An English Dependency Treebank à la Tesnière

Federico Sangati[1] Chiara Mazza[2]

[1]University of Amsterdam

[2]University of Pisa

December 4, 2009

Lucien Tesnière

50 years ago

LUCIEN TESNIÈRE

ÉLÉMENTS
DE
SYNTAXE
STRUCTURALE

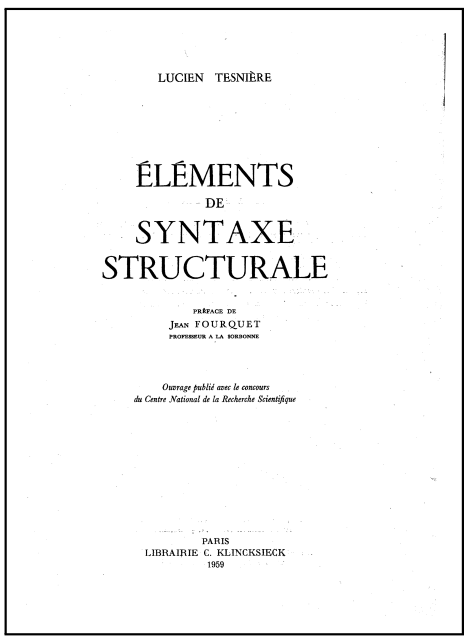
PRÉFACE DE
JEAN FOURQUET
PROFESSEUR A LA SORBONNE

Ouvrage publié avec le concours
du Centre National de la Recherche Scientifique

PARIS
LIBRAIRIE C. KLINCKSIECK
1959

Lucien Tesnière

3. — On peut ainsi comparer le verbe à une sorte d'**atome crochu** susceptible d'exercer son attraction sur un nombre plus ou moins élevé d'actants, selon qu'il comporte un nombre plus ou moins élevé de crochets pour les maintenir dans sa dépendance. Le nombre de crochets que présente un verbe et par conséquent le nombre d'actants qu'il est susceptible de régir, constitue ce que nous appellerons la **valence** du verbe.



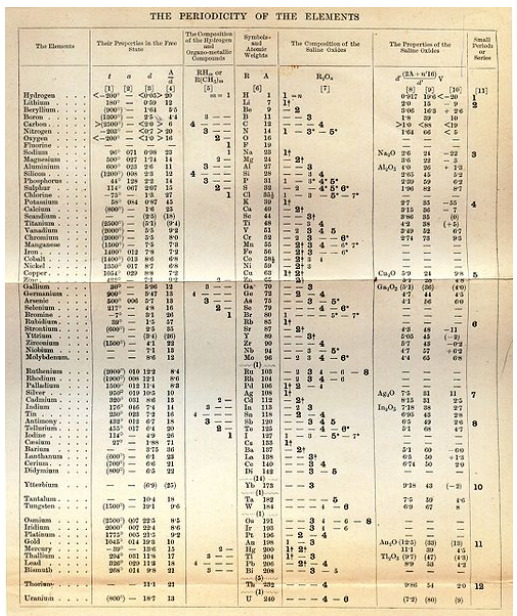
LUCIEN TESNIÈRE

ÉLÉMENTS
DE
SYNTAXE
STRUCTURALE

PRÉFACE DE
JEAN FOURQUET
PROFESSEUR A LA SORBONNE

Ouvrage publié avec le concours du Centre National de la Recherche Scientifique

PARIS
LIBRAIRIE C. KLINCKSIECK
1959

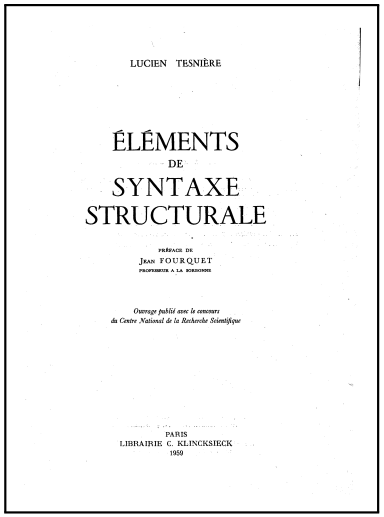


THE PERIODICITY OF THE ELEMENTS

Mendeleev's periodic table, 1891

Outline

Features and Operations


LUCIEN TESNIÈRE

ÉLÉMENTS
DE
SYNTAXE
STRUCTURALE

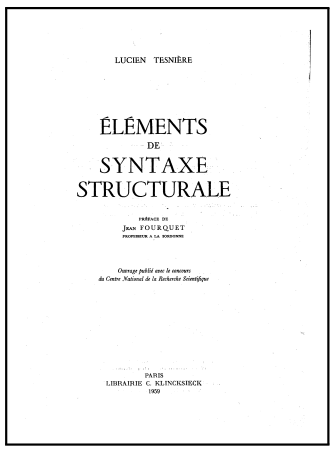
PRÉFACE DE
JEAN FOURQUET
PROFESSEUR À LA SORBONNE

Ouvrage publié avec le concours du Centre National de la Recherche Scientifique

PARIS
LIBRAIRIE C. KLINCKSIECK
1959

Connexion	dependency relation between words
Words types	empty/full words
Nucléus	block of words as intermediate linguistic units
Catégories	noun, verb, adjective, adverb
Jonction	coordination and other types of conjoined structures
Translation	transferences operation to generalize over the categories of the linguistic elements

Dependency Relations (Connexion)


LUCIEN TESNIÈRE

ÉLÉMENTS
DE
SYNTAXE
STRUCTURALE

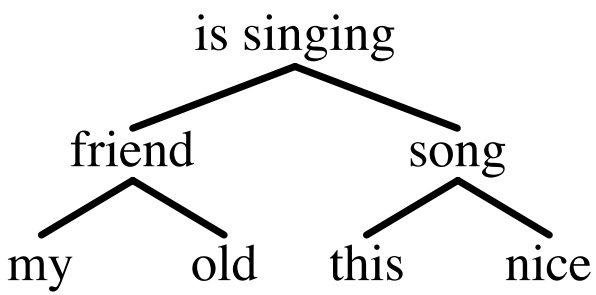
PRÉFACE DE
JEAN FOURQUET

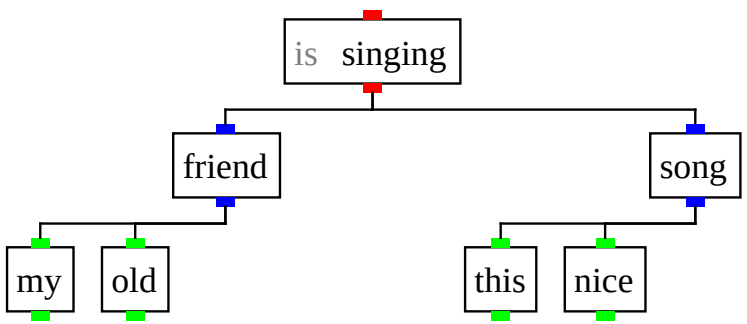
PARIS
LIBRAIRIE C. KLINCKSIECK
1959

1. — Les connexions structurales établissent entre les mots des rapports de **dépendance.** Chaque connexion unit en principe un terme **supérieur** à un terme **inférieur.**

7. — La connexion est **indispensable** à l'expression de la pensée. Sans la connexion, nous ne saurions exprimer aucune pensée continue et nous ne pourrions qu'énoncer une succession d'images et d'idées isolées les unes des autres et sans lien entre elles [1].

8. — C'est donc la connexion qui donne à la phrase son caractère **organique** et **vivant,** et qui en est comme le **principe vital.**



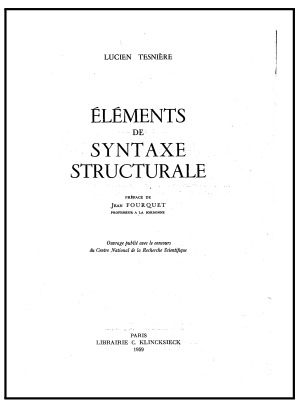


Word types

All words are divided into two classes:

- **Full content words**: nouns, verbs, adjectives, etc.
- **Empty functional words**: determiners, prepositions, etc.

e.g. Snoopy is flying on the doghouse


LUCIEN TESNIÈRE

ÉLÉMENTS
DE
SYNTAXE
STRUCTURALE

PRÉFACE DE
Jean FOURQUET

Ouvrage publié avec le concours
du Centre National de la Recherche Scientifique

PARIS
LIBRAIRIE C. KLINCKSIECK
1959

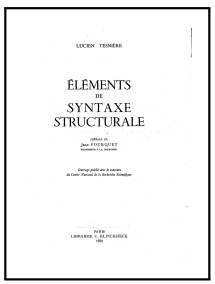
2. — Les mots **pleins** sont ceux qui sont **chargés d'une fonction sémantique,** c'est-à-dire ceux dont la forme est associée directement à une idée, qu'elle a pour fonction de représenter et d'évoquer. Ainsi fr. *cheval*, all. *Pferd*, angl. *horse*, lat. *equus*, etc... sont des mots pleins, parce que leur forme, c'est-à-dire les phonèmes (ou les lettres) qui les composent suffisent à évoquer l'idée d'un cheval.

3. — Les mots **vides** sont ceux qui ne sont pas chargés d'une fonction sémantique. Ce sont de simples **outils grammaticaux** [1] dont le rôle est uniquement d'indiquer, de préciser ou de transformer la catégorie des mots pleins et de régler leurs rapports entre eux.

Block of Words (Nucléus)

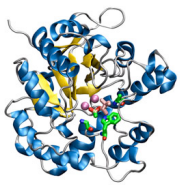
A *block* always includes a single full word and any number of empty words (possibly none).

e.g. [Snoopy] [is flying] [on the doghouse]


LUCIEN TESNIÈRE

ÉLÉMENTS DE SYNTAXE STRUCTURALE

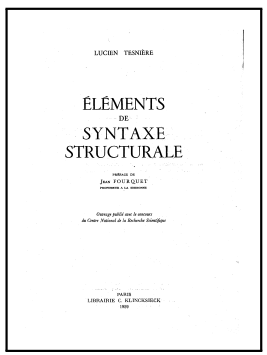
14. — Le nucléus est donc en dernière analyse l'entité syntaxique élémentaire, le matériau fondamental de la charpente structurale de la phrase, et en quelque sorte la **cellule** constitutive qui en fait un organisme vivant.



Word level | Block | Sentence level

Categories (Catégories)

Tesnière distinguishes four *block categories*: nouns, adjectives, verbs, adverbs.


LUCIEN TESNIÈRE

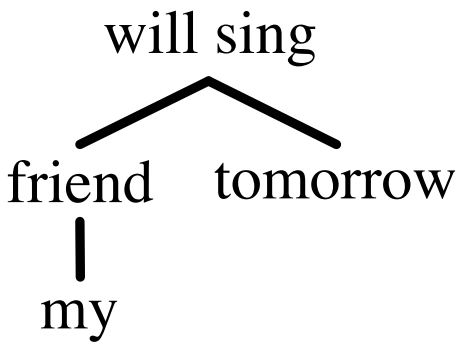
ÉLÉMENTS
DE
SYNTAXE
STRUCTURALE

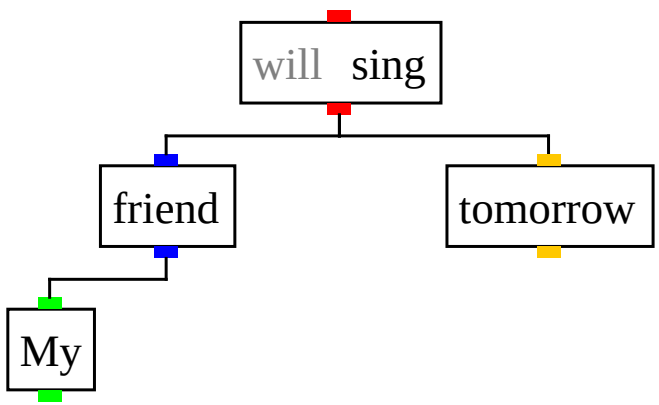
PARIS
LIBRAIRIE C. KLINCKSIECK

2. — Nous adopterons les **représentations symboliques** suivantes :

O = Substantif.
A = Adjectif.
I = Verbe.
E = Adverbe.

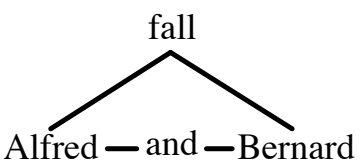
3. — On notera que les quatre lettres adoptées correpondent aux terminaisons des quatre espèces de mots correspondantes en **espéranto** : *-o* pour le substantif, *-a* pour l'adjectif, *-i* pour l'infinitf, *-e* pour l'adverbe.

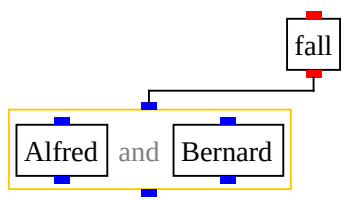


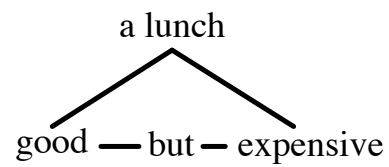


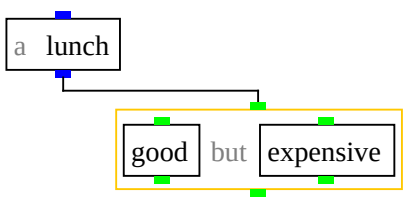
Junction (Jonction)

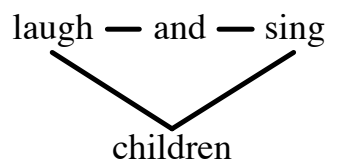
- It groups blocks, the ***conjuncts***, into a unique block entity.
- The conjuncts are connected horizontally by means of empty words, the ***conjunctions*** (possibly missing).

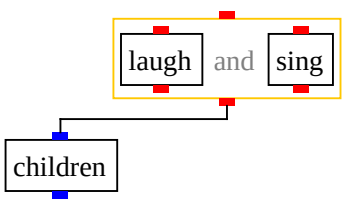




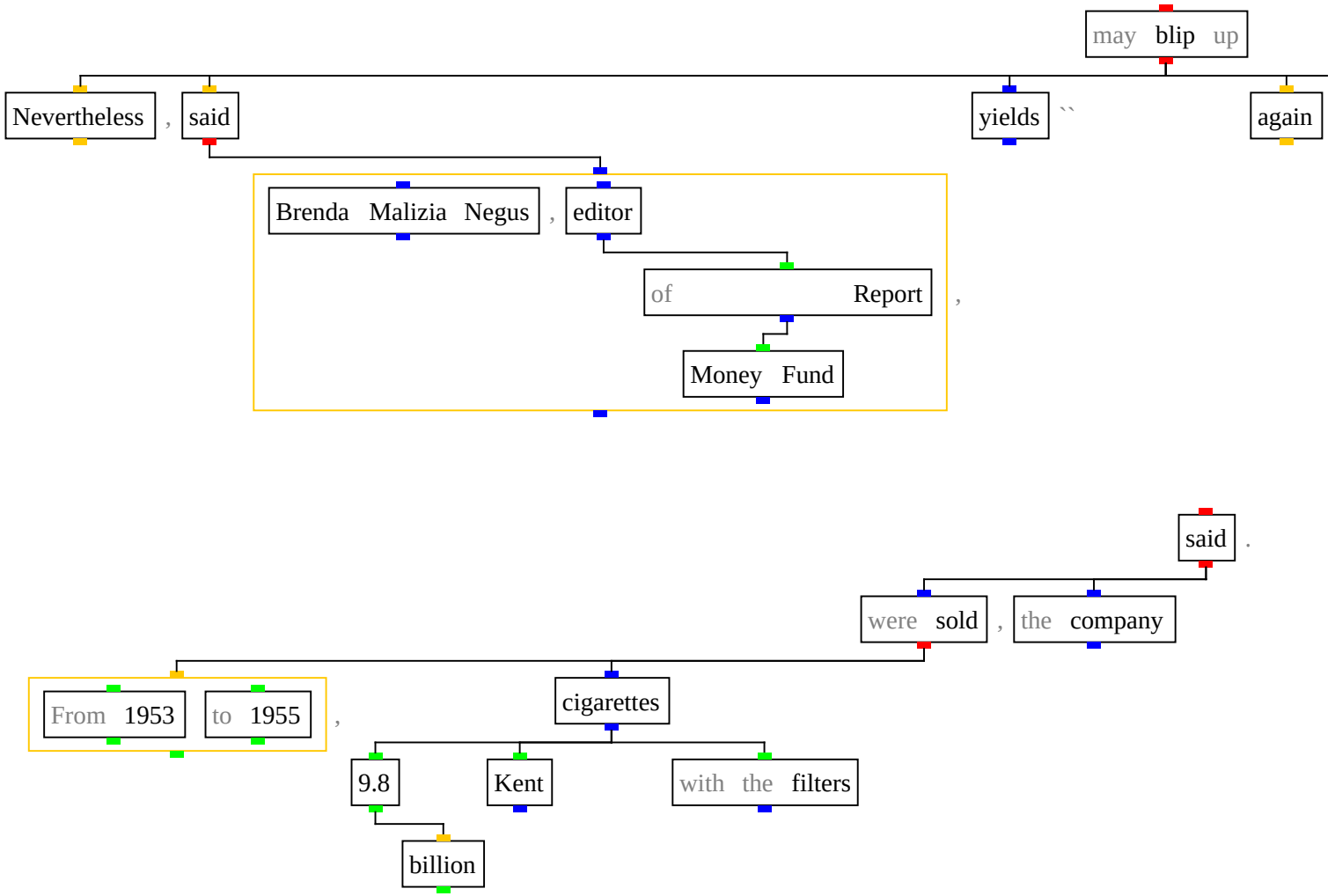






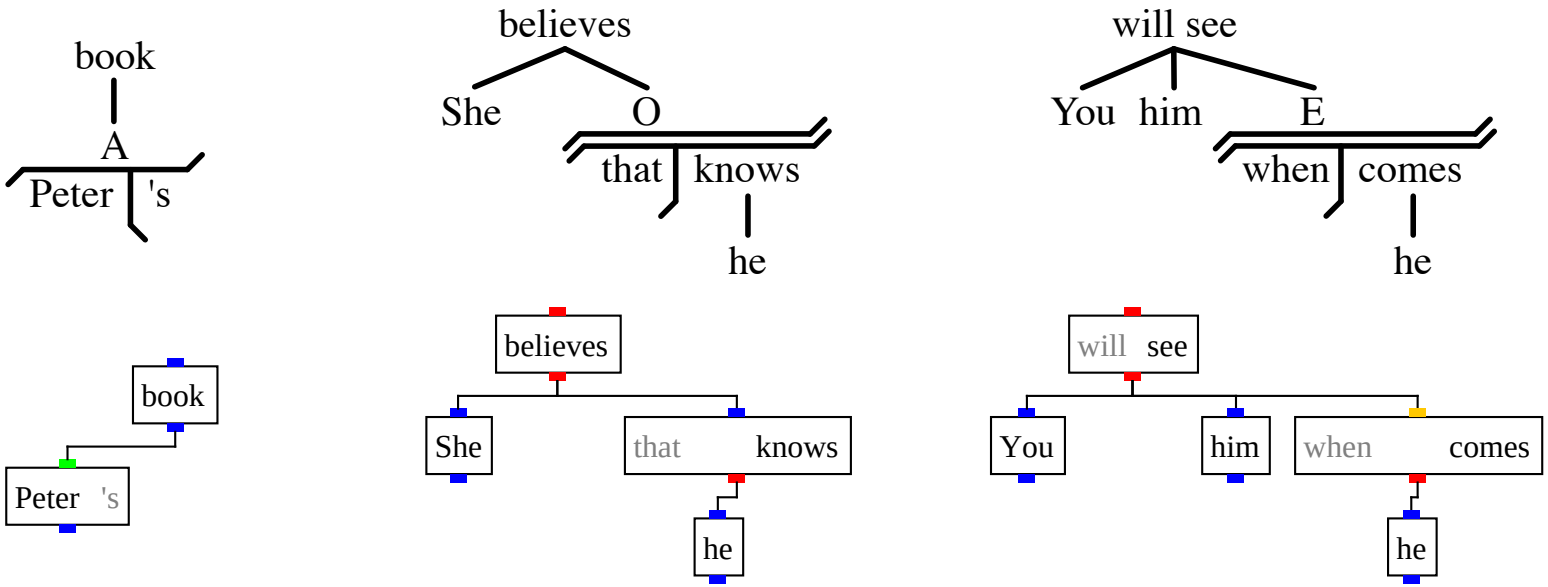


Junction (besides coordination)



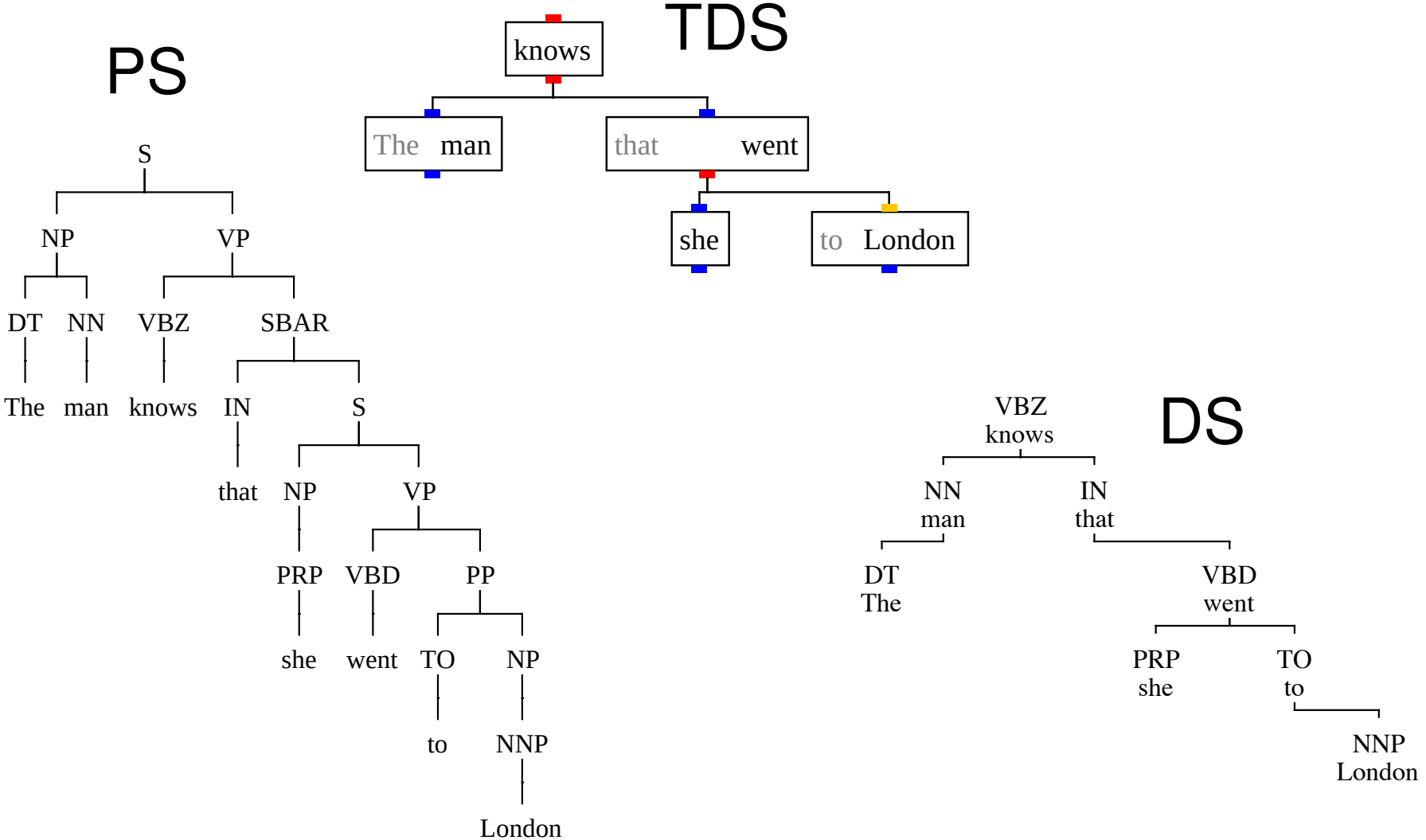
Transference (Translation)

A **shifting process** which makes a block change from the original category of the content word, to another category, by means of zero or more empty words belonging to the same block, called ***transferrers***.

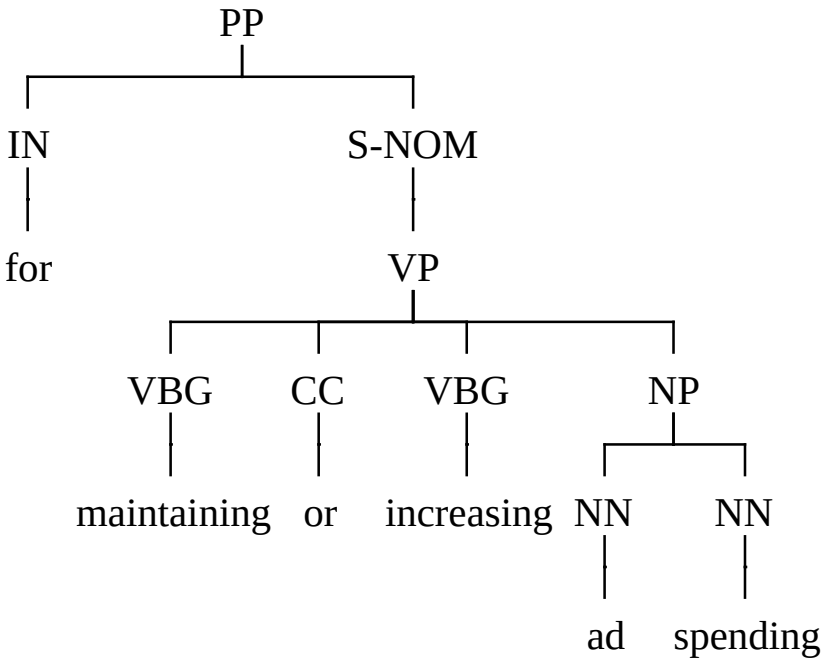


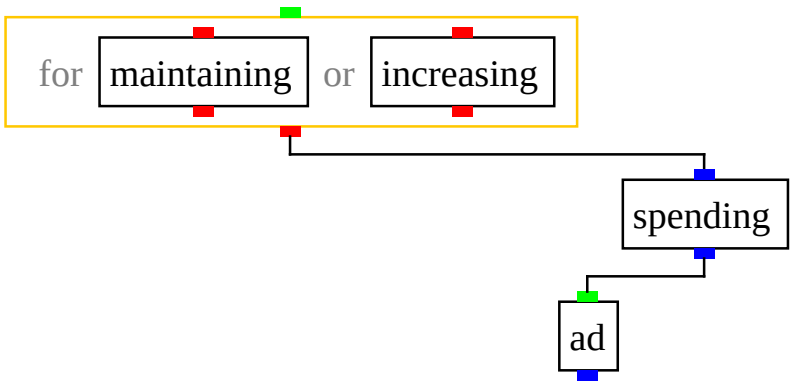
Advantages of TDS

In between PS and DS

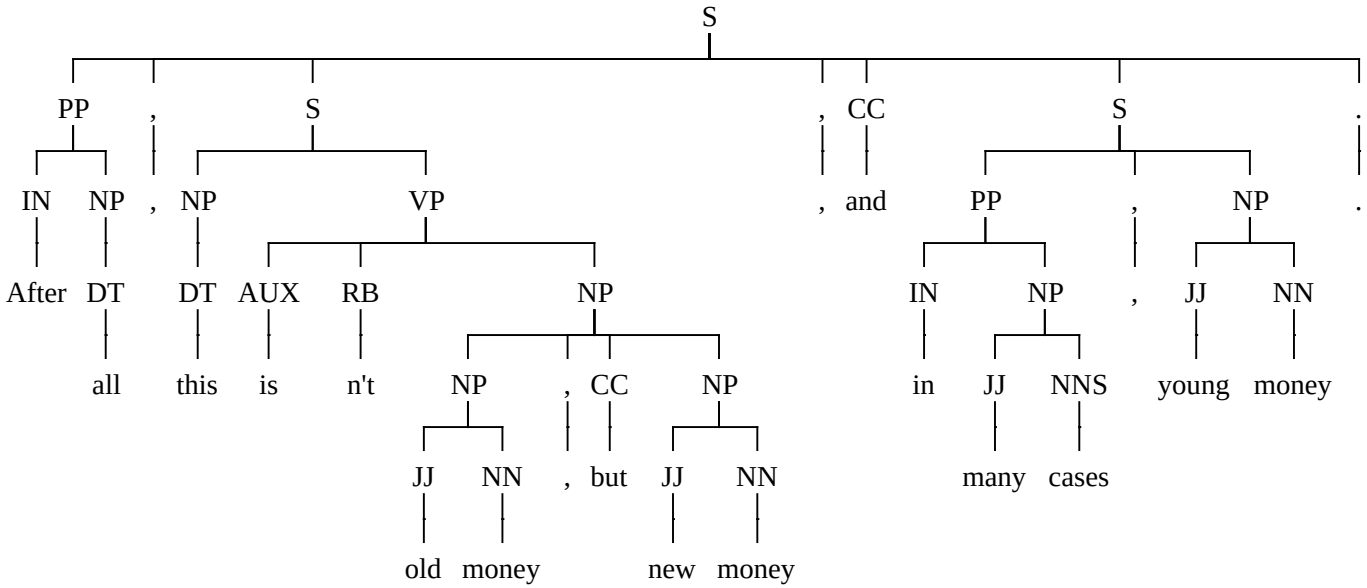


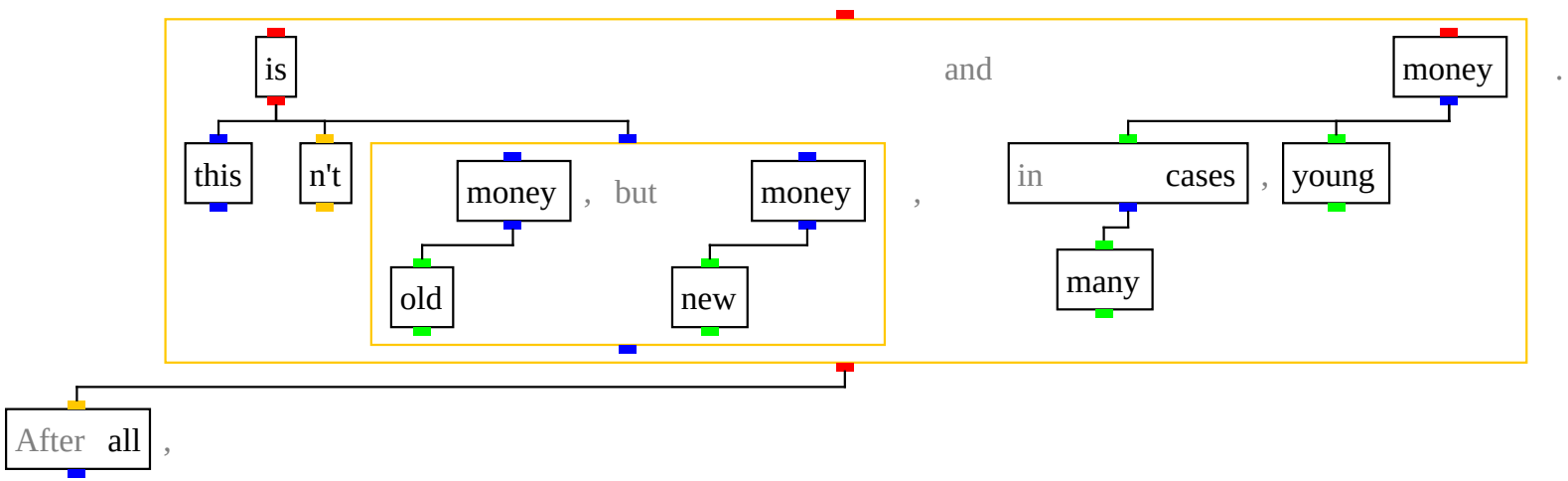
TDS vs. PS: Coordination



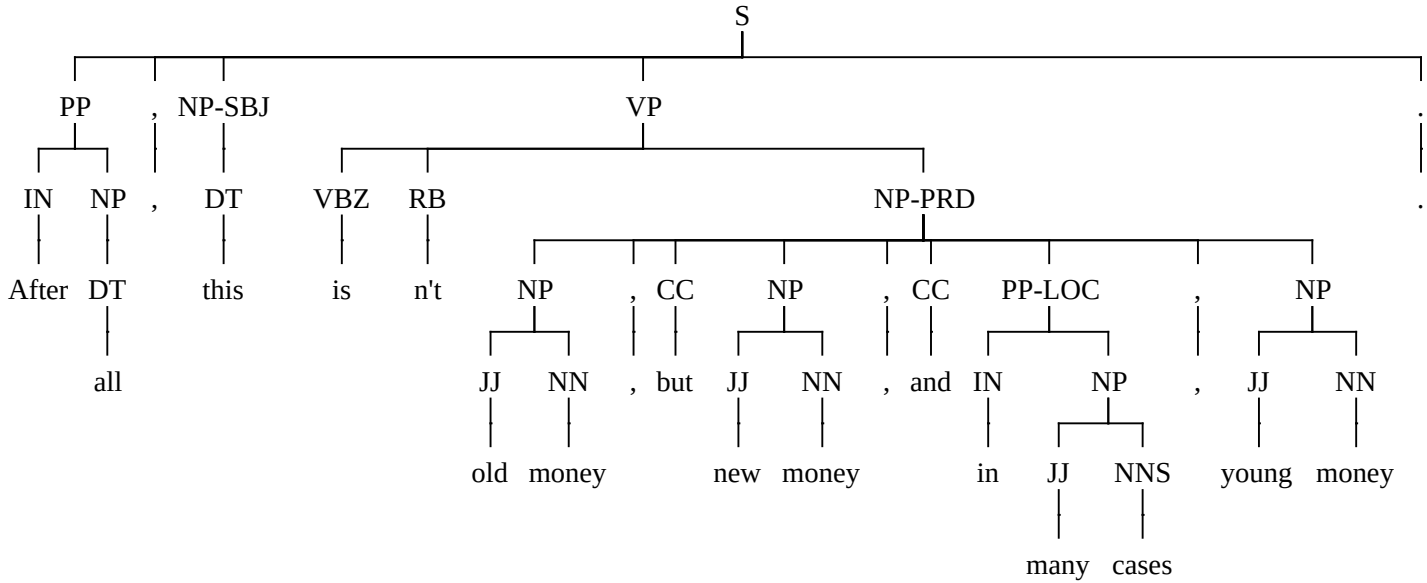


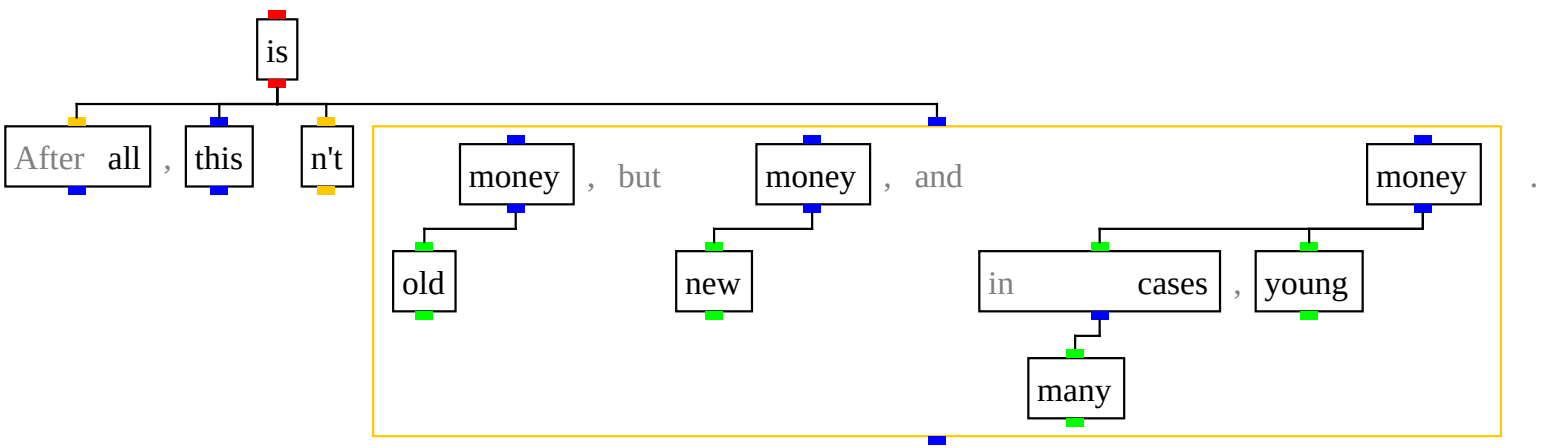
TDS vs. PS: Charniak's bad parses





TDS vs. PS: Charniak's bad parses

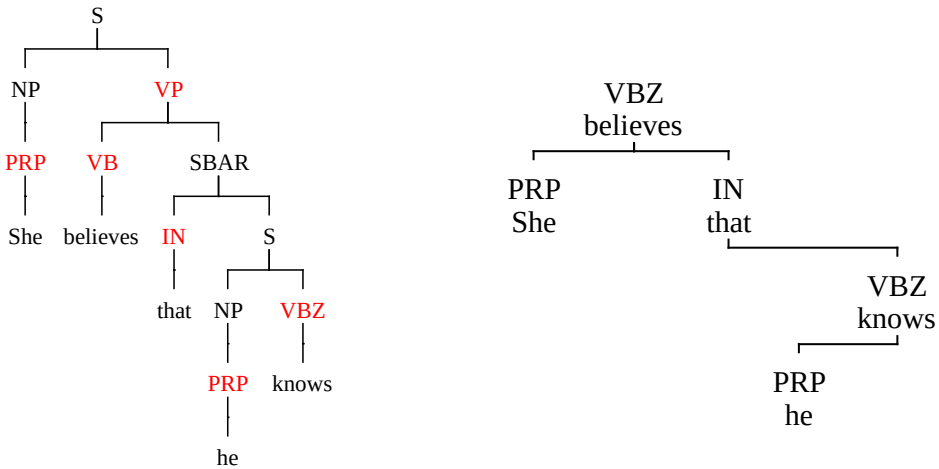




TDS vs. DS: Choosing the correct heads

First step in PS-to-DS conversion: annotate the PS with **heads**.

Sangati & Zuidema, Unsupervised methods for head assignments, EACL 2009.

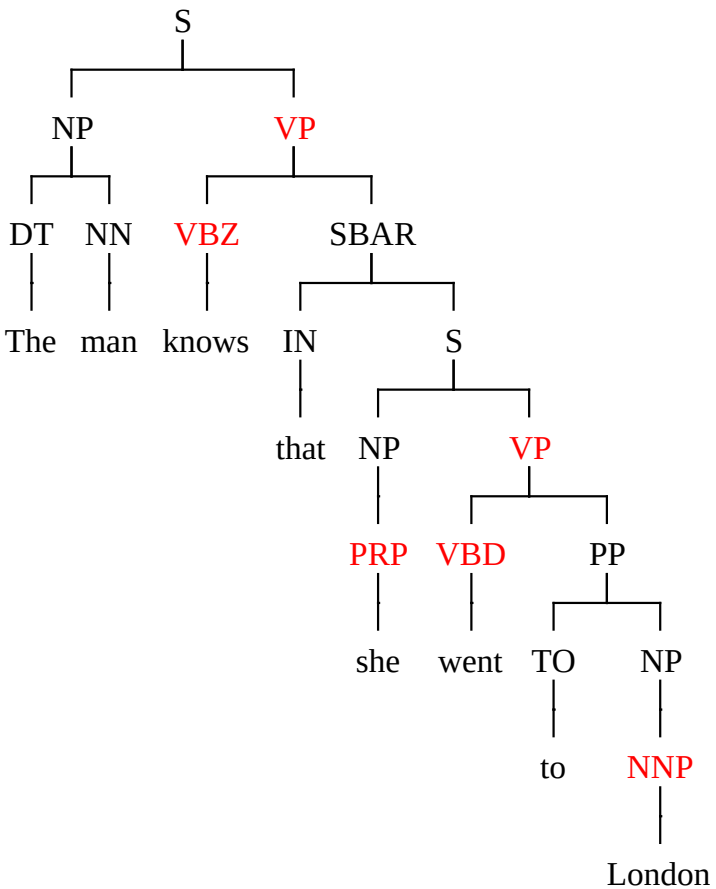


Easy choices e.g. the verb is the head of S.

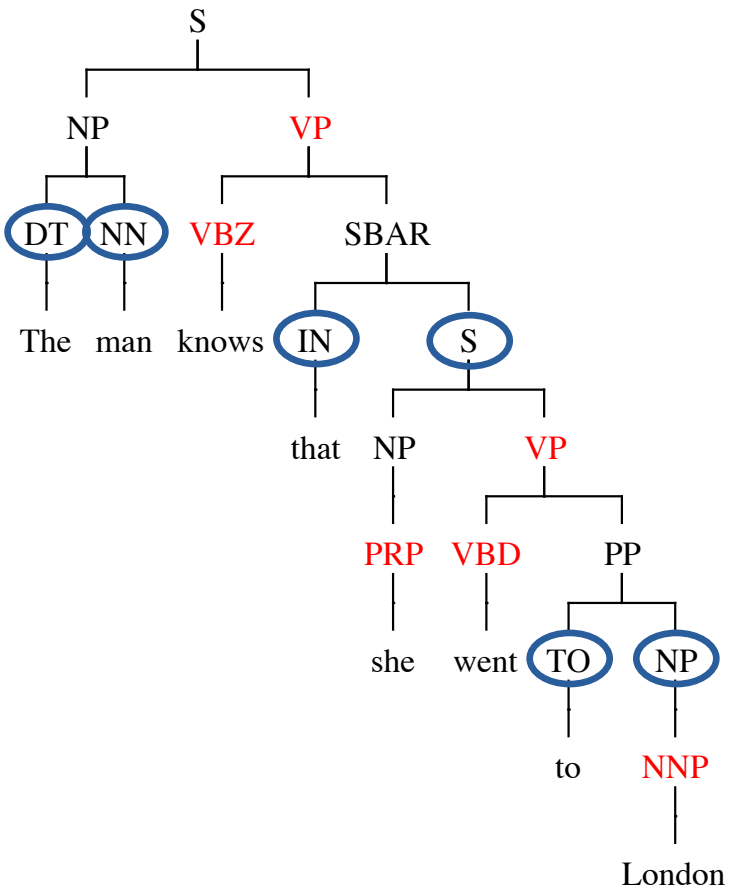
Arbitrary choices should be consistent, most common cases:

- Determiner vs. noun in NP (*the man*).
- Preposition vs. noun in PP (*on paper*).
- Complementizer vs. verb in SBAR (*I believe that she knows*).

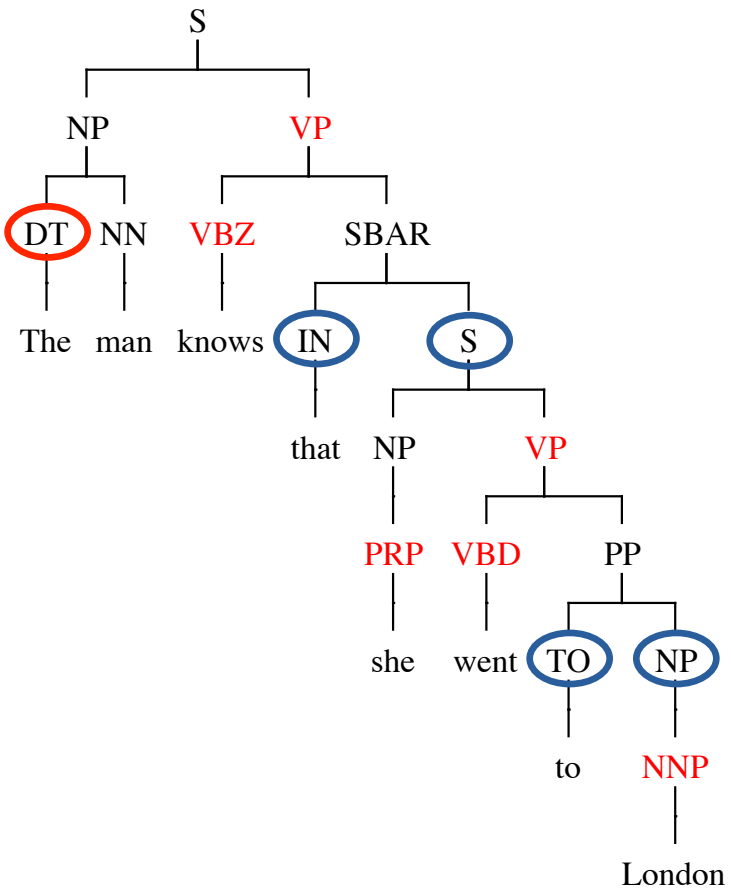
TDS vs. DS: Choosing the correct heads

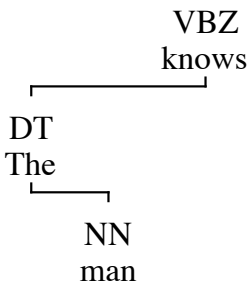


TDS vs. DS: Choosing the correct heads

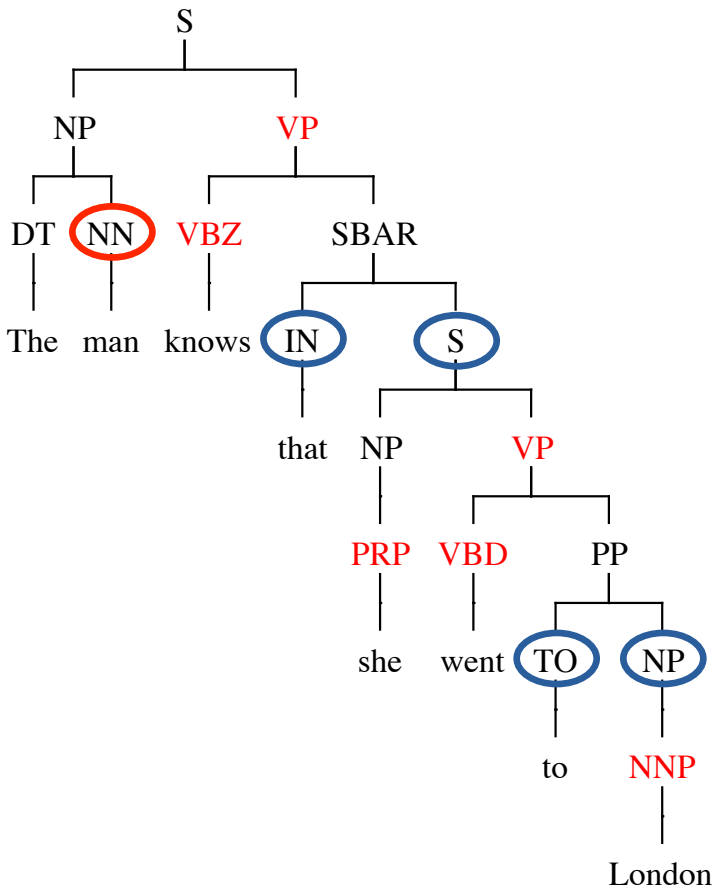


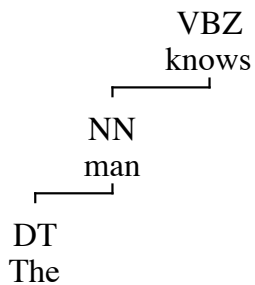
TDS vs. DS: Choosing the correct heads



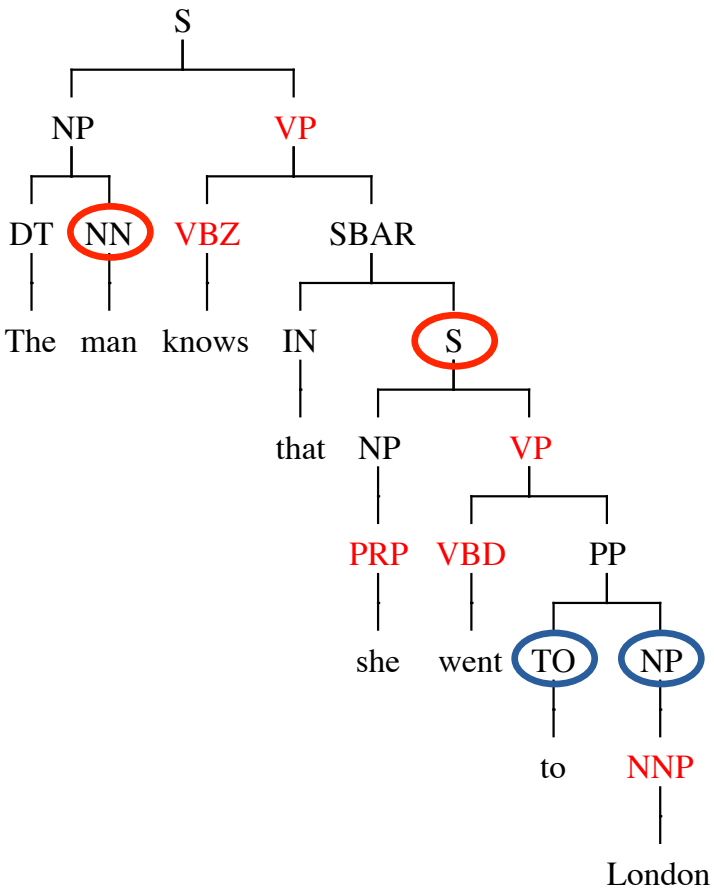


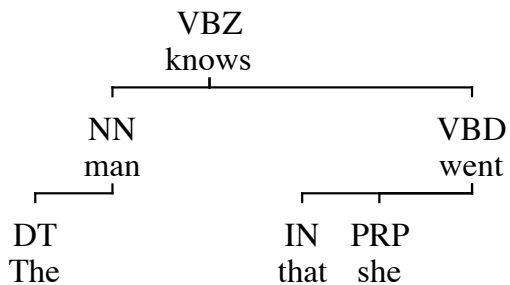
TDS vs. DS: Choosing the correct heads



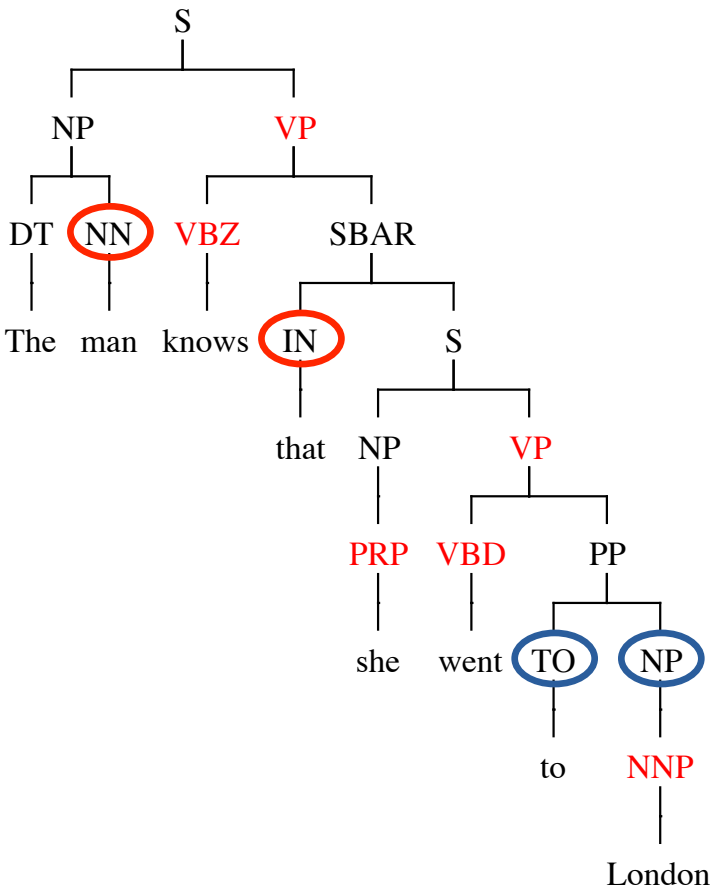


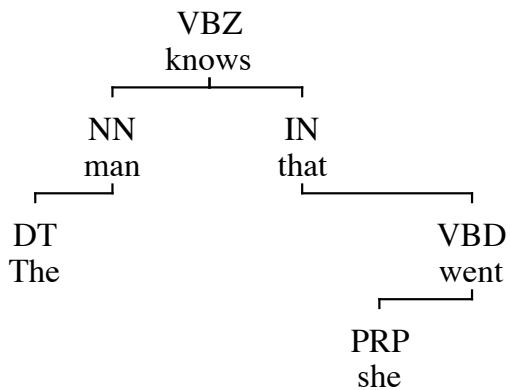
TDS vs. DS: Choosing the correct heads



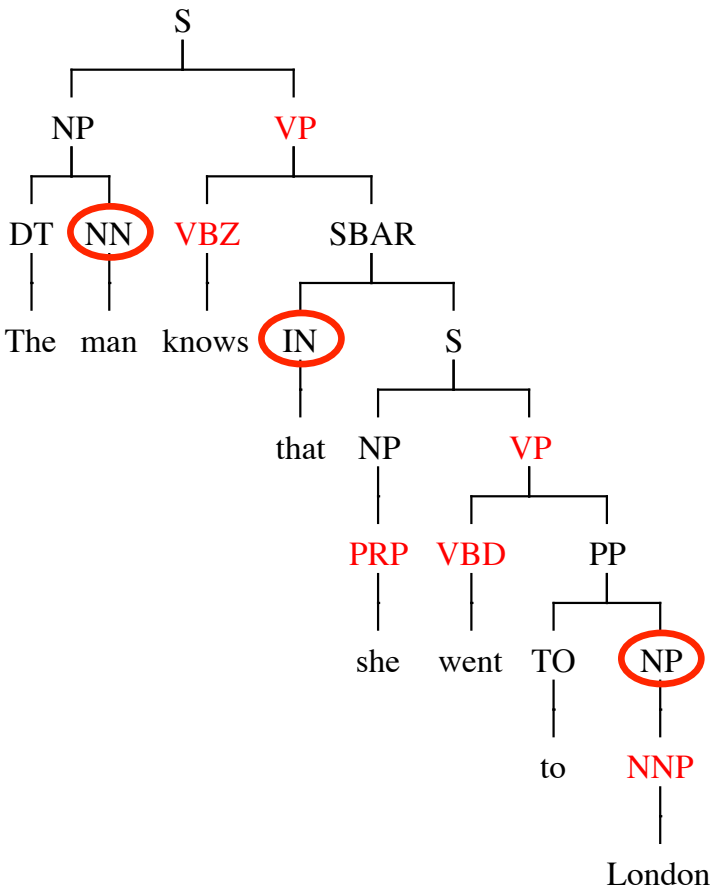


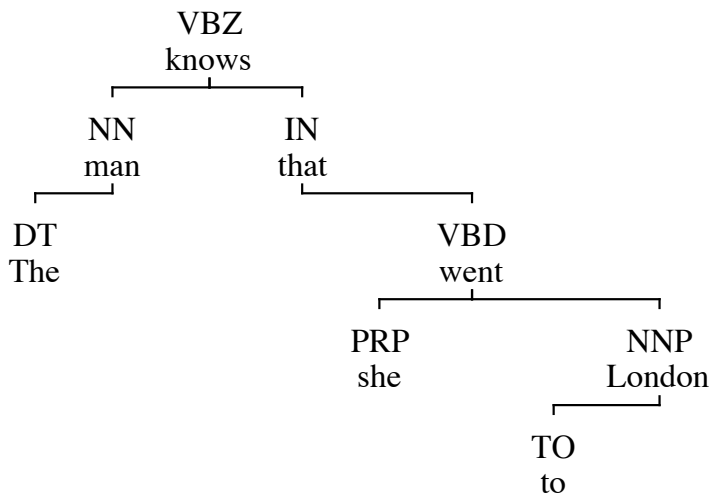
TDS vs. DS: Choosing the correct heads



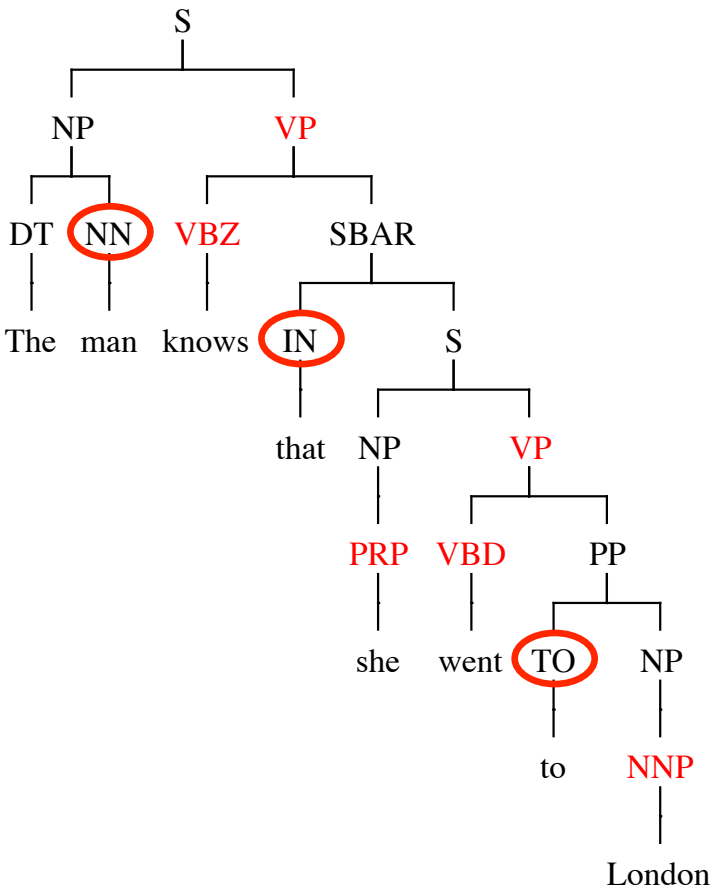


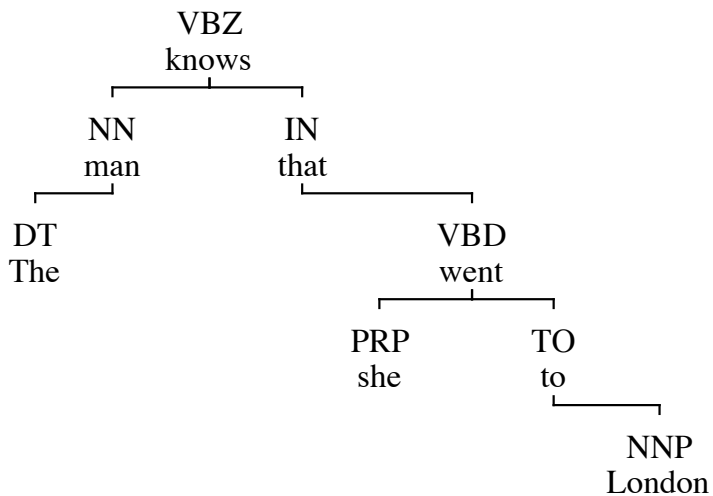
TDS vs. DS: Choosing the correct heads



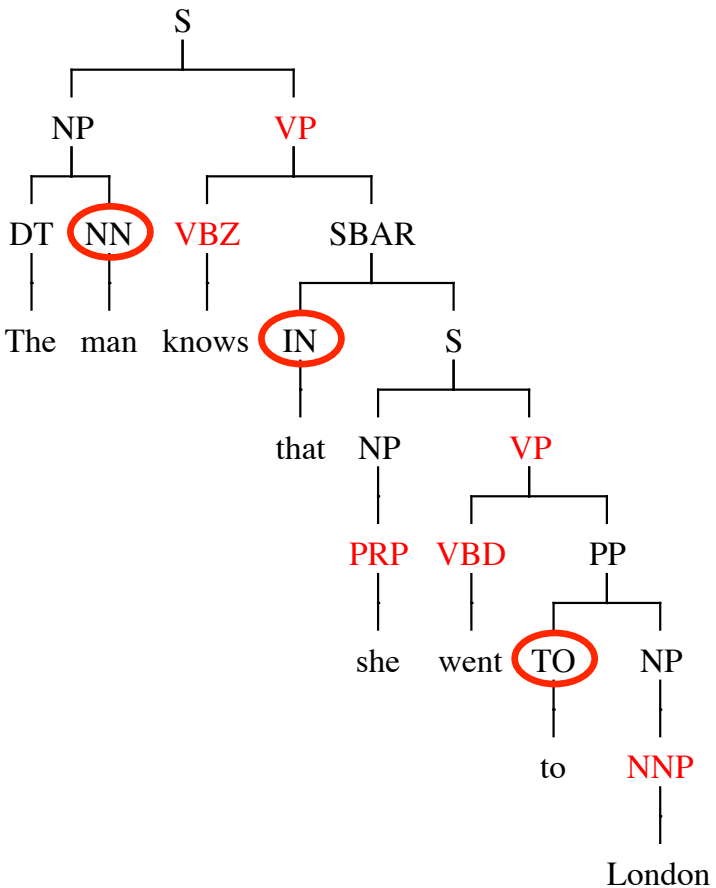


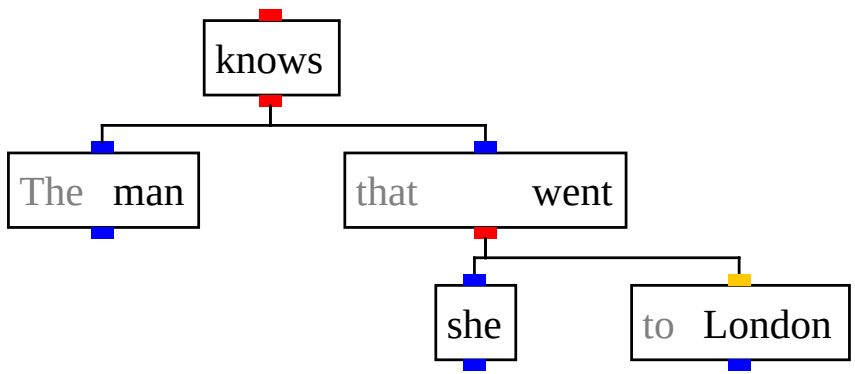
TDS vs. DS: Choosing the correct heads





TDS vs. DS: Choosing the correct heads

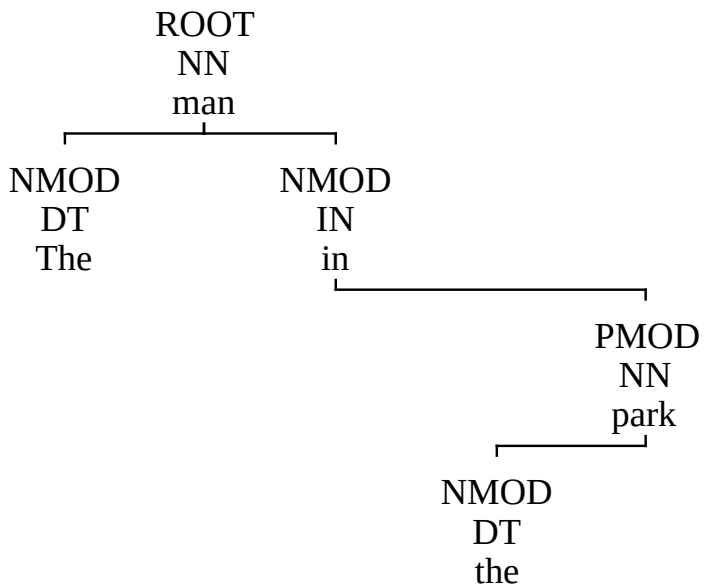




TDS vs. DS: Categories and Blocks

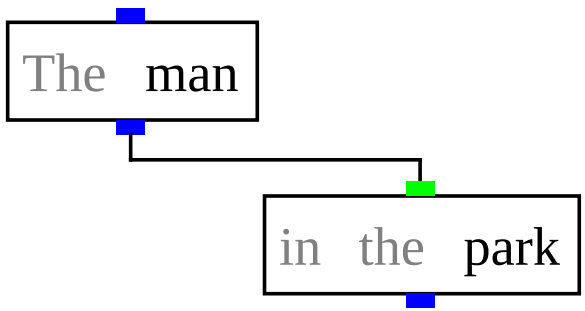
DS

ROOT, SBJ, OBJ, PRN, P, COORD, ADV, VC, VMOD, NMOD, AMOD, PMOD, DEP



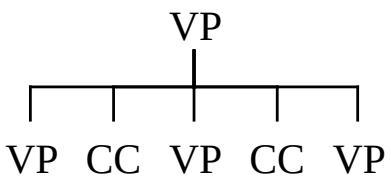
TDS

nouns, adjectives, verbs, adverbs



TDS vs. DS: Coordination structures

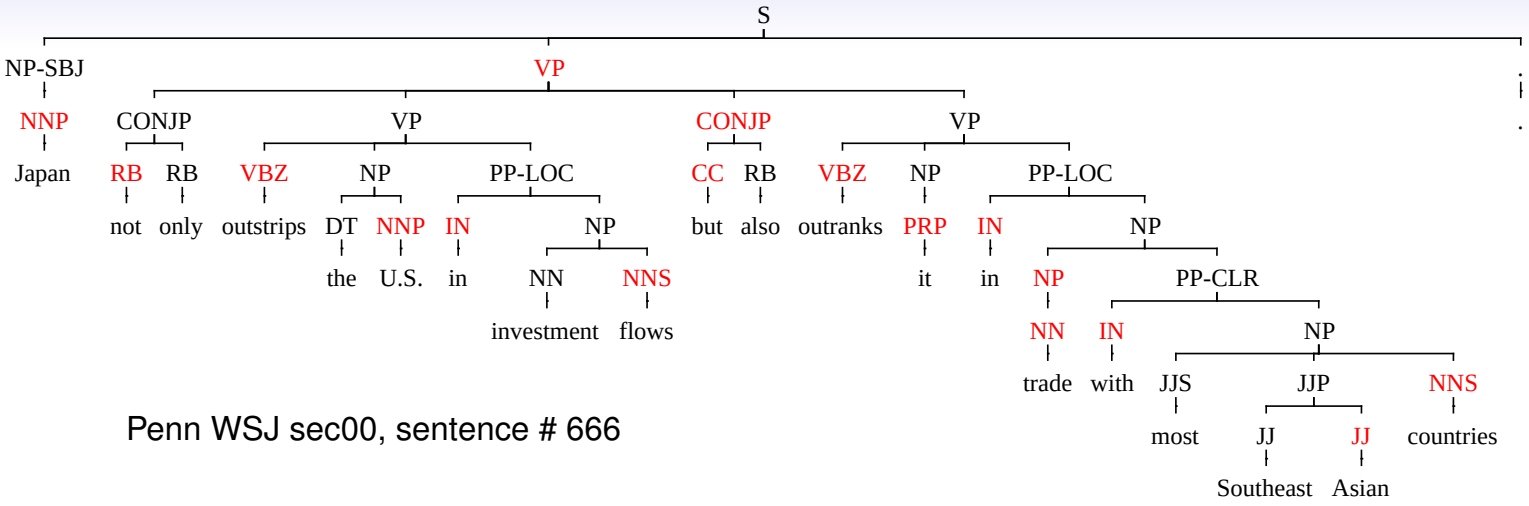
Coordination represents one of the major problems in currently used DS representations, especially when **multiple conjunctions** are present.



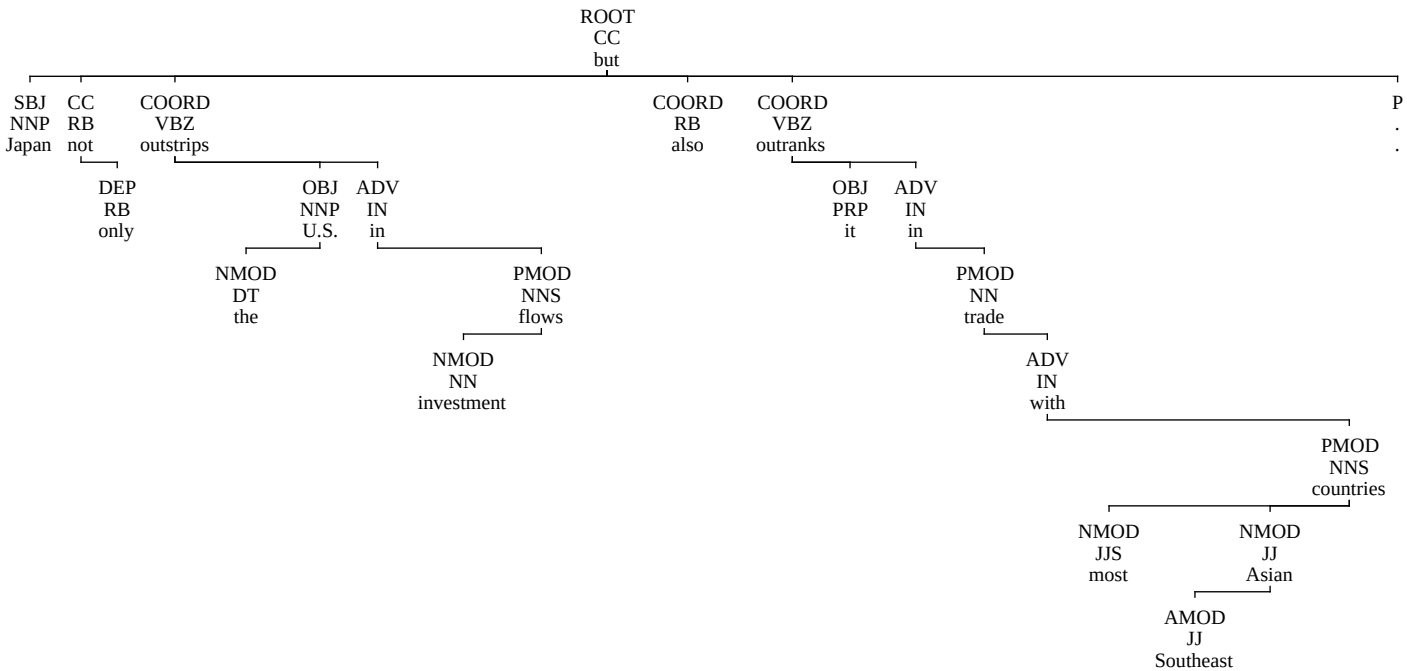
Possible solutions:

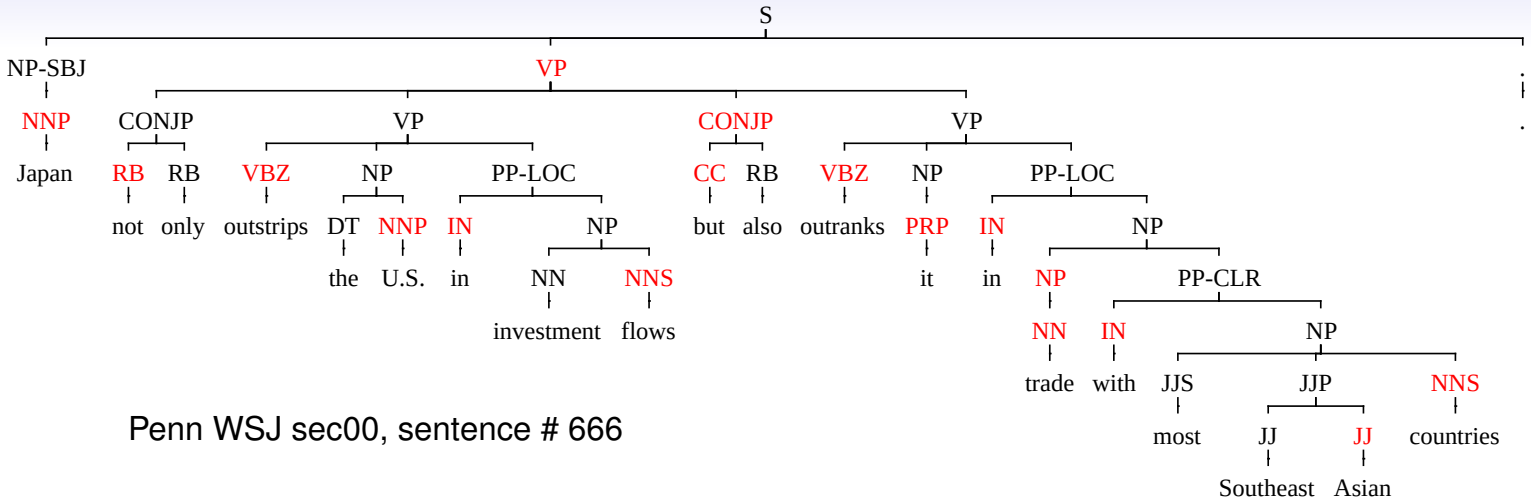
1. One conjunction (or conjunct) is the head of the other elements (most common case).
2. Each element (conjunction or conjunct) is the head of the adjacent element which follows (cf. Mel'čuk, 1988).

Both solutions are problematic.

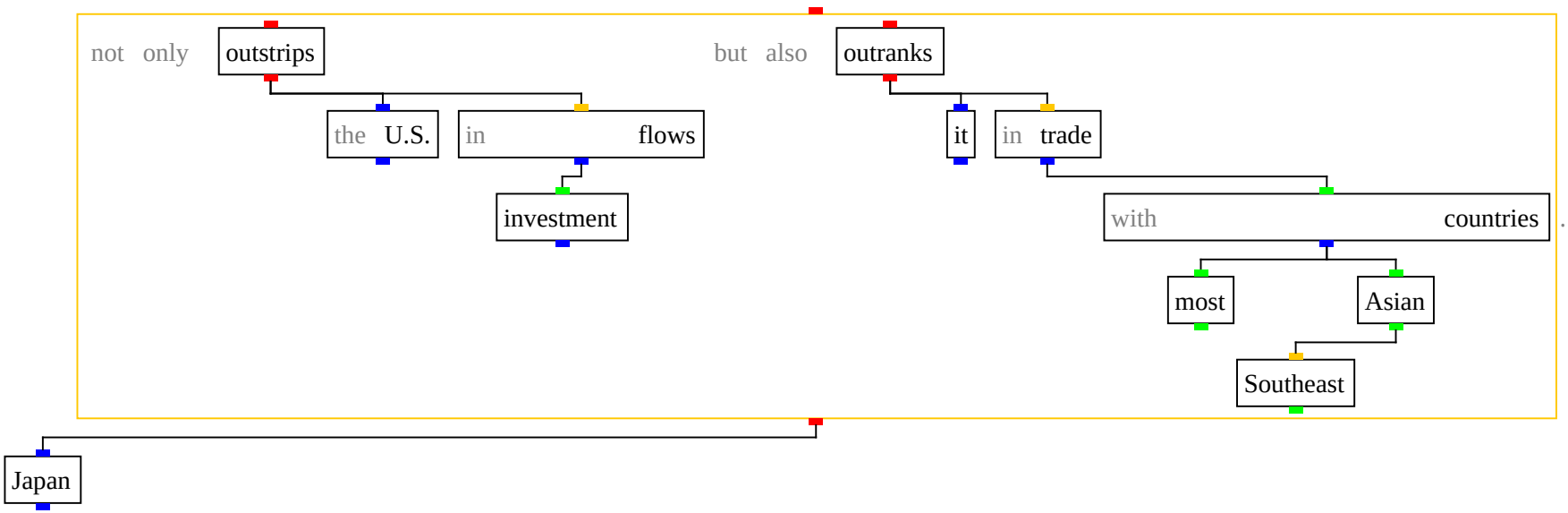


Penn WSJ sec00, sentence # 666



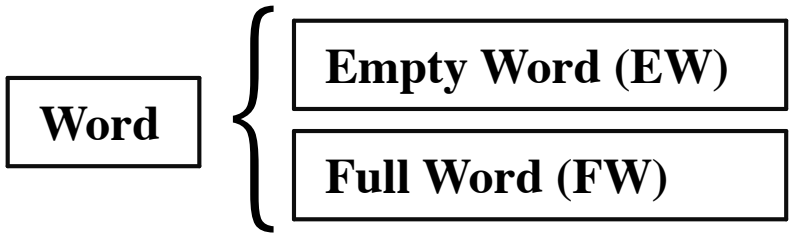


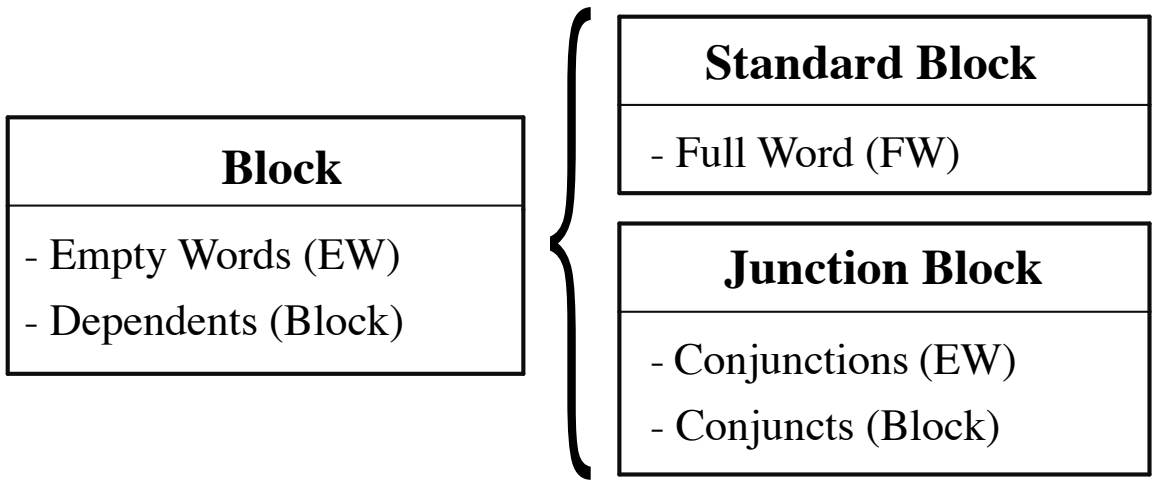
Penn WSJ sec00, sentence # 666



Converting the PTB

Elements of a TDS

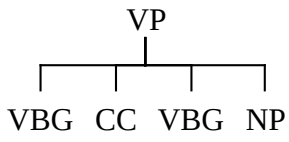


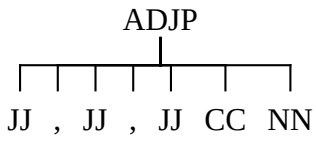


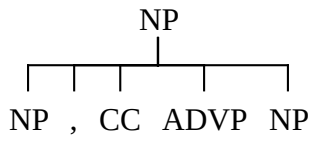
The conversion algorithm

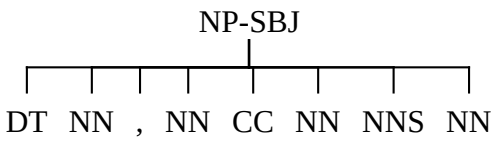
Algorithm: *Convert*(N_{PS})
Input: A node N_{PS} of a PS tree
Output: A block N_{TDS} of a TDS tree
begin
  instantiate N_{TDS} as a generic block
  if N_{PS} *is a junction* **then**
    instantiate N_{TDS} as a junction block
    foreach *node D in children of* N_{PS} **do**
      if *D is a conjunct* **then**
        $D_{TDS} \leftarrow Convert(D)$
        add D_{TDS} as a conjunct block in N_{TDS}
      else
        $D_{lex} \leftarrow$ lexical yield of D
        if D_{lex} *is a conjunction* **then**
          add D_{lex} as a conjunction in N_{TDS}
        else
          add D_{lex} as empty word(s) in N_{TDS}
  else
    $N_h \leftarrow$ head daughter node of N_{PS}
    if N_h *yield only one word* w_h **then** instantiate N_{TDS} as a standard block with w_h as its full word
    else $N_{TDS} \leftarrow Convert(N_h)$
    foreach *node D in children of* N_{PS} **do**
      if $D == N_h$ **then continue**
      $D_{lex} \leftarrow$ lexical yield of D
      if D_{lex} *are only empty words* **then**
        add D_{lex} as empty word(s) in N_{TDS}
      else
        $D_{TDS} \leftarrow Convert(D)$
        add D_{TDS} as a dependent of N_{TDS}
  return N_{TDS}
end

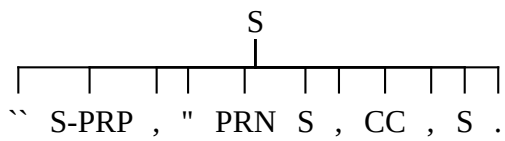
Detecting Conjuncts in Junction Structures

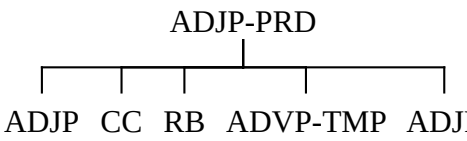


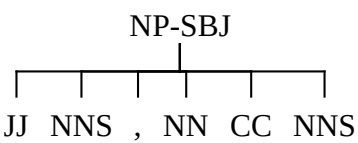


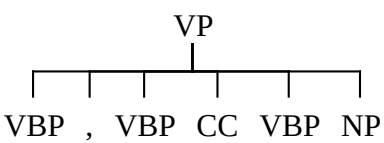


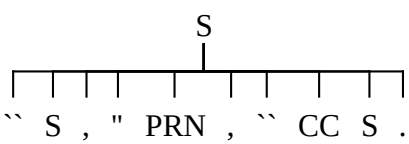


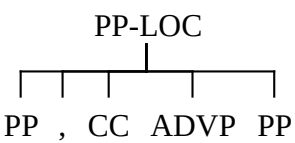


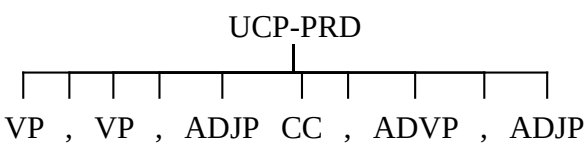


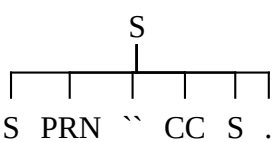


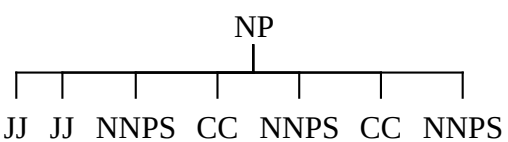


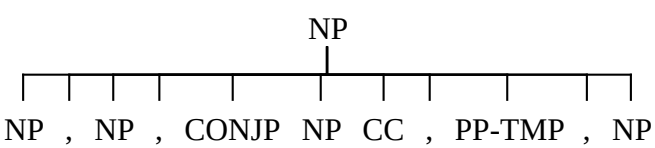


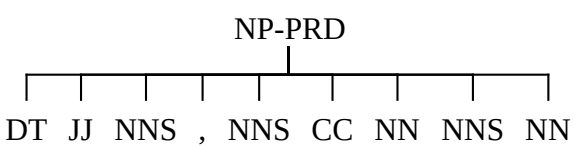


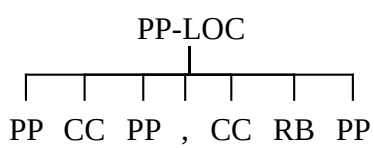


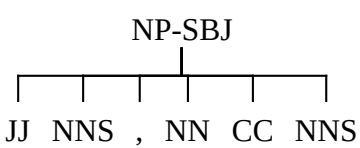


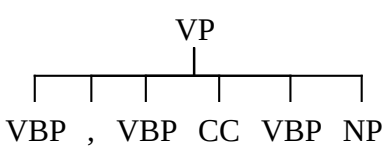


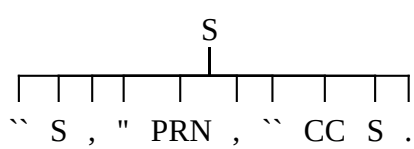


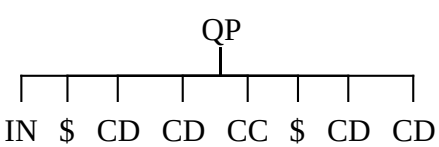












Error Analysis

Qualitative analysis on section 00, mistakes in:

- **Coordinated structures**: blocks that are dependent on the whole coordination are wrongly identified as conjuncts or dependents of one of the conjunct.
- Wrongly assigned **categories**.

Possible **inadequacies** of the TDS formalism:

- Junction between empty words
  (e.g. *[on and off] the court*, *[up to and through] the crash*).
- Junction between coordination of compound verbs
  (e.g., *He was [eating and drinking]*). SOLVED!
- Full word as transferrers (e.g. *A forum likely to bring attention*).

Conclusions

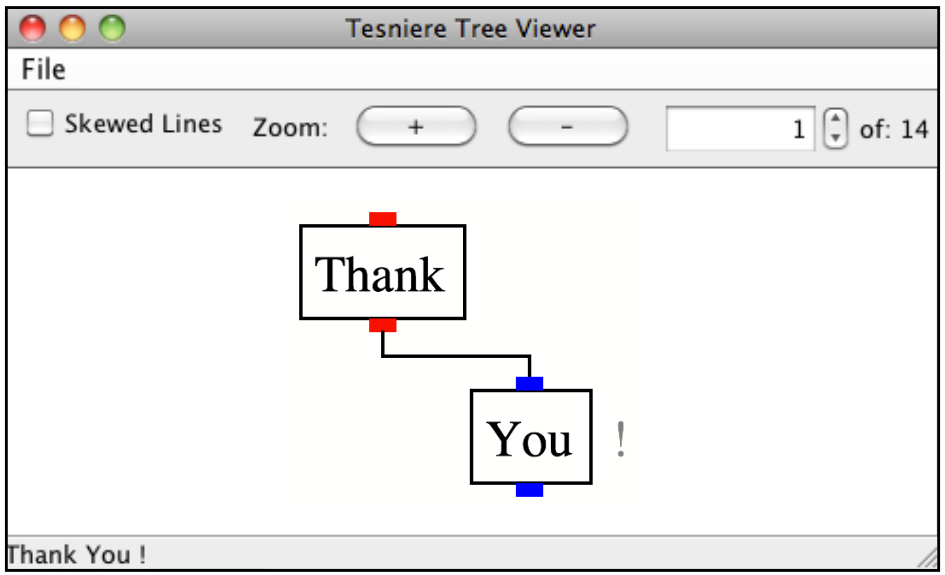
- Ongoing project of converting the Penn WSJ treebank into TDS representation.
- CL has valued the simplicity of DS but inadequate to handle coordination.
- **Junction** is a dedicate operation in TDS to handle junction structures.
- **Blocks**, **transference**, and **categories** further simplify and generalize the model.

Further Work

- Improve on systematic mistakes (we need your help!).
- Currently working on a **probabilistic language model** for parsing and generating TDS structures.
- A good language model needs a good **modeling of coordination**.

Conversion and visualization tool available at:

staff.science.uva.nl/~fsangati/TDS



f.sangati@uva.nl
chiara.mazza@gmail.com