A Corpora Based Toy Model for DisCoCat

Stefano Gorgioso

University of Oxford

The Plan

What we assume:

(i) an abstract corpus, as a set/sequence of sentences

(ii) each sentence annotated with a constituent structure tree

- we consider context-free grammars à la Chomsky.

What we do with it:

(i) obtain a toy pregroup grammar from the annotated corpus

- entirely object-oriented, no sentence type

(ii) obtain semantics in a category of R-semimodules[1]

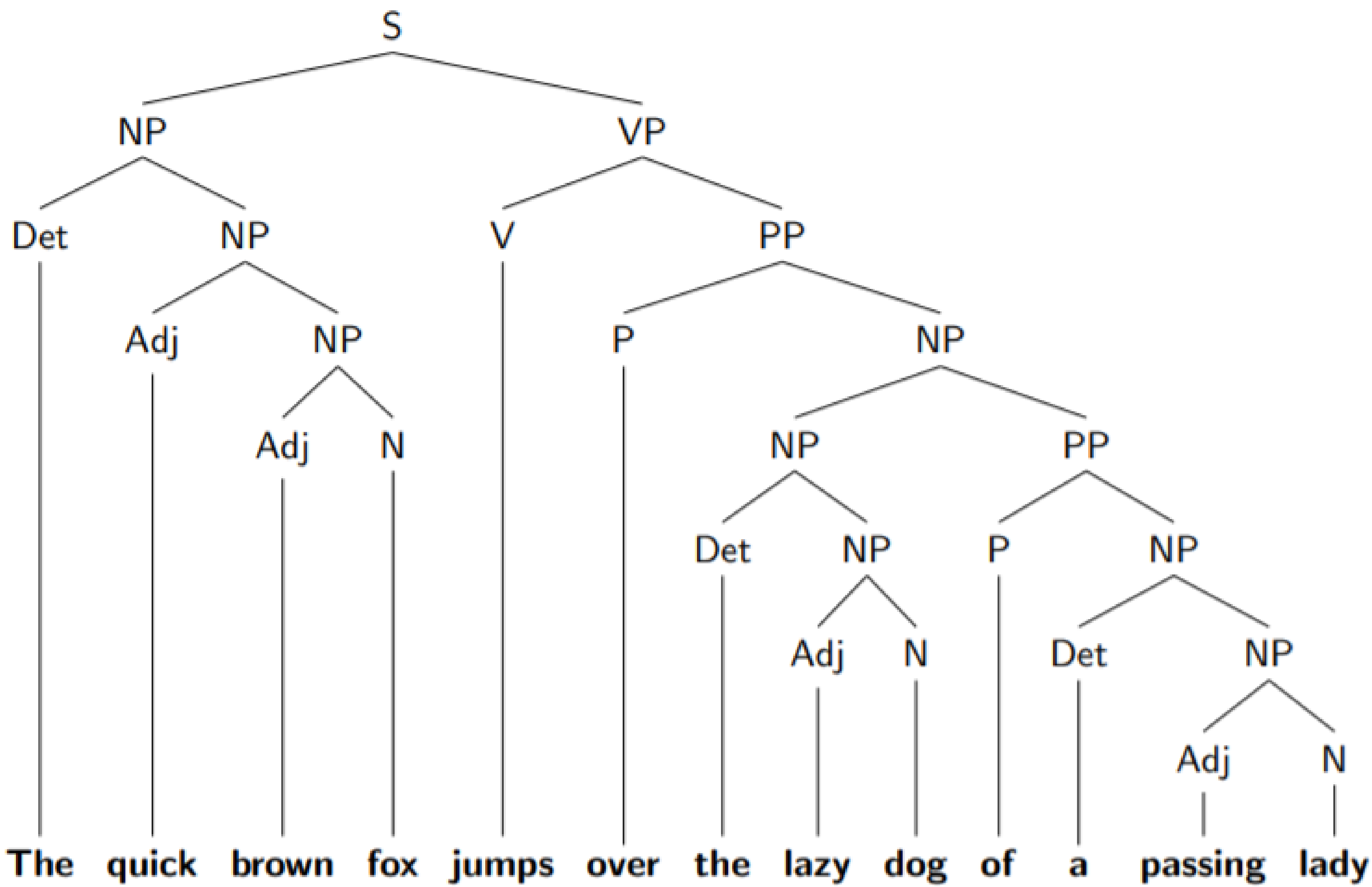
- any involutive commutative semiring R, but here we focus on $\mathbb{N}$

Our semantics are free/minimal, in a certain sense explained later.

---

[1] Free and finite-dimensional.

Constituent Structure Trees



The Pregroup Grammar

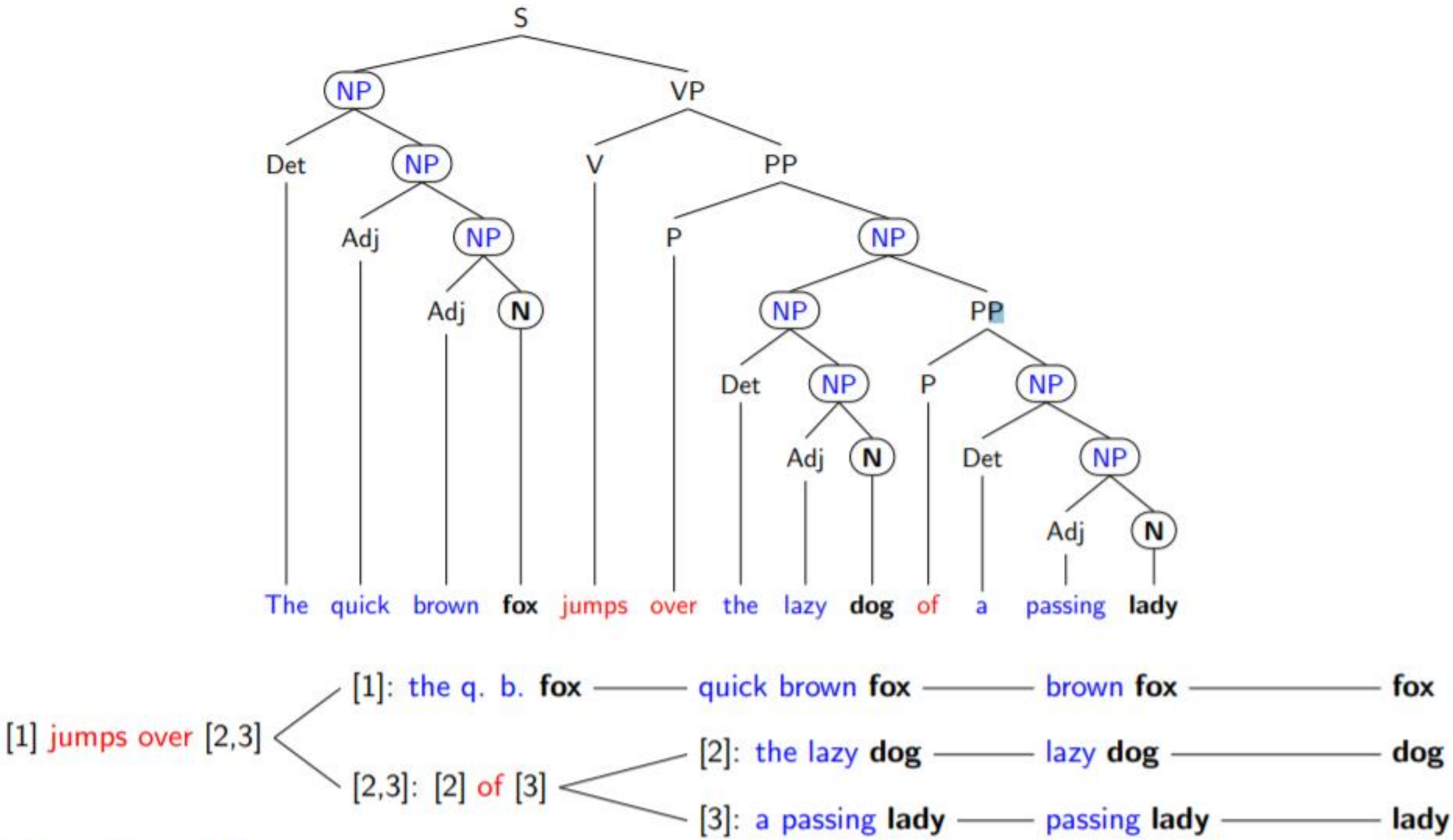
A single atomic type n for *objects*, together with:

(i) object words, of type n
   - the nouns in the corpus sentences

(ii) modifying words, of type $n^r \cdot n$ or $n \cdot n^l$
   - the words modifying nouns/NPs into other NPs

(iii) interaction fragments, of type $n^r \cdot n \cdot n^l$
   - sentence fragments connecting noun phrases

Objects and their Interactions



Interaction words
Modifying words
Interaction fragments

The Category of R-Semimodules

To obtain our semantics, we consider:

(i) an involutive, commutative semiring R

(ii) the category R-Mod of free finite-dim R-semimodules
- objects in the form R^X, for finite sets X
- morphisms $R^X \to R^Y$ are $Y \times X$ R-valued matrices

Many examples of interest are in this form:

- finite sets and relations, for $R = \mathbb{Bool}$
- finite-dim real/complex vector spaces, for $R = \mathbb{R}, \mathbb{C}$
- finite-dim convex cones[2], for $R = \mathbb{R}^+$
- finite multi-sets and "multi-relations", for $R = \mathbb{N}$

[2] Including probability distributions and stochastic maps.

The Category of R-Semimodules

Some desirable features of the category of R-semimodules:

(i) R-Mod is a †-symmetric monoidal category
(ii) R-Mod is compact closed, with self-dual objects
(iii) R-Mod has classical structures associated to canonical bases

The Distributional Part

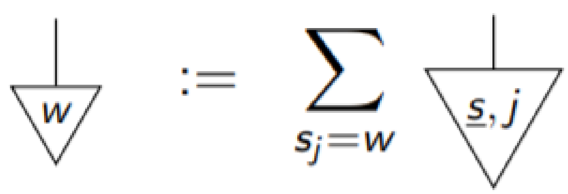
We construct our semantic space from the corpus:

(i) consider the set X of all word instances in all sentences

$$X = \{(\underline{s}, j) \mid \underline{s} \text{ sentence}, j \text{ index of word instance in } \underline{s}\}$$

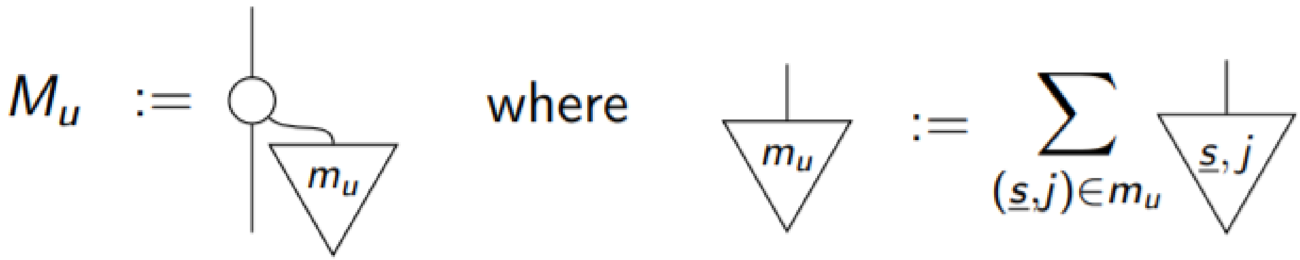
(ii) take R^X as the semantic space

(iii) embed a word w as the indicator function of all its instances



The Compositional Part

Modifier words are mapped to projectors:



We define m_u to be the set of instances of object words which appear in objects modified by u. For example, from our sentence we'd have:

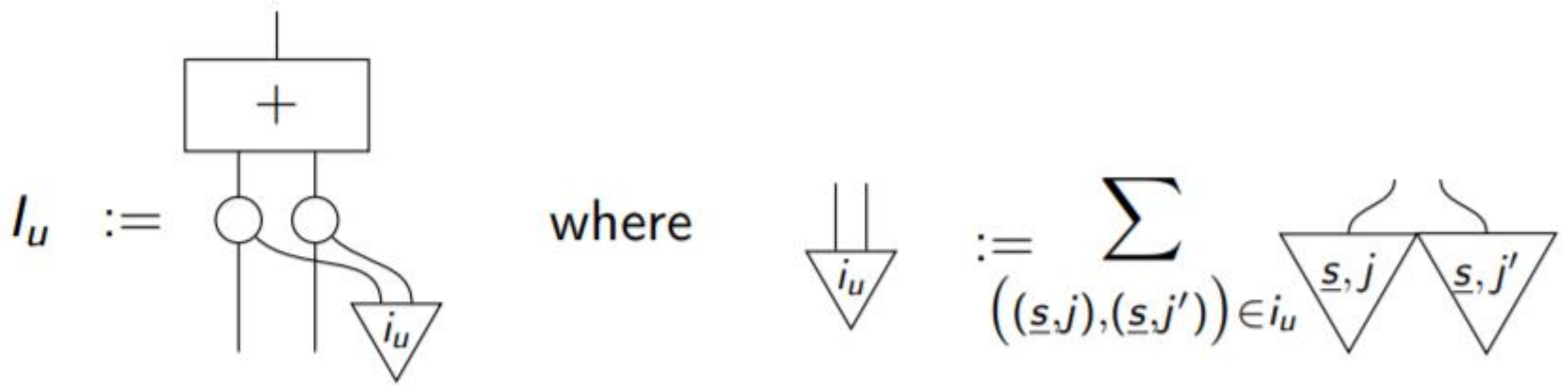
$$\{\text{fox}\} \subseteq m_{\text{quick}}$$
$$\{\text{lady}\} \subseteq m_{\text{passing}}$$
$$\{\text{fox}, \text{dog}\} \subseteq m_{\text{the}}$$

We automagically get some logic out of operator algebra[3].

[3] As long as R satisfies the additive cancellation law.

The Compositional Part

Interaction fragments are mapped to binary operations. First we construct the operation for single-word fragments:

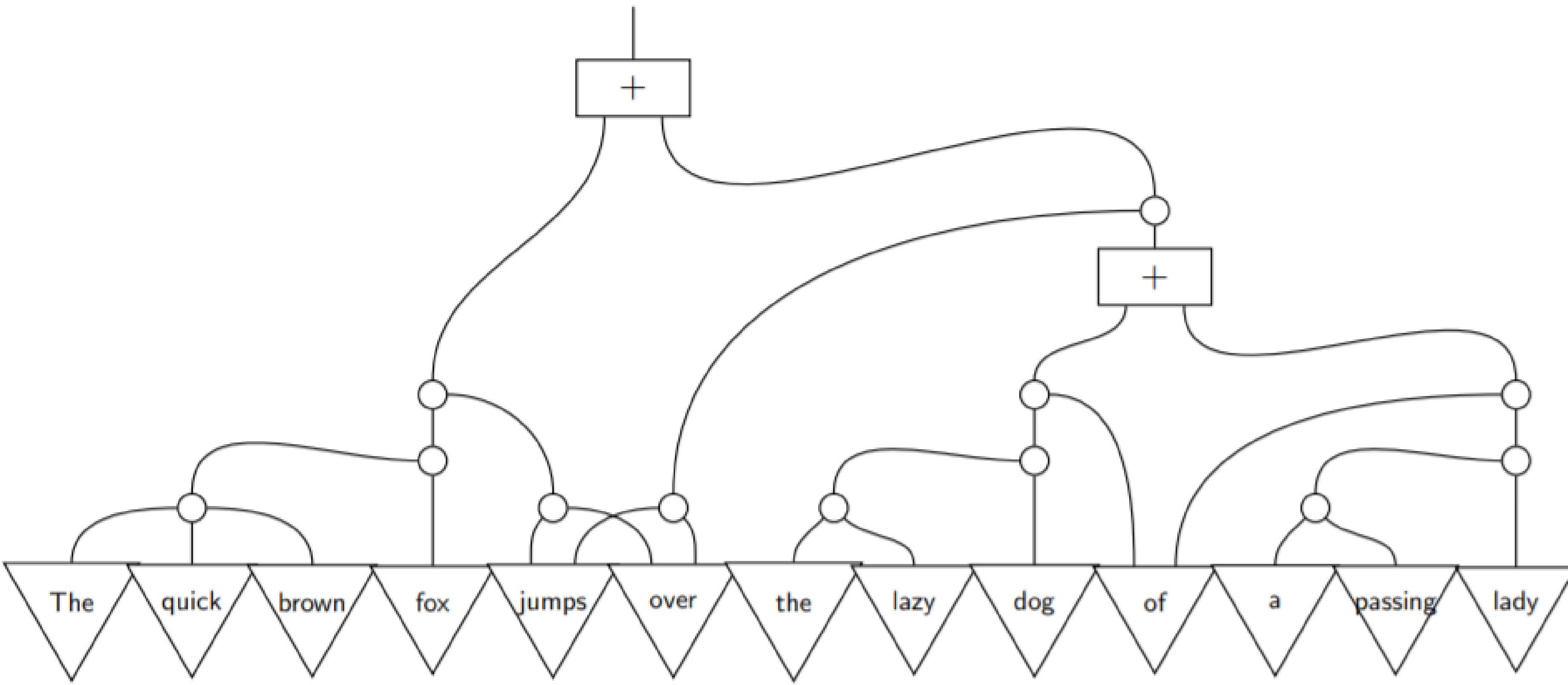


We define i_u to be the set of pairs of instances which appear in objects put into relation by u. For example, from our sentence we'd have:

$$\{(\text{fox},\text{dog}),(\text{fox},\text{lady})\} \subseteq i_{\text{jumps}} \tag{0.1}$$

The End Result

Here is the resulting[4] semantics for our sentence:



[4] After some applications of the spider theorem, to group modifier words together.

Future work

A lot of things to do!

(i) Toy model needs a number of improvements
  - treatment of personal/possessive pronouns
  - treatment of conjunctions

(ii) More sophisticated choice of semiring
  - encoding of polarity, modality and inflection

(iii) Compressing the free model to obtain concrete models
  - change of semiring + linear compression $\Rightarrow$ more semantics?

(iv) CPM construction (possibly iterated)
  - treatment ambiguity and entailment

(v) Enriched/higher order categories
  - encode simplicial structure extracted from the corpus

Thank you for your attention!