ESSLLI 2002 Workshop on

Machine Learning Approaches in Computational Linguistics

August 5-9 2002

Trento, Italy

# Learning Lexical Inheritance Hierarchies with Maximum Entropy Models


**Caroline Sporleder**

Division of Informatics

University of Edinburgh, Scotland

# Overview

1. Lexical Inheritance Hierarchies
2. The Task
3. Minimal Redundancy vs. Linguistic Plausibility
4. Galois Lattices & Maximum Entropy Pruning
5. Experiments
6. Conclusion

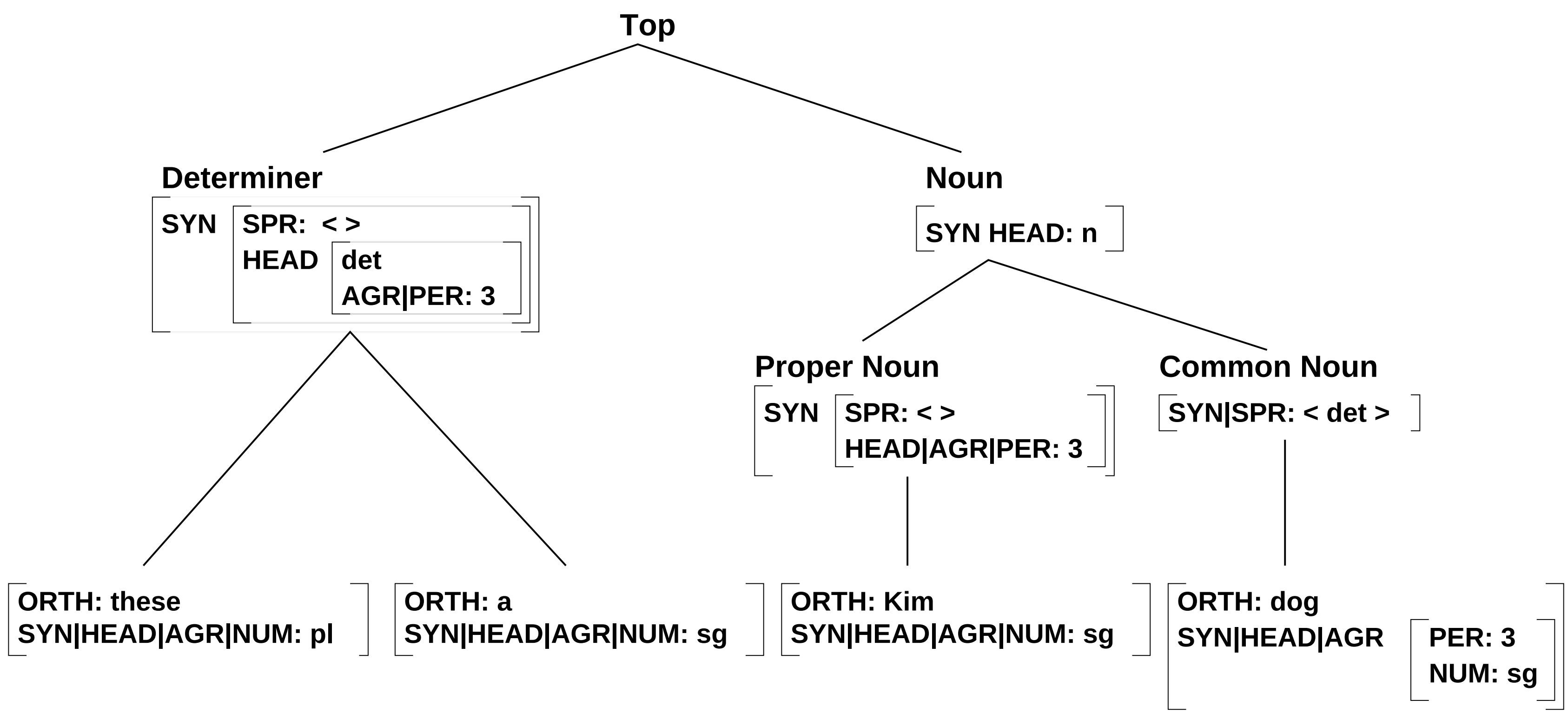
# Lexical Inheritance Hierarchies

- hierarchical representation of lexical knowledge
- capture generalisations
- reduce redundancy

$\Rightarrow$ popular in many modern (esp. lexicalist) grammar theories

# Lexical Inheritance Hierarchies

Example:



# The Task

Constructing a “good” lexical inheritance hierarchy for a flat lexicon

| ORTH | SYN$\vert$SPR | SYN$\vert$HEAD | SYN$\vert$HEAD$\vert$AGR$\vert$PER | SYN$\vert$HEAD$\vert$AGR$\vert$NUM |
| --- | --- | --- | --- | --- |
| these | $\langle\ \rangle$ | det | 3 | pl |
| a | $\langle\ \rangle$ | det | 3 | sg |
| Kim | $\langle$det$\rangle$ | noun | 3 | sg |
| Mary | $\langle$det$\rangle$ | noun | 3 | sg |
| dog | $\langle$det$\rangle$ | noun | 3 | sg |
| cats | $\langle$det$\rangle$ | noun | 3 | pl |

# Minimal Redundancy vs. Linguistic Plausibility

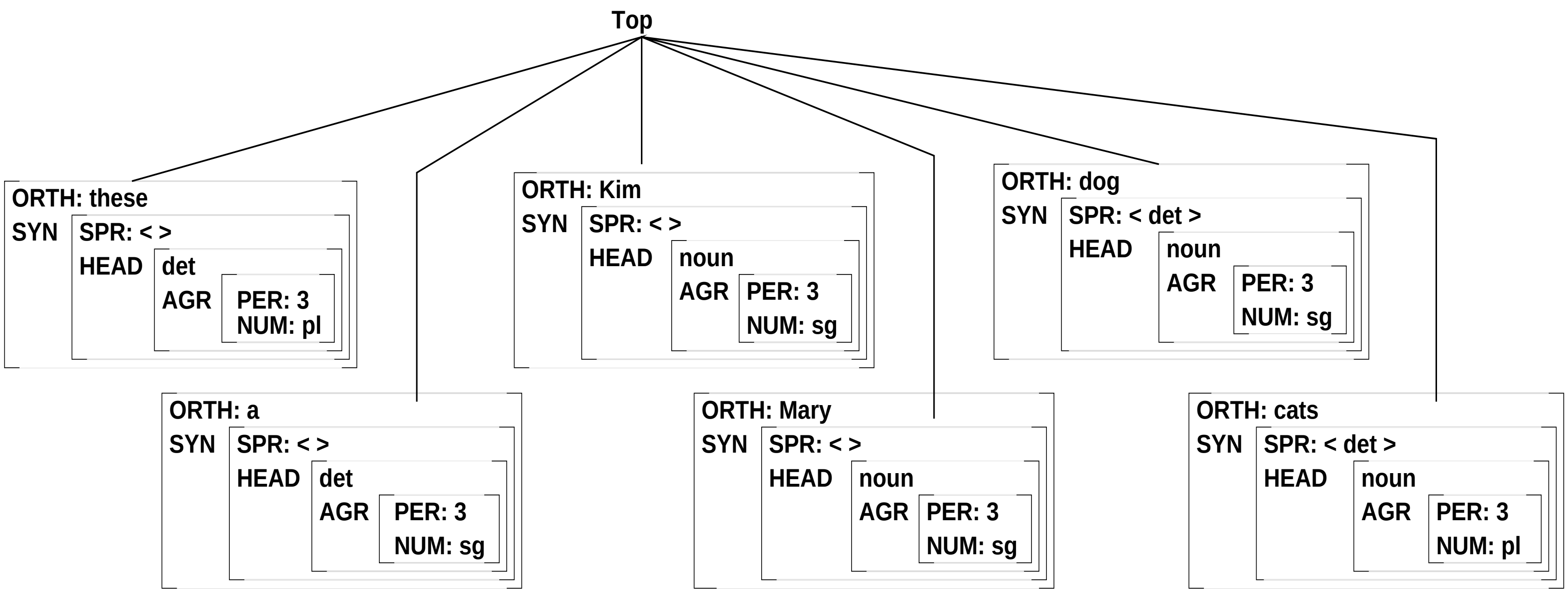
Previous approaches: aim for minimal redundancy

How is redundancy defined?

- number of nodes?
- number of attribute-value pairs?
- number of inheritance links?
- some combination of the above?
- something else?

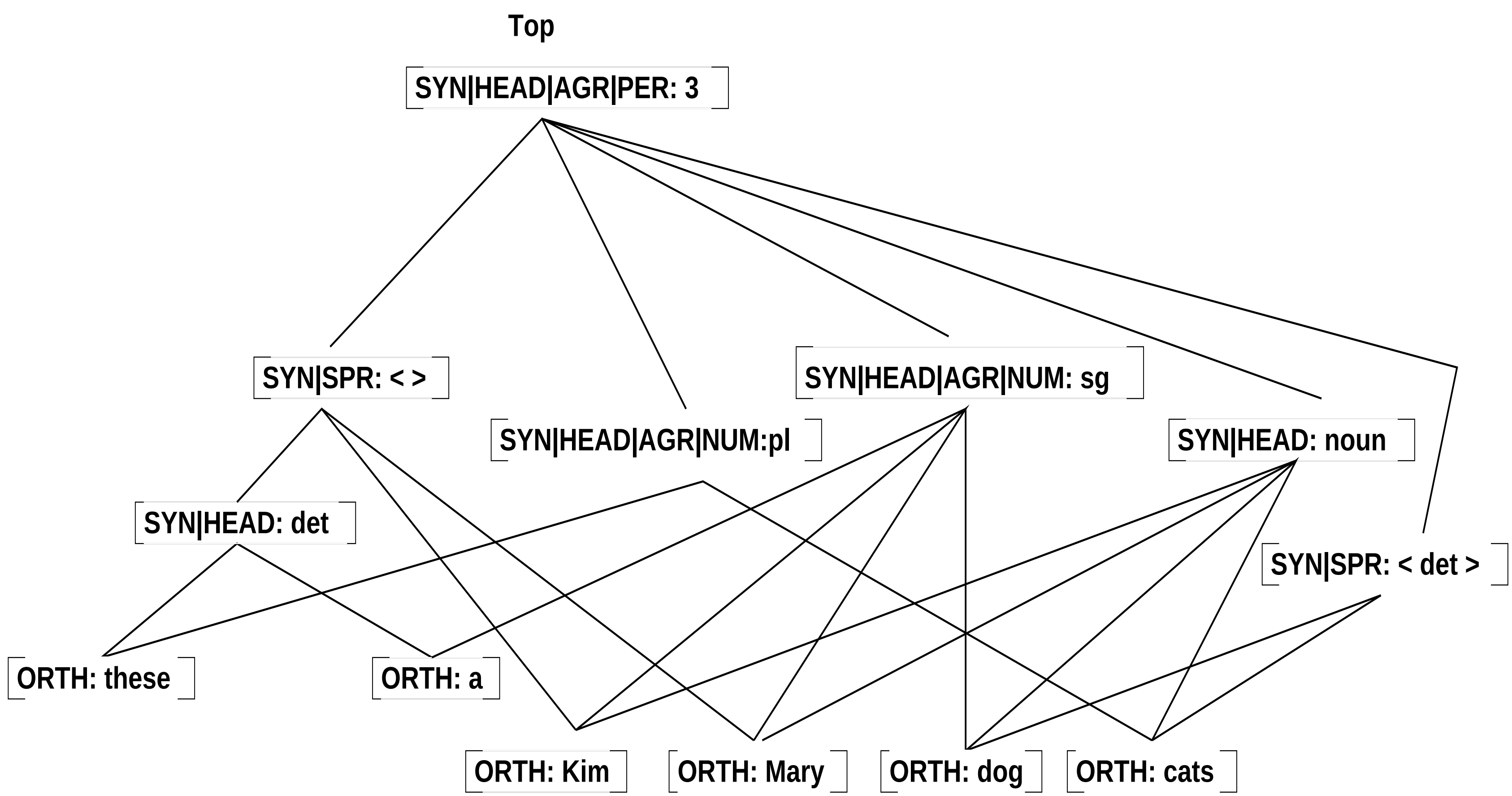
# Minimal Redundancy vs. Linguistic Plausibility

Minimising nodes:



# Minimal Redundancy vs. Linguistic Plausibility

Minimising attribute-value pairs:



# Minimal Redundancy vs. Linguistic Plausibility

What about a combination of simple redundancy criteria?

For example:

redundancy $\stackrel{\text{def}}{=}$ # nodes + # attribute-values pairs

| | nodes | attribute-value pairs | sum |
|---|---|---|---|
| plausible hierarchy | 11 | 19 | 30 |
| min. nodes | 6 | 30 | 36 |
| min. AVPs | 13 | 13 | 26 |

# Minimal Redundancy vs. Linguistic Plausibility

Minimal redundancy criteria:

- conflict with each other
- don't lead to linguistically plausible hierarchies
- simple combination doesn't help either

Better:

focus on linguistic plausibility

$\Rightarrow$ plausibility of a hierarchy fragment depends on its context, e.g.:

- its surrounding nodes etc.
- interdependencies in the data

# Galois Lattices & Maximum Entropy Pruning

1. find all generalisation contained in the lexicon

   $\Rightarrow$ non-empty intersections between lexical entries

   $\Rightarrow$ build Galois lattice for the lexicon

2. decide which generalisations are "good" (linguistically plausible)

   $\Rightarrow$ classification problem

   $\Rightarrow$ supervised learning using maximum entropy models

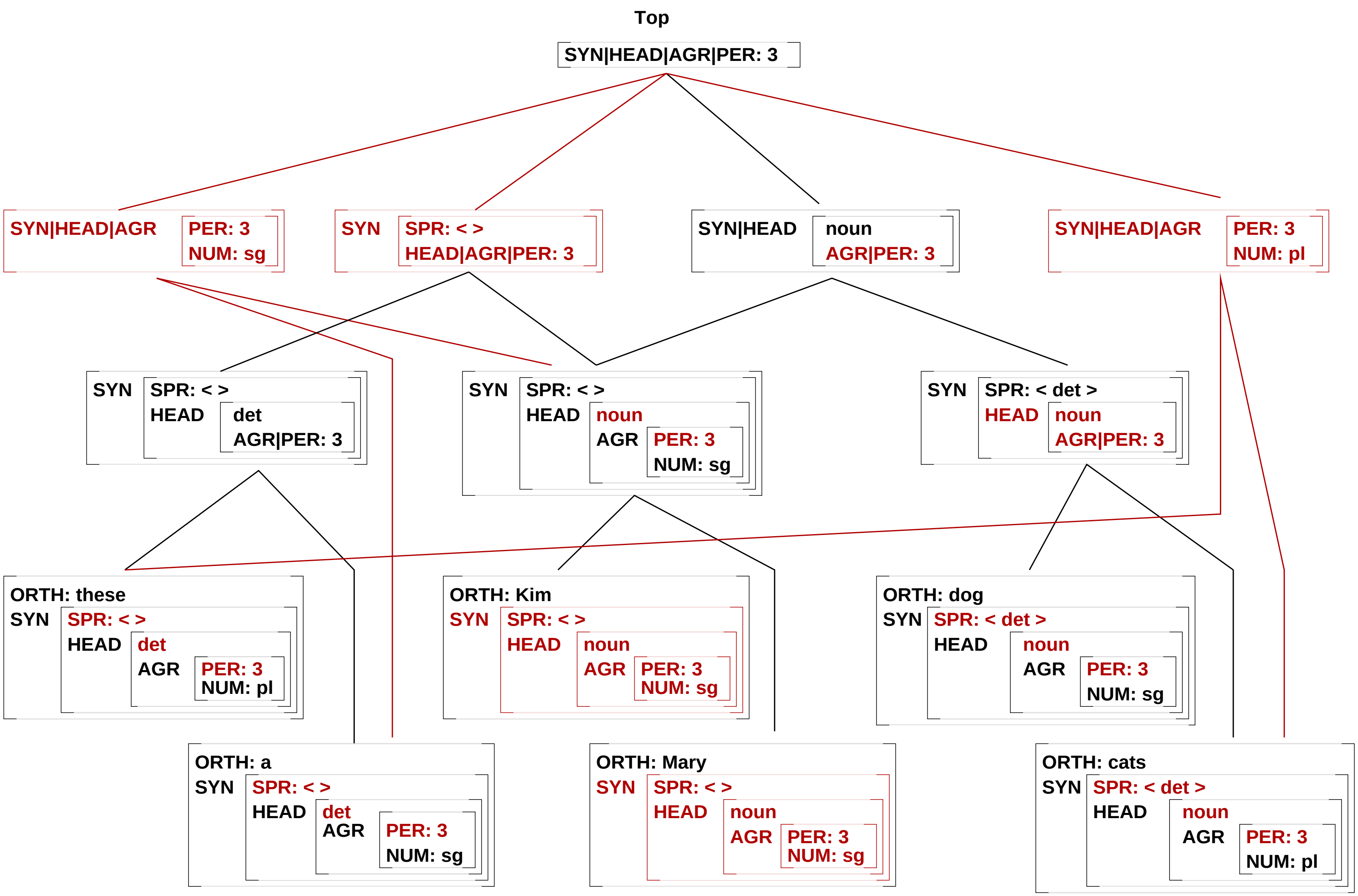
   $\Rightarrow$ fine-grained context-dependent modelling of linguistic plausibility

3. prune bad generalisations (& redundant avps)

   $\Rightarrow$ lexical inheritance hierarchy

# Galois Lattices & Maximum Entropy Pruning

Galois lattice:



# Galois Lattices & Maximum Entropy Pruning

Statistical Modelling:

so far 14 contextual feature sets, relating to:

- immediate ancestors & descendants
- terminal descendants
- attribute-value pairs
- level in hierarchy

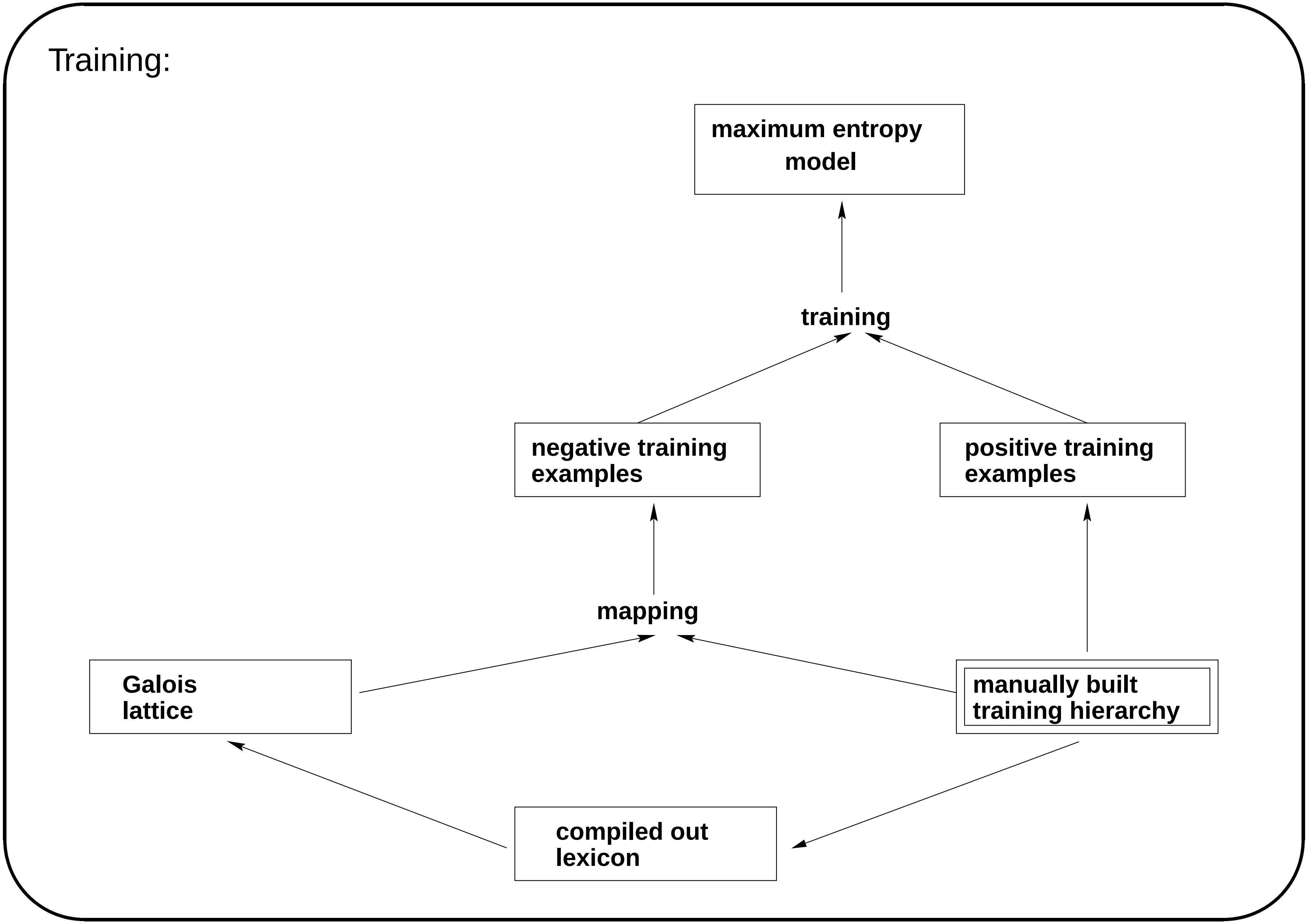
Statistical Modelling (cont'd):

Interdependencies between attributes (not yet implemented):

- Value Dependence: each value of attribute $a1$ implies a particular value of attribute $a2$
- Value-Set Dependence: each value of attribute $a1$ implies that the value of attribute $a2$ is taken from a particular subset of values
- Appropriateness Dependence: the attribute-value pair $a1 : v1$ implies the appropriateness of the attribute $a2$

$\Rightarrow$ interdependencies between avps of a node should increase $P(retain)$

# Galois Lattices & Maximum Entropy Pruning



# Galois Lattices & Maximum Entropy Pruning

Evaluation:

- partial matching of automatically constructed hierarchy to original hierarchy
- calculation of precision and recall based on the proportion of matched nodes/attribute-value pairs

! assumes one “ideal” hierarchy (i.e. the original one)

# Experiments

How good is a basic maximum entropy model compared to other pruning methods?

Data (manually built hierarchies):

- English (Sag & Wasow 1999), 501 entries
- Spanish (Quirino Simões 2001), 405 entries

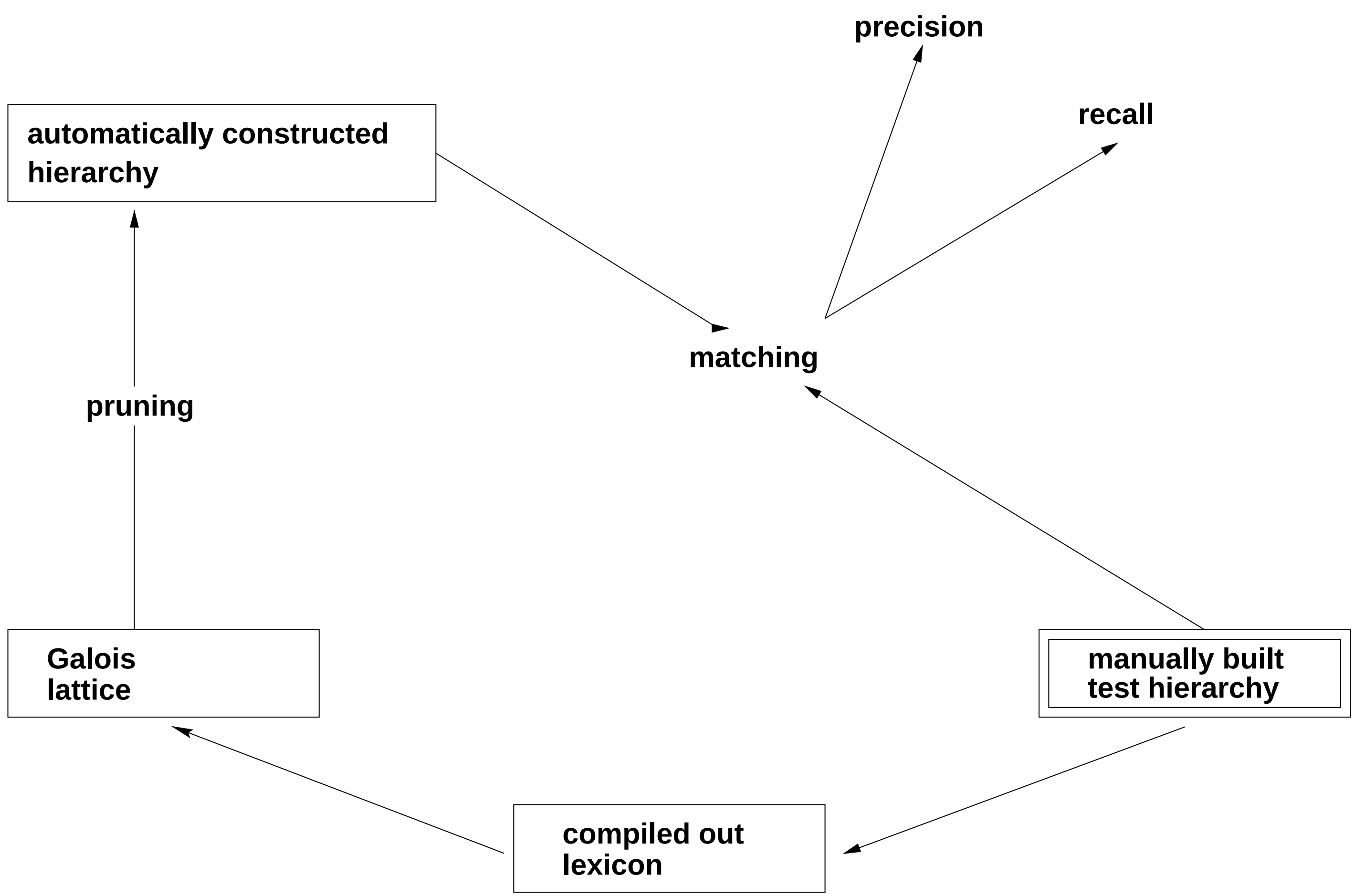
# Experiments

4 pruning methods:

- Maximum Entropy: trained on Spanish and applied to English & *vice versa*
- Minimal AVPs: every attribute-value pair occurs exactly once in the pruned lattice (cf. Petersen 2001)
- Random, uniform: nodes are pruned randomly, *P(prune)=0.5*
- Random, n-best: randomly keep *n* nodes, where *n* is the number of nodes in the original hierarchy

# Experiments

Set-Up for Testing:



# Experiments

Results (English):

- low f-score for all pruning methods ($\rightarrow$ task is hard)
- MaxEnt & minimal pruning better than random pruning

| ENGLISH | f-score | precision | recall | retained nodes |
|---|---|---|---|---|
| MaxEnt | 22.16% | 18.59% | 27.44% | 51 |
| minimal | 22.14% | 15.79% | 37.05% | 238 |
| rnd, uni | 18.37% | 12.21% | 37.19% | 287 |
| rnd, n-best | 21.93% | 23.65% | 20.65% | 43 |

# Experiments

Results (Spanish):

- English lexicon doesn't provide enough training data for MaxEnt

  $\rightarrow$ bad results for training on English & testing on Spanish

| SPANISH | f-score | precision | recall | retained nodes |
|---|---|---|---|---|
| MaxEnt | 0.29% | 0.62% | 0.19% | 25 |
| minimal | 23.99% | 19.84% | 30.32% | 330 |
| rnd, uni | 16.90% | 12.01% | 28.59% | 556 |
| rnd, n-best | 9.54% | 11.36% | 8.28% | 100 |

# Conclusion

Summary:

- linguistic plausibility not minimal redundancy
- simple minimal redundancy criteria are not enough
- statistical modelling based on many contextual properties
- simple MaxEnt model beats baseline

Improvements:

- more training data
- better modelling of context

## Acknowledgements

This research was funded by a University of Edinburgh Faculty of Science and Engineering Scholarship.

I am also grateful for a Division of Informatics Graduate School Travel Grant.