Integrating a Phrase Structure Corpus Grammar and a Lexical-Semantic Network: the HOLINET Knowledge Graph

LREC 2022, Marseille, France

Jean-Philippe Prost

Laboratoire Parole et Langage (LPL), CNRS – Aix-Marseille Université, France
Jean-Philippe.Prost@univ-amu.fr

Motivation

Based on the fact that...

Knowledge Graphs (KGs) have become a main source of knowledge for many NLP applications, whether based on symbolic or sub-symbolic approaches,

and given that...

KGs can be found that integrate ontological, morphological, lexical, semantic knowledge, *but rarely, if ever, grammar knowledge*

Question

Would the integration of, ultimately, all the possible linguistic dimensions contribute to improve the KG-based NLP applications?

Motivation

Possible applications

What for?

E.g.,

- Graph embeddings
- Knowledge graph reasoning

Motivation

Focus

Integration of a phrase structure grammar and lexical-semantic knowledge

Contributions

- a property graph model for a phrase structure grammar, to be integrated in a multi-layered Knowledge Graph
- a procedure for the creation and integration of such a grammar layer with a lexical-semantic network,
- an implementation of the model and procedure with the French Treebank (FTB) and the JeuxDeMots lexical-semantic network (JDM)

Outcome

- The HOLINET Knowledge Graph, available in the LRE map
- The software involved

Outline

The (preliminary) Alpha model

Example: PTB-annotated phrase from the FTB

```
(NP
 (DET##lem=un|cpos=D|g=f|n=s|s=ind## une)
 (NC##lem=bataille|cpos=N|g=f|n=s|s=c## bataille)
 (AP
  (ADJ##lem=politique|cpos=A|n=s|s=qual## politique))
 (AP
  (ADV##lem=extrêmement|cpos=ADV|_## extrêmement)
  (ADJ##lem=ardu|cpos=A|g=f|n=s|s=qual## ardue)))
```

an extremely arduous political struggle

The Alpha model

Example: PTB-annotated phrase from the FTB

The phrase structure can be represented as a set of rewrite rules:

(1) NP $\longrightarrow$ DET#fsi NC#fsc AP AP

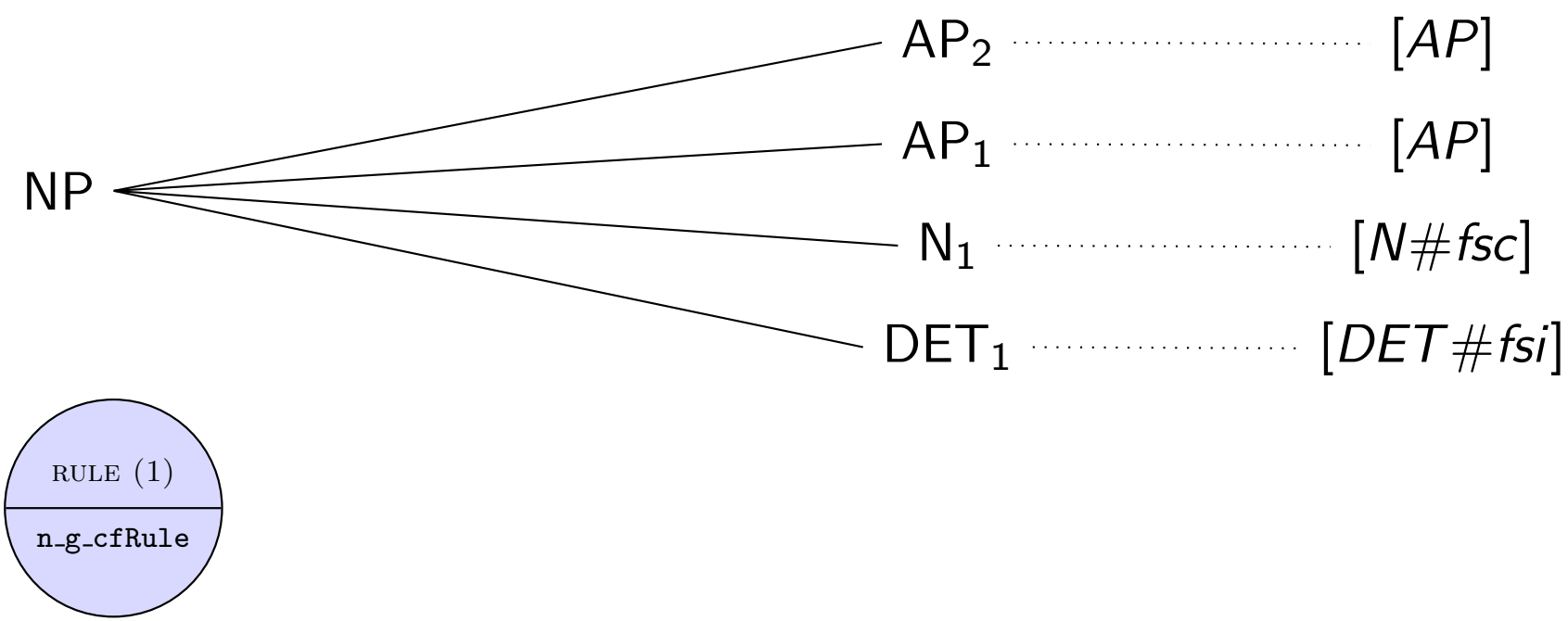
$$
\textit{rule1}\left[\begin{array}{l}\textsc{type}\ \ \texttt{n\_g\_cfRule}\\ \textsc{rule}\left[\begin{array}{l}\textsc{lhs}\ \left[\textsc{cat}\ \ \textit{NP}\right]\\ \textsc{rhs}\left\langle \underset{\textit{DET\#fsi}}{\left[\begin{array}{ll}\textsc{cat} & \textit{DET}\\ \textsc{cpos} & \textit{D}\\ \textsc{gend} & \textit{f}\\ \textsc{num} & \textit{s}\\ \textsc{subcat} & \textit{ind}\end{array}\right]}, \underset{\textit{N\#fsc}}{\left[\begin{array}{ll}\textsc{cat} & \textit{NC}\\ \textsc{cpos} & \textit{N}\\ \textsc{gend} & \textit{f}\\ \textsc{num} & \textit{s}\\ \textsc{subcat} & \textit{c}\end{array}\right]}, \left[\textsc{cat}\ \ \textit{AP}\right], \left[\textsc{cat}\ \ \textit{AP}\right]\right\rangle\end{array}\right]\end{array}\right]
$$

(2) AP $\longrightarrow$ ADJ#sq

(3) AP $\longrightarrow$ ADV ADJ#fsq

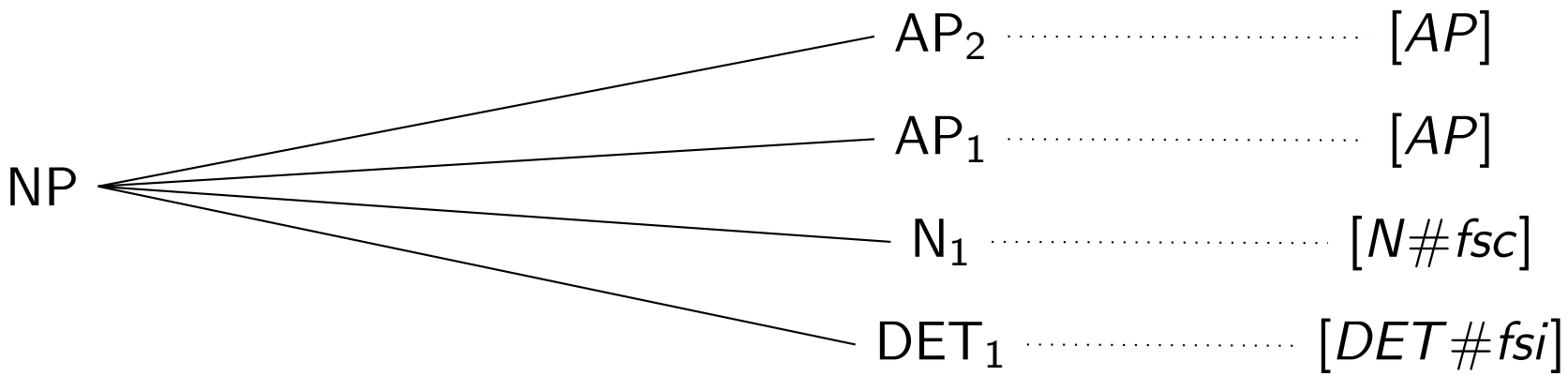
The Alpha model

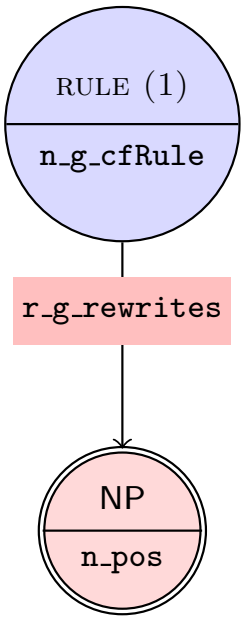
Rule (1) modelled as a *property graph*



The Alpha model

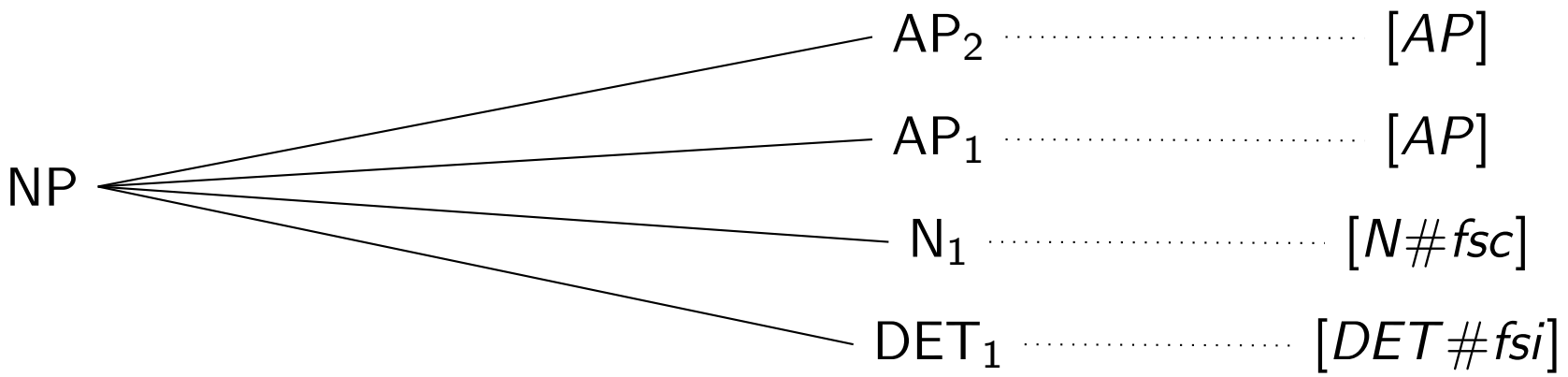
Rule (1) modelled as a *property graph*

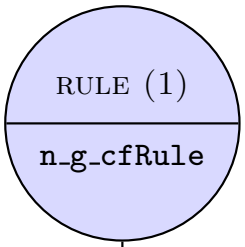


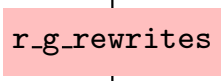


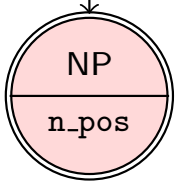
The Alpha model

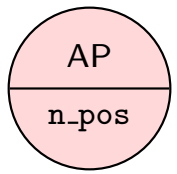
Rule (1) modelled as a *property graph*

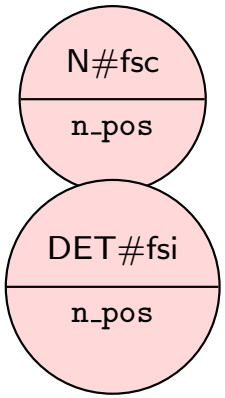






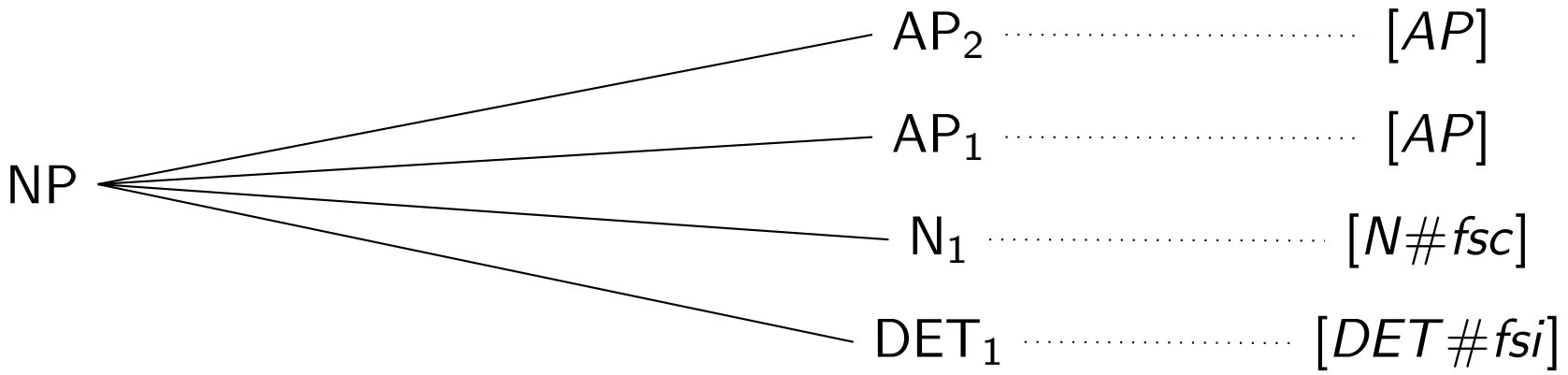


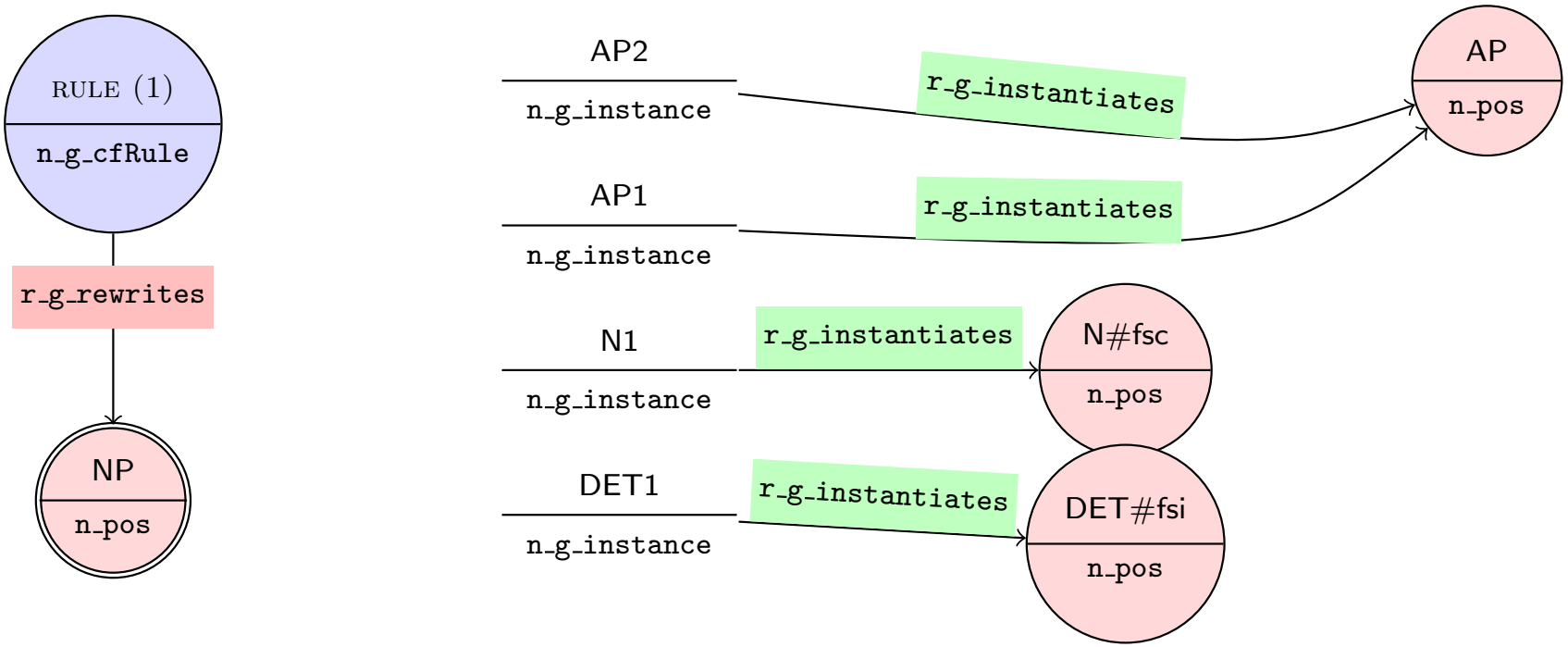




The Alpha model

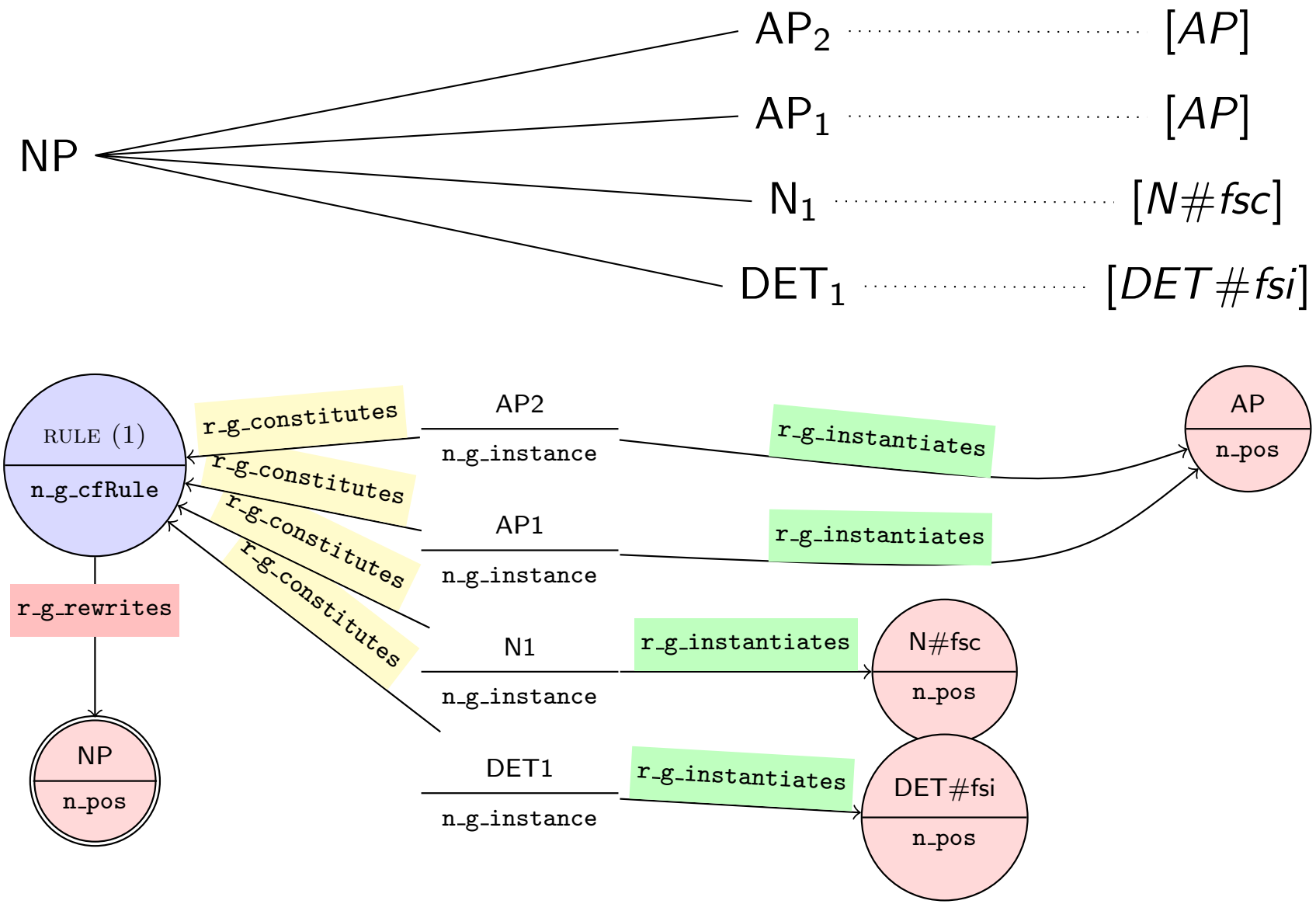
Rule (1) modelled as a *property graph*



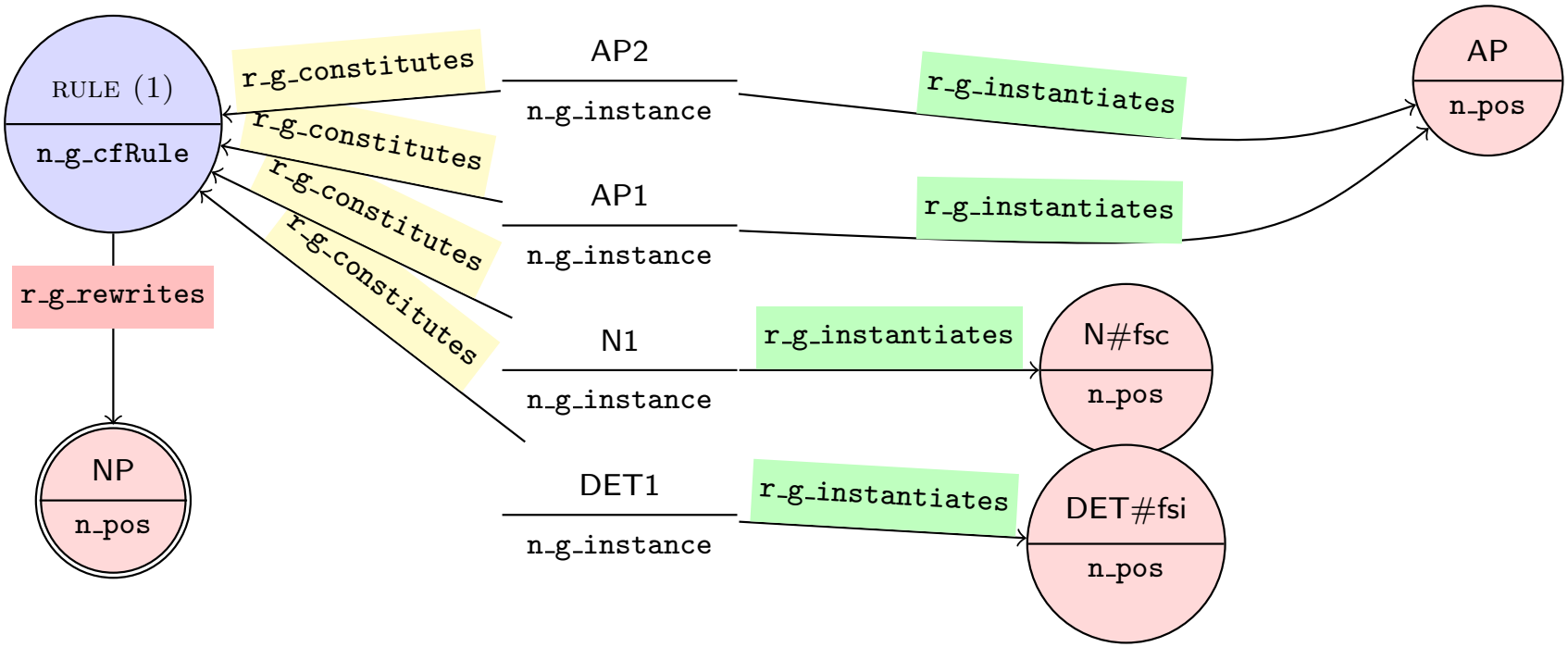


The Alpha model

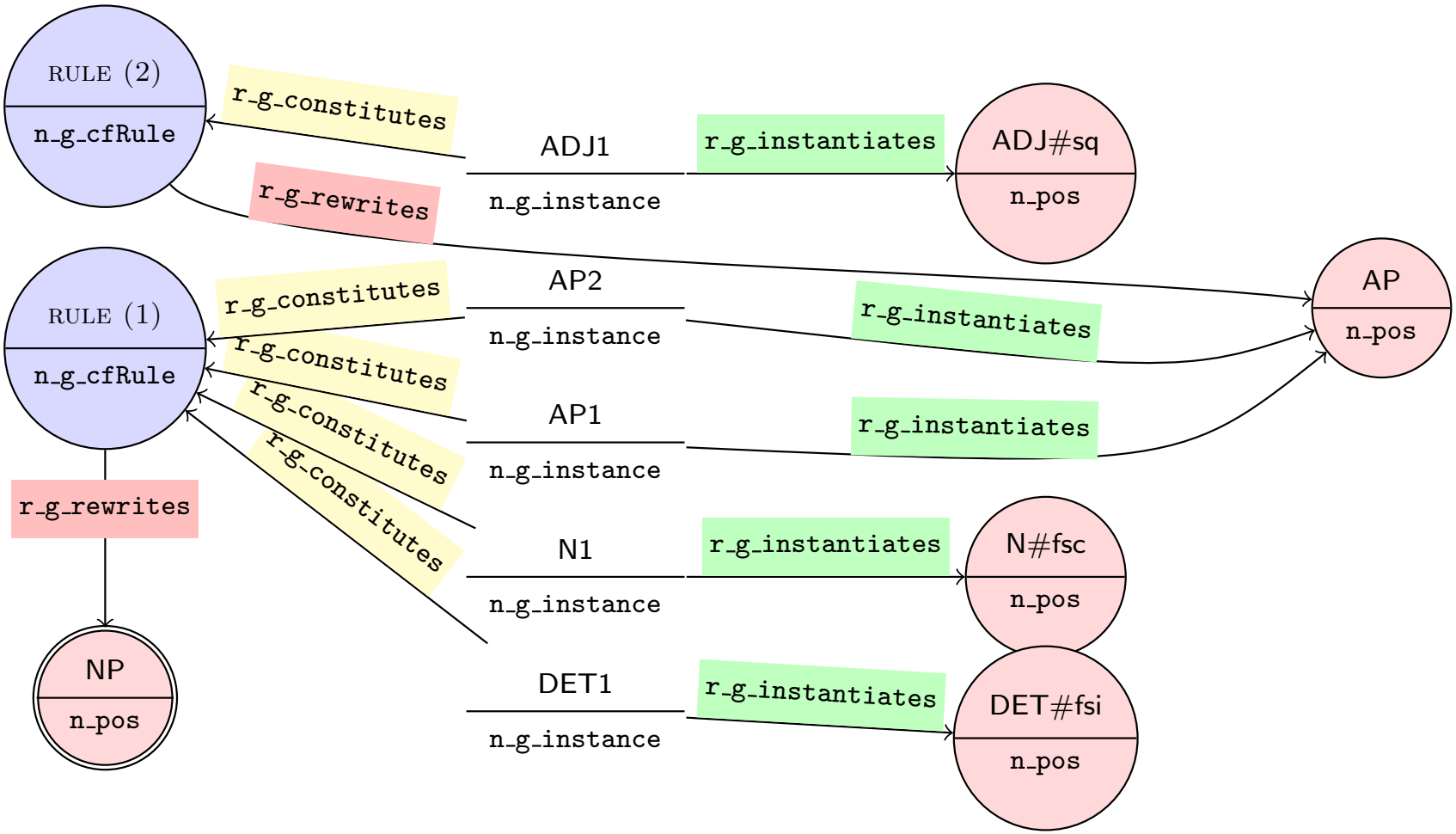
Rule (1) modelled as a *property graph*



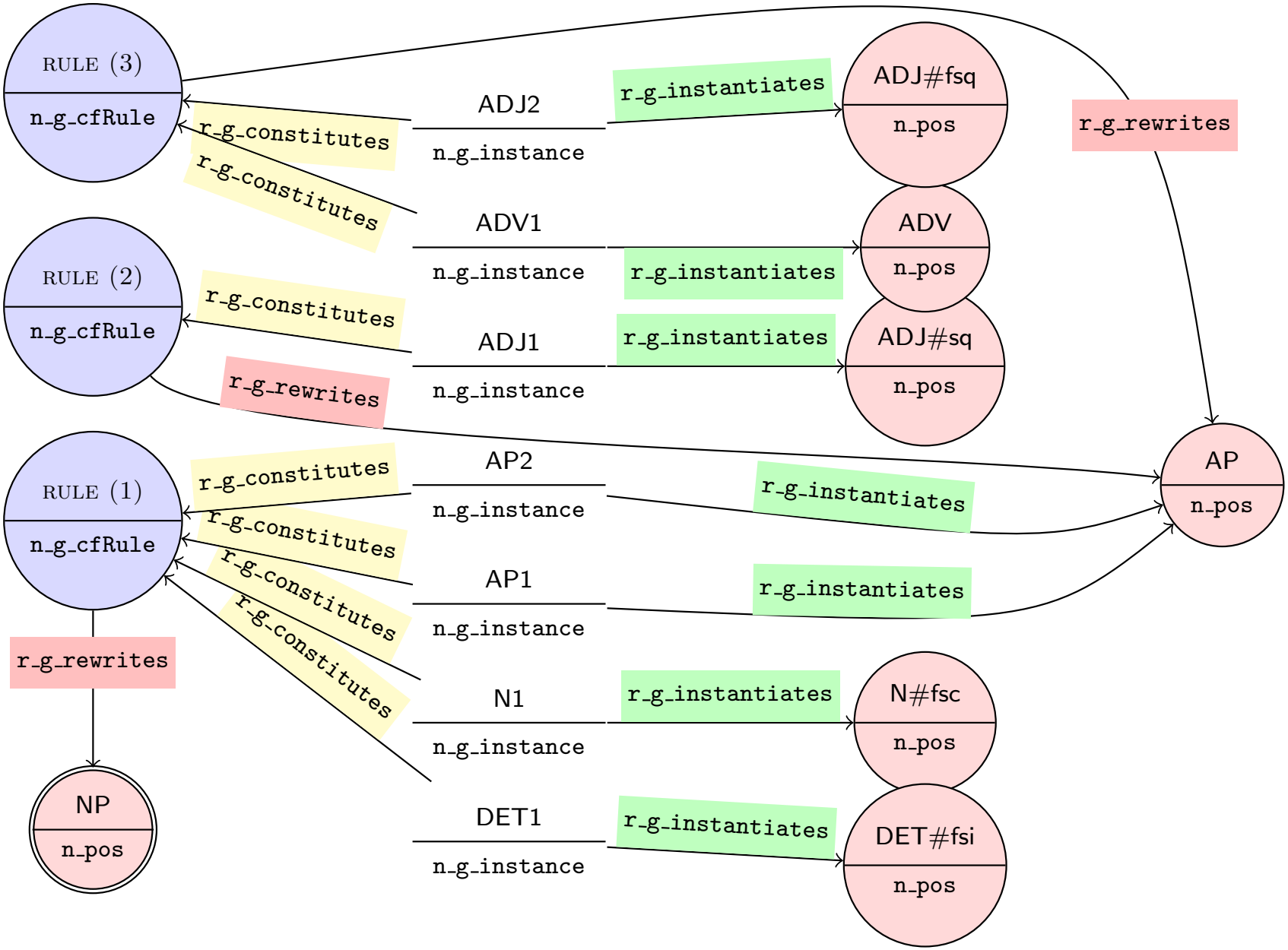
The Alpha model



The Alpha model



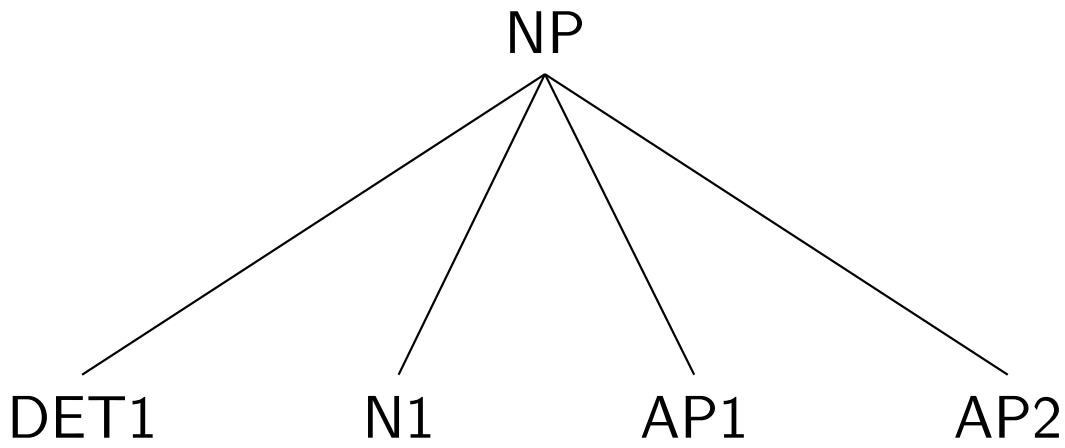
The Alpha model



Shortcomings of the Alpha model

Rule 1

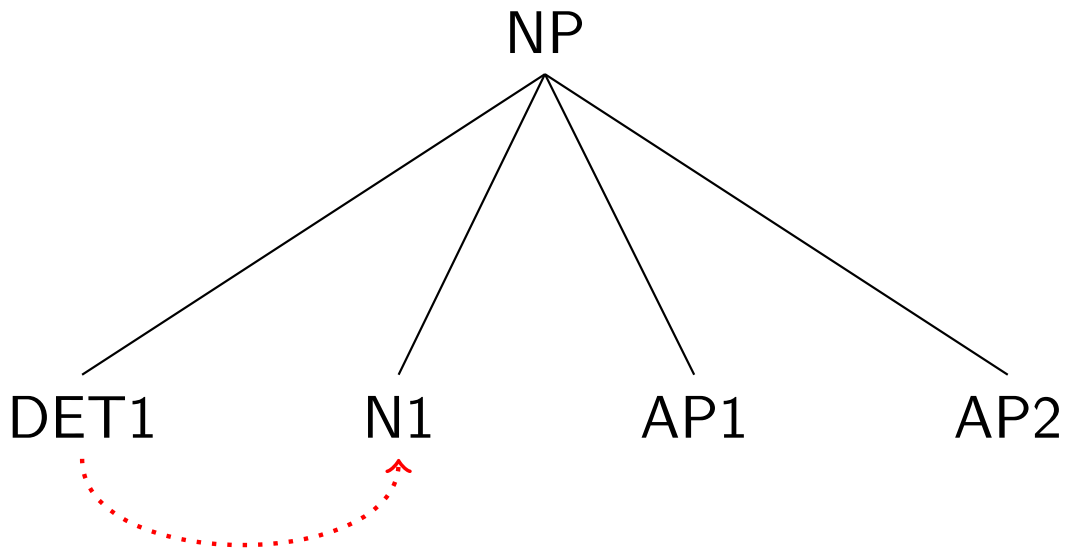
The Alpha model accounts for Immediate Dominance,



Shortcomings of the Alpha model

Rule 1

The Alpha model accounts for Immediate Dominance,
but it fails to account for other implicit relationships, such as:

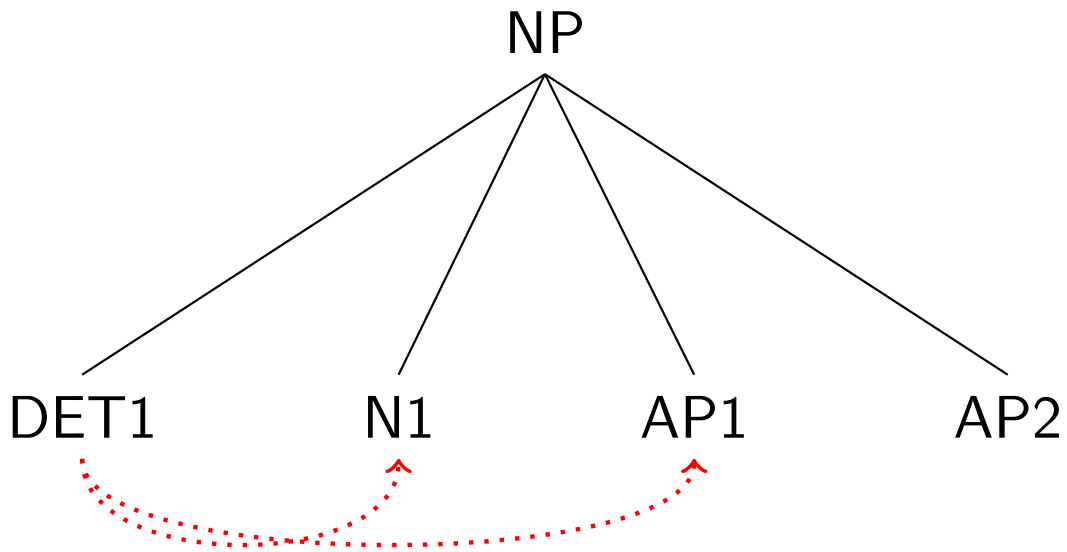


- Linear Precedence (local to a rule)

Shortcomings of the Alpha model

Rule 1

The Alpha model accounts for Immediate Dominance,
but it fails to account for other implicit relationships, such as:

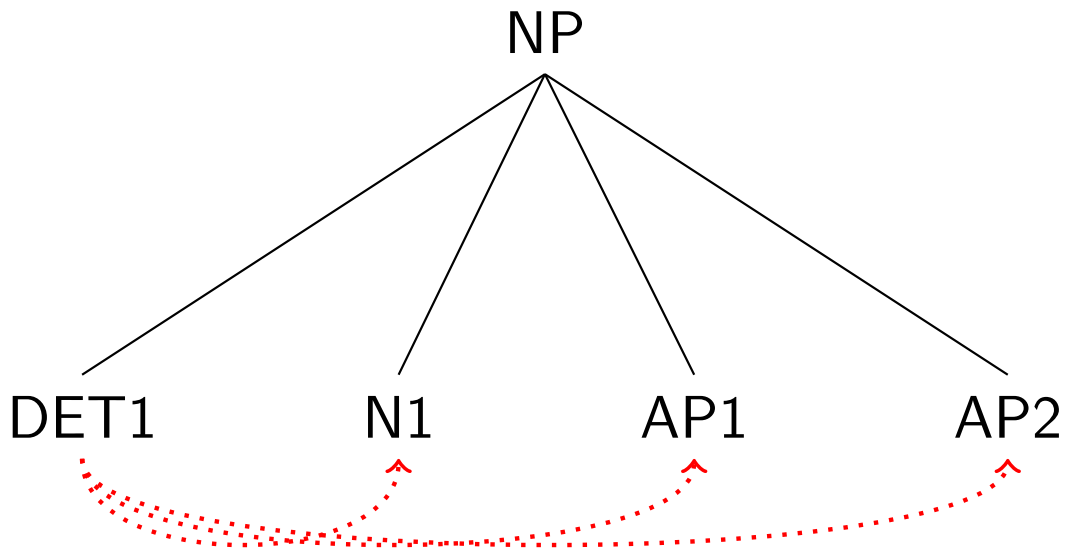


- Linear Precedence (local to a rule)

Shortcomings of the Alpha model

Rule 1

The Alpha model accounts for Immediate Dominance,
but it fails to account for other implicit relationships, such as:

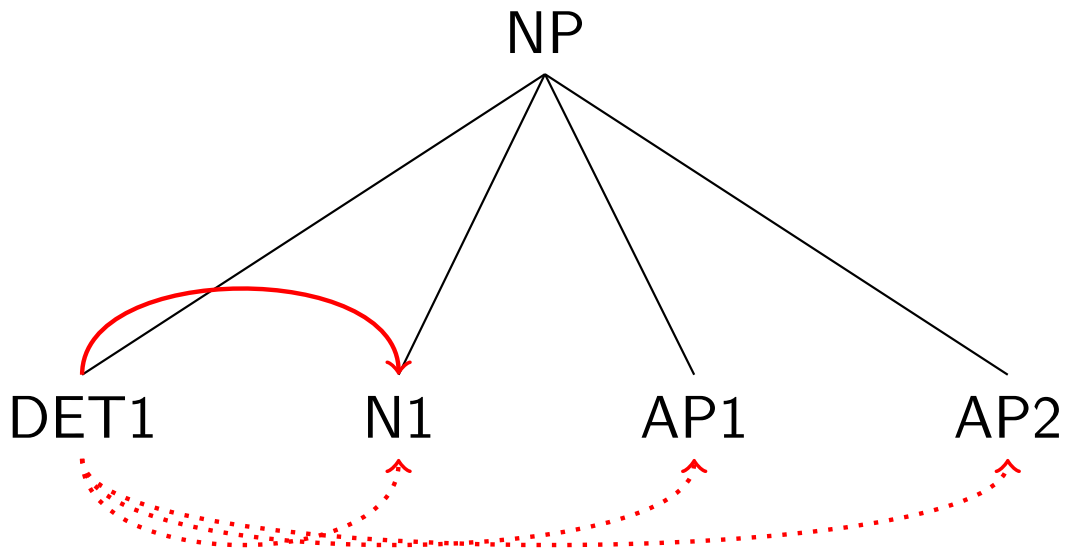


- Linear Precedence (local to a rule)

Shortcomings of the Alpha model

Rule 1

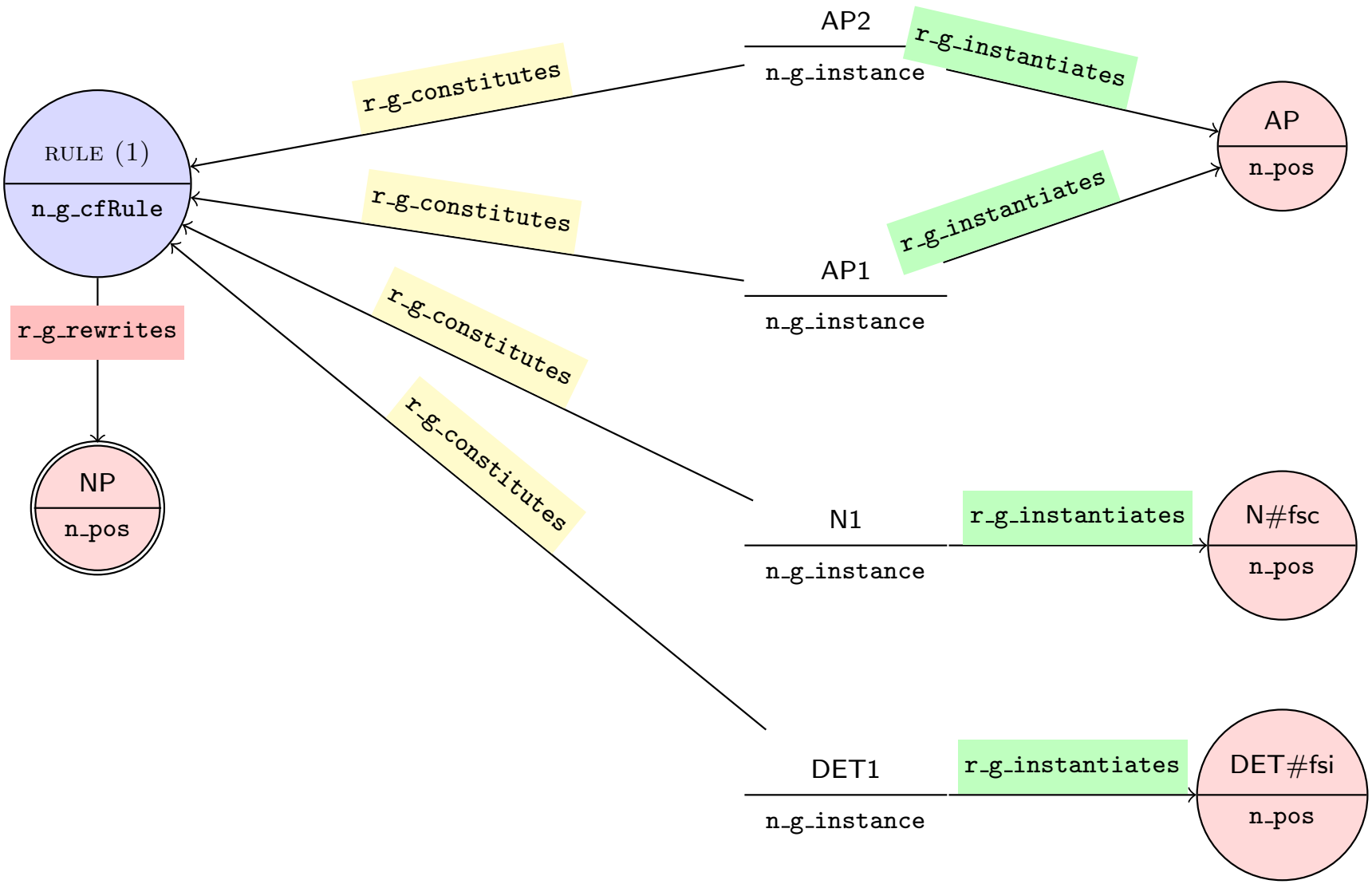
The Alpha model accounts for Immediate Dominance,
but it fails to account for other implicit relationships, such as:



- Linear Precedence (local to a rule)
- (oriented) co-occurrence (global to the CFG)

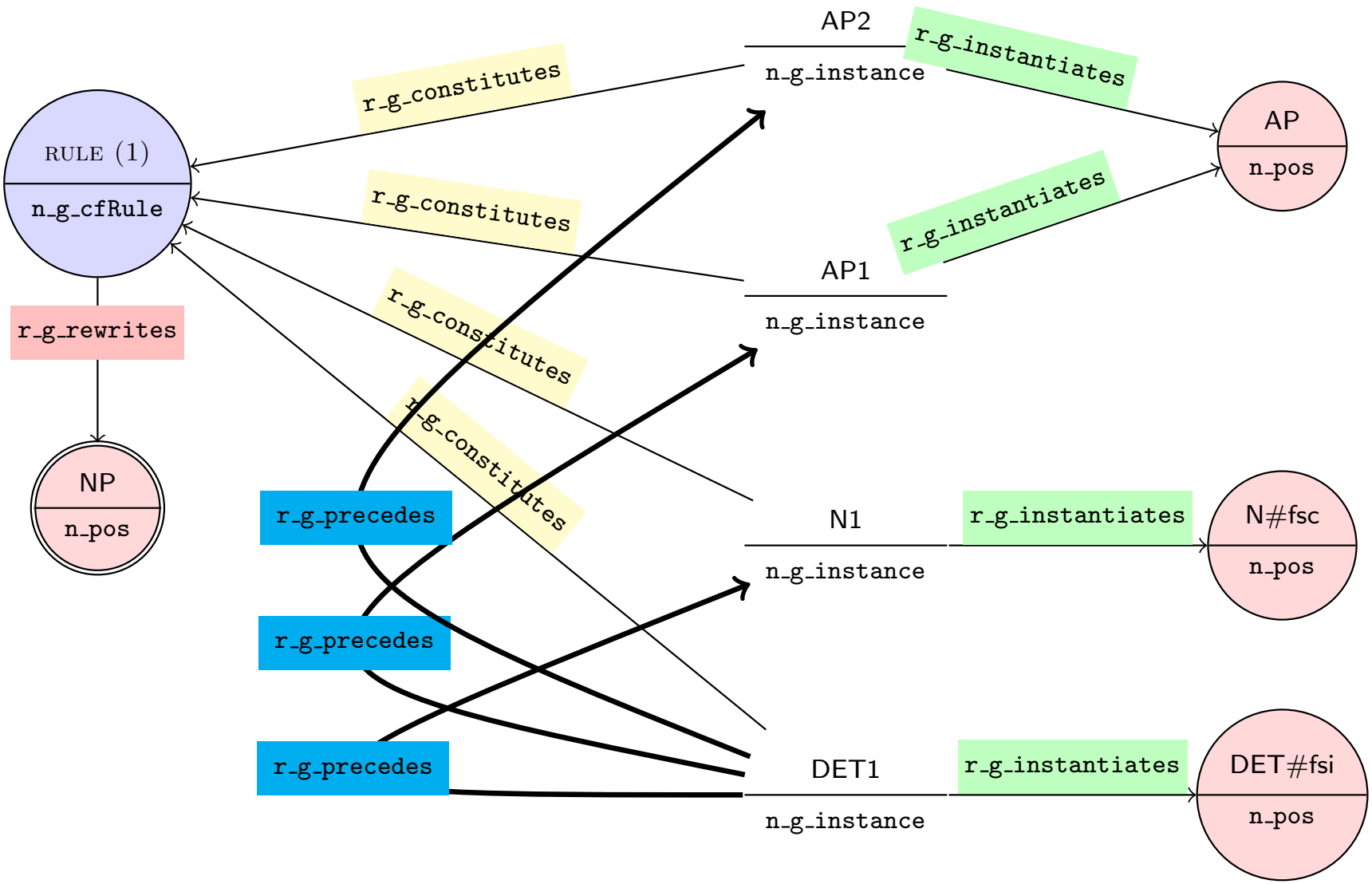
The Beta (and final) model

Making implicit relationships explicit



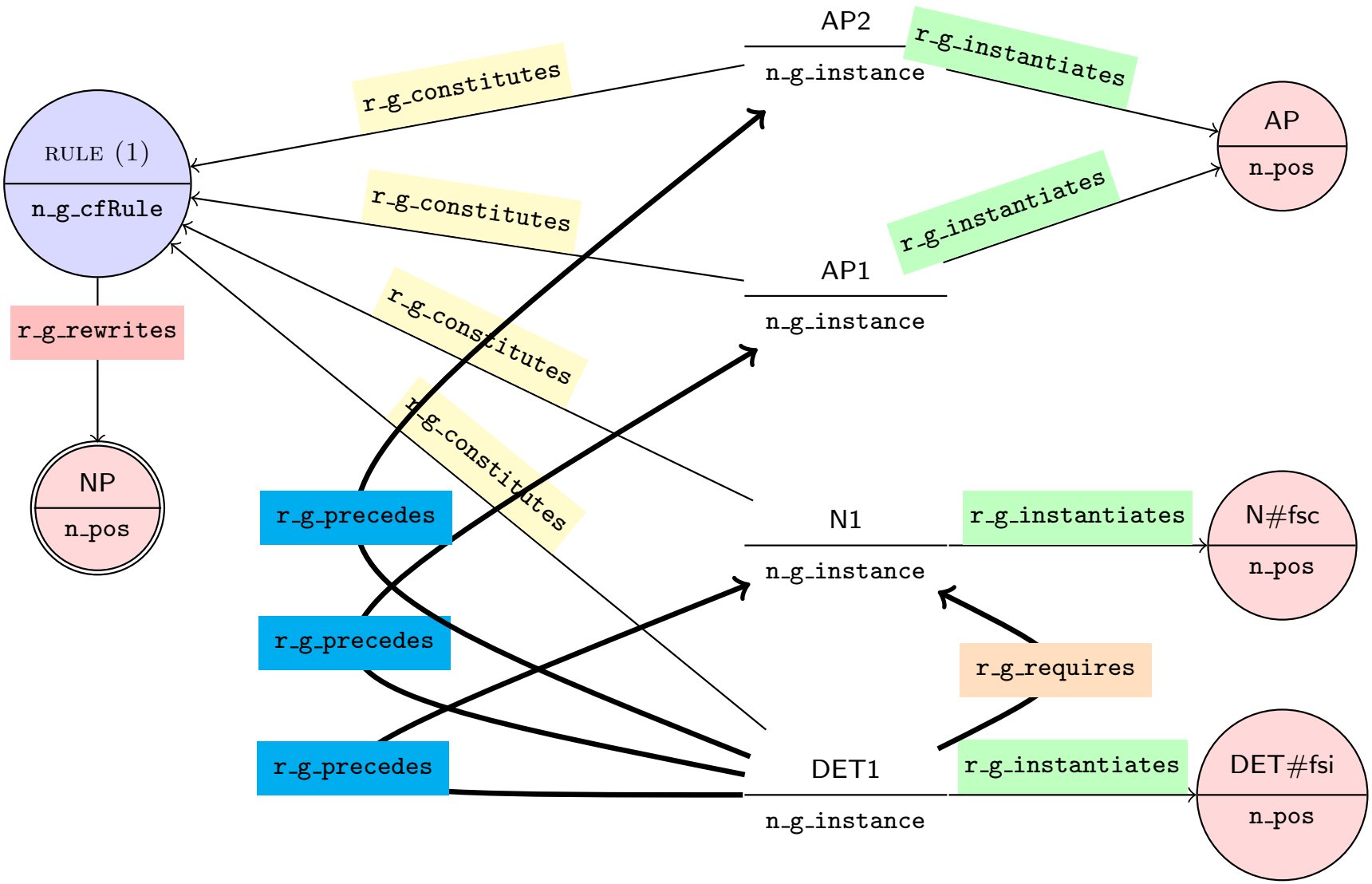
The Beta (and final) model

Making implicit relationships explicit



The Beta (and final) model

Making implicit relationships explicit



The Beta model

How to define the grammar relationships' semantics?

We borrow the semantics defined in Duchier et al. (2009) for the following Property Grammar (PG) (Blache, 2001) relationship types:

- Constituency
- Linear Precedence
- Requirement

The creation and integration procedure

Pre-requisites

We assume:

- a treebank annotated with phrase structures
- a lexical-semantic network, where
  - lexical entries
  - POS categories
  - `r_pos` relationships between the lexical entries and their POS nodes
- if necessary, a conversion table: treebank POS labels $\leftrightarrow$ network POS labels

The creation and integration procedure

Implementation

We experiment with:

- The French Treebank (FTB) (Abeillé et al., 2003)
- JeuxDeMots (JDM) (Lafourcade, 2007)

The creation and integration procedure

In short

1. *extract* the corpus CFG from the treebank
2. *derive* the PG properties from the corpus CFG
3. *assess* the PG properties
4. *convert* the corpus tagset as required
5. *create* the sets of nodes and edges for the grammar layer
6. *merge* the grammar and the lexical layers

The creation and integration procedure

In short

1. *extract* the corpus CFG from the treebank
2. *derive* the PG properties from the corpus CFG
3. *assess* the PG properties
4. *convert* the corpus tagset as required
5. *create* the sets of nodes and edges for the grammar layer
6. *merge* the grammar and the lexical layers

The creation and integration procedure

In short

1. *extract* the corpus CFG from the treebank
2. *derive* the PG properties from the corpus CFG
3. *assess* the PG properties
4. *convert* the corpus tagset as required
5. *create* the sets of nodes and edges for the grammar layer
6. *merge* the grammar and the lexical layers

The creation and integration procedure

Step 2: derive PG properties from CFG

- The automated derivation relies on previous work (Prost, 2016)
- Each relationship type corresponds to a derivation rule

  Constituency **if** a rule can be found in the CFG where A rewrites B, **then** B is a possible constituent of A

  Linear Precedence **if** a rule can be found where A rewrites B and C with $B \prec C$, and no rule can be found where A rewrites $C \prec B$ **then** $(A : B \prec C)$

  Requirement **if** no rule can be found where A rewrites B without C **then** $(A : B \Rightarrow C)$

The creation and integration procedure

Step 3 (in short): assess the PG properties

At this stage, all the relationships are deemed "satisfied" (i.e., `true`), since they were observed on corpus, but the step is needed because the model allows for more sophisticated grades.

The creation and integration procedure

Step 4: convert the FTB tagset the JDM tagset

- working out the conversion table

Sample of the conversion table

FTB	JDM
ADJ##cpos=A|g=f|n=p|s=ord|pred=y##	Adj:Fem+PL+Ord
DET##cpos=D|g=f|n=s|s=def|pred=y##	Det:Fem+SG+Def
P+D##cpos=P+D|s=def|pred=y##	Pre+Det:
VPP+	Ver:PPas

Evaluation

Integration

How many of the POS in the corpus CFG pre-exist in JDM?

	Connected	Disconnected	Null	Total
Num. nodes	22,742	163,210	9,408	195,360
%	11.6%	83.5%	4.8%	100

$\Leftarrow$ many missing POS nodes in JDM

Perspectives

Grammar Engineering

- Dependency Grammar
- Construction Grammar
- Discourse
- ...

Perspectives

Does an integrated representation of linguistic knowledge help capture interactions across heterogeneous dimensions?

Encoding of graph embeddings

- When syntax is encoded, it rarely (never?) comes from a KG
- very few works on constituency
- How would such embeddings compare to other approaches wherever syntax is required/desired?

Knowledge graph reasoning

- Inference at the interface between syntax, lexical semantics, and other linguistic and ontological dimensions

Conclusion

How to integrate grammar knowledge and lexical-semantic knowledge within a homogeneous knowledge graph?

- we introduced a graph model a phrase structure grammar to enable its integration with a lexical-semantic network
- we proposed a creation process for the model, we implemented it, and experimented with the FTB and JDM
- our implementation shows room for improvement to better integrate the grammar layer with JDM, but the property graph model as such is not to be blamed