Generating Verbal Humour

Graeme Ritchie

University of Aberdeen

This talk....

- Why study humour?
- Review of humour generation programs
- Some comments on the field

Long Term Perspective

Humour is complex and unexplained.

A precise, detailed, computationally testable theory would be desirable.

For the moment, small scale investigations are all we can manage.

Why Study Humour? (1)

Could throw light on:

- how people categorise the world
- how people reason
- how people interact
- factors affecting health

Why Study Humour? (2)

To enable the design of better human-computer interfaces and intelligent agents.

Computer systems that:

- **are "lifelike"**
- **detect potentially funny situations**
- **understand user's remarks**

Why build computer models of humour?

Science:

- to explore the workings of humour theoretically
- to gain insight into human intelligence

Engineering:

- to produce practical applications

Practical Applications?

Automatically creating:

- advertising slogans
- jokes for Christmas crackers, greetings cards

In computer user interfaces:

- making error messages less brusque
- general lightening of atmosphere
- more life-like "agents"

Interactive education:

- teaching simple concepts, particularly language, e.g. to children.

Methodological Consideration 1.

Different ROLES for the program

A humorous agent: able to create (or recognise) humour in context. Arguably, it must be "intelligent".

A model-tester: explores some abstract theory of humour, showing its effects (e.g. what jokes can be constructed from a given set of rules?).

Methodological Consideration 2.

Different TASKS for the program

A joke recogniser: Since it is likely to be using knowledge that corresponds to only a small subset of all jokes, realistic testing is difficult to arrange.

A joke generator: The only claim to be tested here is that all the output which the program offers as a joke is indeed a joke. This is much more amenable to testing.

Since 1992, a number of word-play programs:

- The "Tom Swifty" generator (Lessard & Levison 1992)
- A simple pun generator (Lessard & Levison 1993)
- The JAPE riddle generator (Binsted 1994,1996)
- The Homonym Common Phrase Pun system (Venour 1999)
- The WISCRAIC pun builder (McKay 2000, 2002)

The State of the Art

Most implemented systems:

- are on a very small scale
- are not strongly oriented towards either "science" or "engineering" motivation
- are not directly tied to any humour theory
- manipulate language
- are generators, not recognisers, of humour
- rely on ad hoc, not general purpose, language processing

Evaluation

A critical question for a joke-generator:

how good are the results?

Usually determined by showing results to human subjects.

It is necessary to do this in a carefully controlled way.

What questions should the "judges" be asked?

e.g.

Is this a joke? (YES/NO/NOT SURE)

How funny is this? (On a scale of...?)

Do you think this joke was computer-produced? (YES/NO/NOT SURE)

How does this joke make you feel?

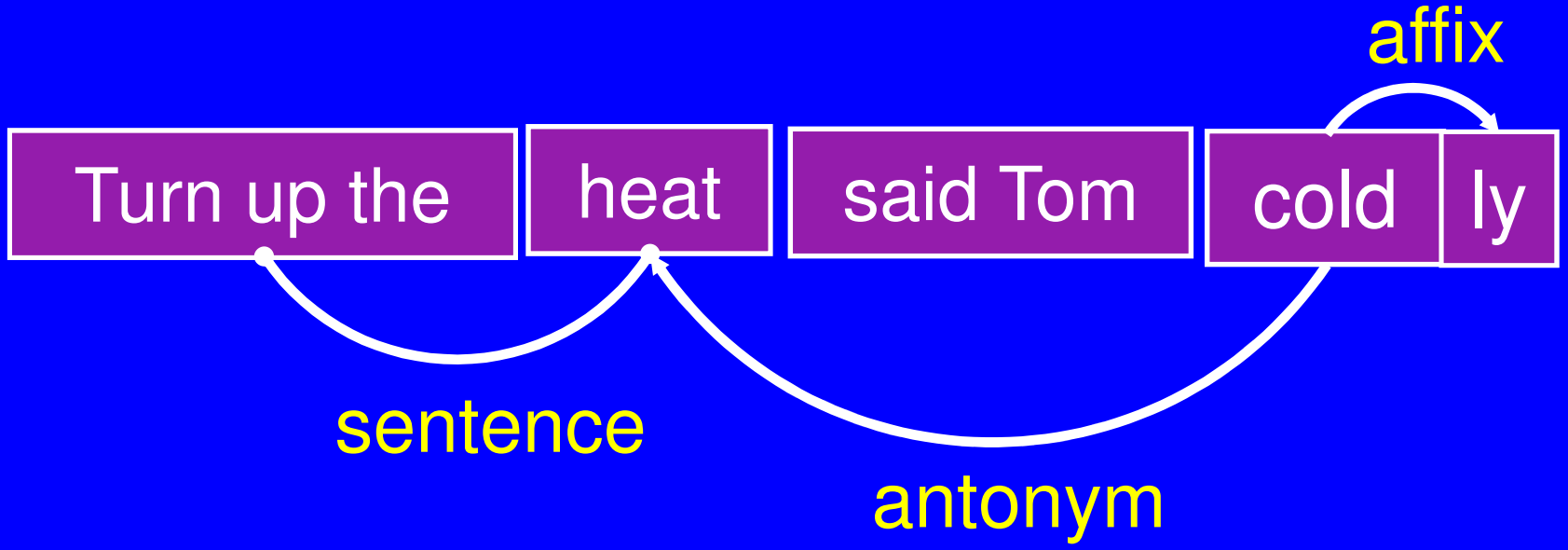
(OFFENDED/AMUSED/ PUZZLED/ETC)

Lessard & Levison 1992

Tom Swifties : quoted sentences with a *'said Tom'* tag and a punning adverb.

"I hate chemistry," said Tom acidly.

"I am not a girl," said Tom boyishly.


affix
Turn up the
heat
said Tom
cold
ly
sentence
antonym

WisCraic (McKay 2000, 2002)

Simple one-sentence puns based on semantic associations of words.

The performing lumberjack took a bough.

The strong fisherman showed the woman his mussels.

The performing lumberjack took a bough.

The [ADJECTIVE] [PERSON] took a bow .

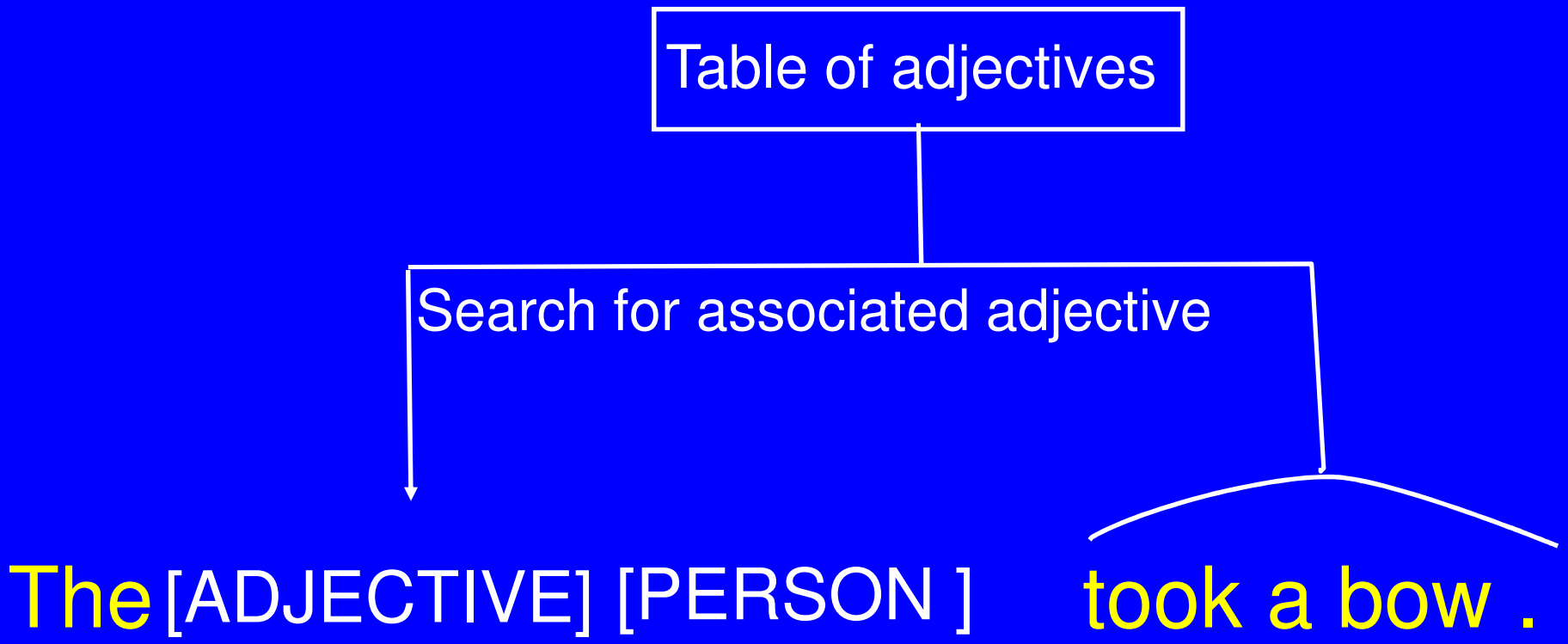


Table of adjectives
Search for associated adjective
The [ADJECTIVE] [PERSON]
took a bow .

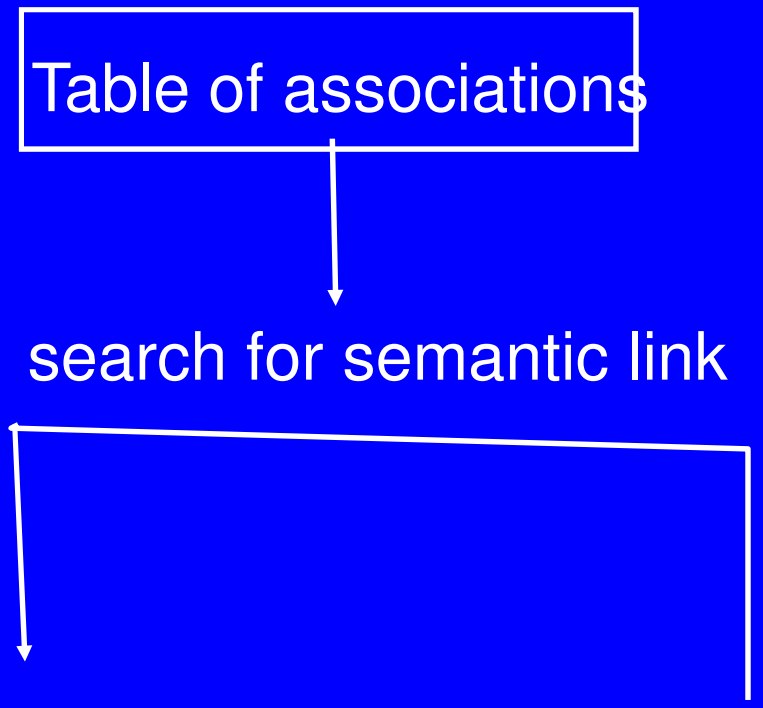
The performing [PERSON] took a bow .

bough

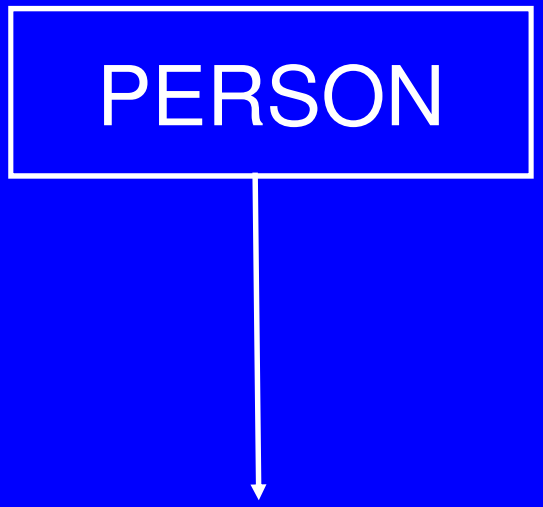
search for homophone

The performing [PERSON] took a bow .

The performing [PERSON] took a bough.



The performing [PERSON] took a bough.



The performing lumberjack took a bough.

The performing lumberjack took a bough.

Wiscraic:

Subjects shown a mixture of type of item.

Subjects asked:

Is it a joke?
If so : Funniness Rating, Cleverness Rating
If not: Obscure words/ Ordinary sentence/ Nonsense/ Failed joke attempt

Results:

84% were jokes
Funniness and Cleverness correlated
40% of those judged to be jokes were low in Funniness

Wiscraic evaluation - some weaknesses

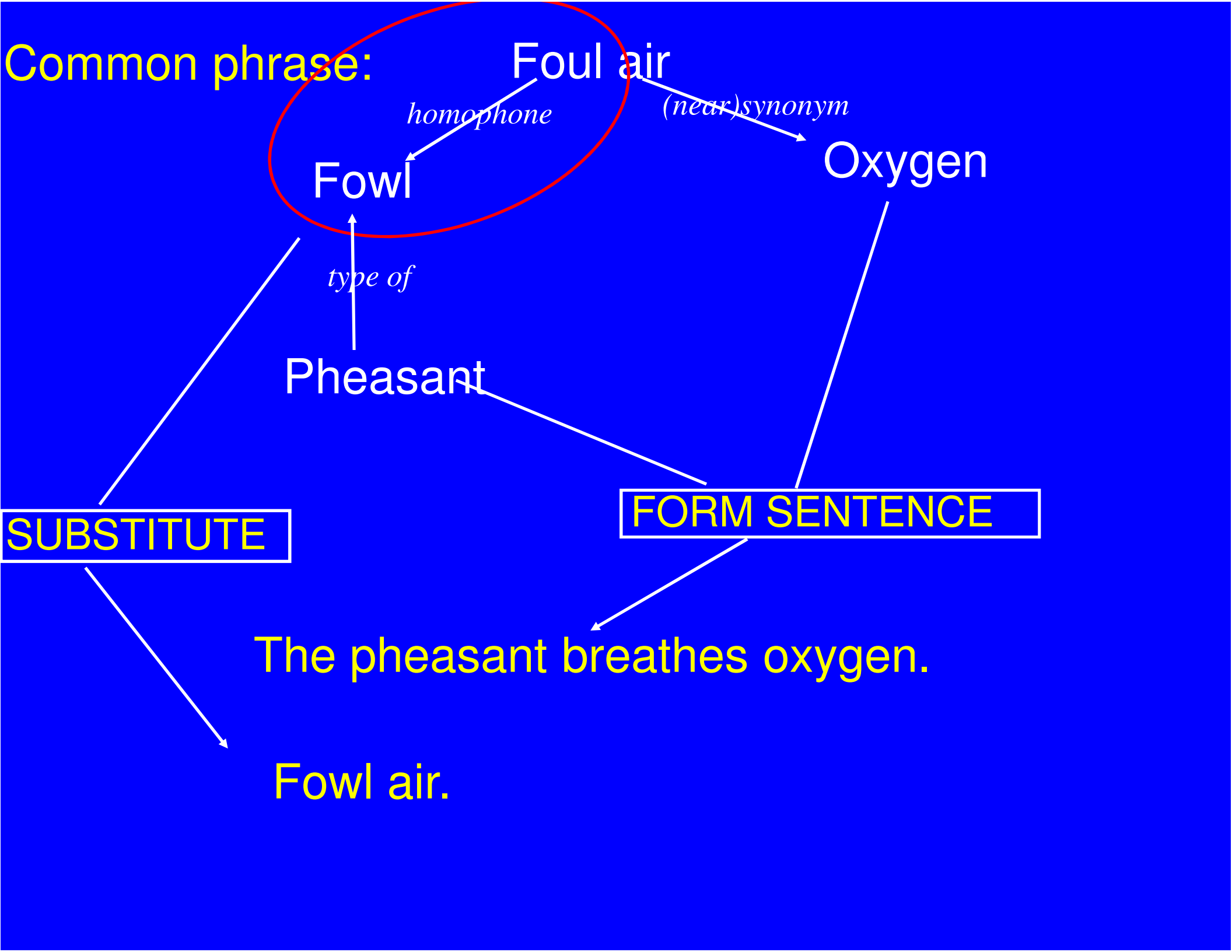
- very small number of subjects
- coverage very patchy/unbalanced
- (hence) no statistical significance

HCPP (Venour 1999)

Simple two-sentence puns based on associations between words.

The sailor bears a stress.

Pier pressure.


Common phrase:
Foul air
homophone
(near)synonym
Fowl
Oxygen
type of
Pheasant
SUBSTITUTE
FORM SENTENCE
The pheasant breathes oxygen.
Fowl air.

HCPP:

50 jokes, 15 judges.

Subjects asked to rate on a scale of:

1. Not a joke. Does not make sense.
2. Recognizably a joke but a pathetic one.
3. OK. A joke you might tell a child.
4. Quite good.
5. Really good.

Results:

Average score 2.81

44% of jokes scored 3, 4 or 5.

About one-third of the total votes were 4 or 5.

HCPP evaluation - some weaknesses

- small number of subjects
- no control items
- no statistical significance computed

JAPE (Binsted 1996)

Produced simple punning riddles.

Punning riddles

What kind of vegetable can jump? *A spring onion.*

What's the difference between leaves and a car?

One you brush and rake, the other you rush and brake.

What kind of tent has hair? *A wig-wam.*

What do you call a lizard on the wall? *A rep-tile.*

What do you get when you cross a monkey and a peach?

An ape-ricot.

What's the difference between money and a bottom?

One you spare and bank, the other you bare and spank.

What kind of leg can shoot? *A bow leg.*

What do you call a ghost summer race? A dead heat.

JAPE evaluation:

50 jokes, 120 judges.

Items mixed with human jokes and some non-jokes.

Subjects asked to:

- say whether a joke or not
- rate on a scale of:

1. Not funny at all.
2. Not very funny.
3. Not sure.
4. Funny.
5. Very funny.

Results:

Human jokes more like jokes, and funnier, than JAPE's.

JAPE's jokes more like jokes, and funnier, than the non-jokes.

Funniest joke was one of JAPE's.

JAPE evaluation - some weaknesses

- doubtful scale for "funniness“
- too many tests on same data?

Some methodological guidelines for joke generation (if aiming for a scientific model)

- Make it clear which aspects are humour-relevant.
- Any linguistic data used (e.g. dictionary) should not accidentally introduce humorous effects.
- Keep irrelevant details out of sight (e.g. text processing).
- Select input data systematically (e.g. randomly).
- Evaluate results properly.
- Select output for evaluation systematically.
- Design evaluation scales carefully.

What are the gains from this work?

- It makes a start on computer modelling of humour.
- It formalises some (very simple) aspects of jokes.
- It clears the ground for further investigation.
- It allows real testing of these (limited) proposals.

Main obstacles

(to building working systems)

Lack of formal and detailed theories of humour.

The need for world-knowledge and powerful inference ("humour-generation is AI-complete").

But...

computational humour could be fruitful, both practically and scientifically.

The Linguistic Analysis of Jokes

Graeme Ritchie (2003)

Routledge, London/New York