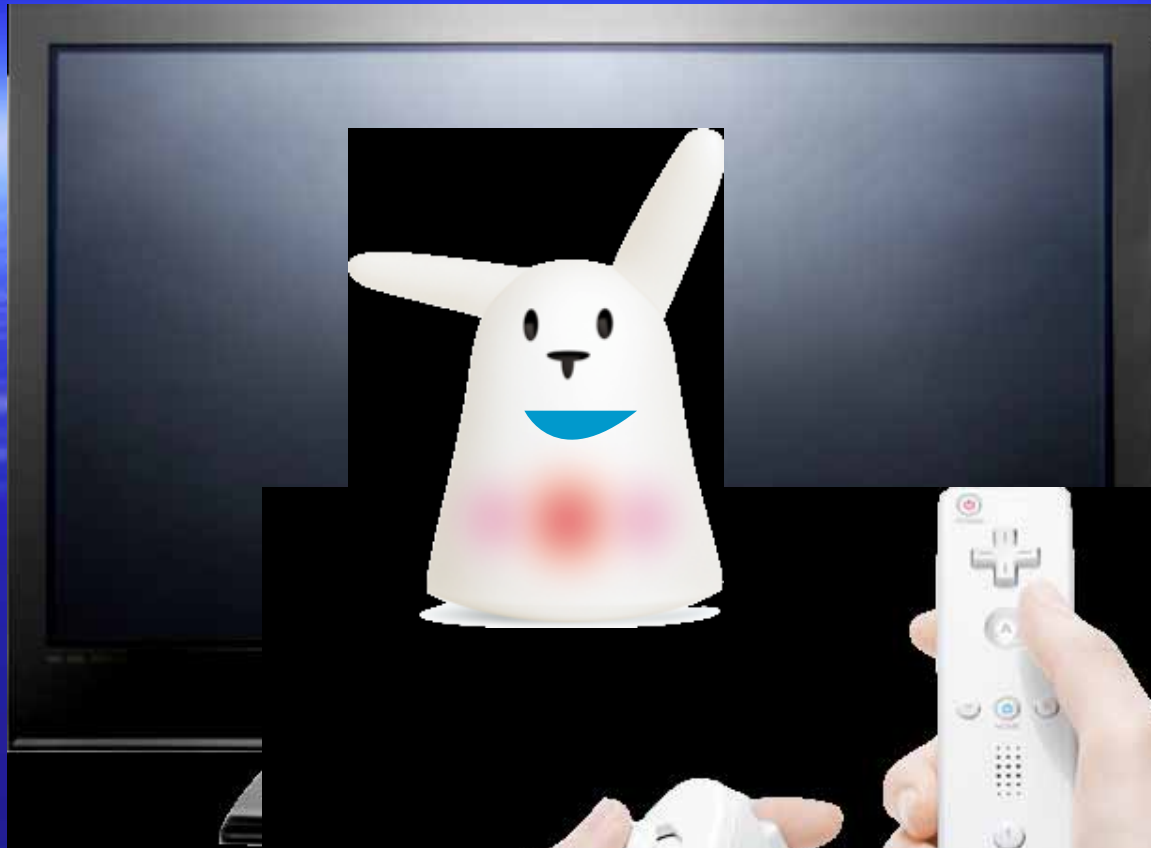
- Remember Tamagochi?
- Quite intelligent people rushed home to feed one (and later Furby) even though they knew it was a simple empty mechanism.
- And Tamagochi could not even talk!
- People with pets live longer.
- Wouldn't you like a warm pet to remind you what happened in the last episode of a TV serial?
- OK, you would not, but perhaps millions of your compatriots would?!

Akino
and
Primo
Puel



So what will these Companions be like?

Persistent
Multimodal
Pro Active
Knowledgeable
Engaging
Rich Personality
Polite
Humorous





A familiar face...and a familiar voice.



A familiar face...and a familiar voice.

The Companions project is really about language, and relationships and emotion expressed through language

- Which means it ISNT about robots
- Which cuts us off from lots of fun slides
- And discussions like David Levy's--->



But seriously, and back to corpora:

- Dialogue corpora for ML in a domain are hard to find so have to be specially constructed (we are);
- We are seeking to induce Dialogue Manager script structures from these corpora;
- We are investigating using SW-style RDF as the knowledge base
 - open to info about topics the user mentions and about which we know nothing
 - And which we can draw down as RDF from wikipedia, say

Conclusion: my longest term research interest has been lexical novelty---new senses of words

- Are there genuinely novel senses of words (metaphors, metonymies, or whatever) that cannot be captured by computational Word Sense Disambiguation (WSD)?
- In a sense, they obviously cannot be captured by WSD because, by definition, we cannot mark up for novel senses.
- Locating novel senses must therefore be another kind of process.

Traditions of coping with usage outside the norm

- Karen Sparck Jones thesis (1964): see if word falls in a new bottom up cluster based on classification.
- Hanks/lexicography (ditto): declare them « exploitations »
- Lakoff/Martin (1970s): fit them into a number of predetermined patterns of standard metaphor
- Briscoe/Peters (after Givon 1966): use a resource (LDOCE/WN) to see if they fit under a generalization of (possibly unrealised) « systematic polysemy or metonymy » (e.g animals become foods)

More outside-norm approaches

- Some form of adapting to the nearest norm sense and interpreting in terms of it:
 - Pustejovsky (The Generative Lexicon)
 - Nirenburg (Mikrokosmos)
 - Fillmore (Framenet)
- Most call on some notion of stereotype
- But if so conformed, how then can a sense be novel?

At the end, and back to the Holy Grail of mapping/discovering novel sense:

- My (then) student. David Guthrie, and I concentrated on verbs and their object words (in the BNC) where both were frequent (I.e. avoiding rare words which give separate problems--the issue here is only combinatorial!)
- We looked for ones not present at all in 1990, once in 1991-2, but occurring more than 8 times in 1993.
- More on this in article “Lexical tuning”
- (with R.Catizone: Jnl. Semantics 2002)

Books made:	358, 15822
Eyes studied:	4040, 483
Police closed:	2551, 1774
Directors make:	340, 3757
Phone began:	328, 3654
Body opened:	1612, 2176
Probe follows:	78, 3581
Mouth became:	816, 2816
Look says:	644, 2976

The desire to be present at the birth of a novel sense!

We shall not cease from exploration
And the end of all our exploring
Will be to arrive where we started
And know the place for the first time.

T.S.Eliot
Four Quartets