CS114 Lecture 13a

Sparser

March 12, 2014

Professor Meteer

Thanks for Jurafsky & Martin & Prof. Pustejovksy for slides

The Sparser system
Architecture and Operation

David McDonald

August 11th 2009

“Sparser” (sparse + parser)

- Does a full, linguistically principled analysis of the parts of the text it understands.

  understand = represent in its semantic model

- Uses a semantic grammar with integrated syntax and interpretation.
- Creates the rules as part of defining the concepts in the model — the concepts/instances are automatically linked to their grammar rules.
- Efficient algorithm (monotonic, indelible); recycled data structures allow it to run fast and indefinitely.

  ~ 5k words/second, hours at a time

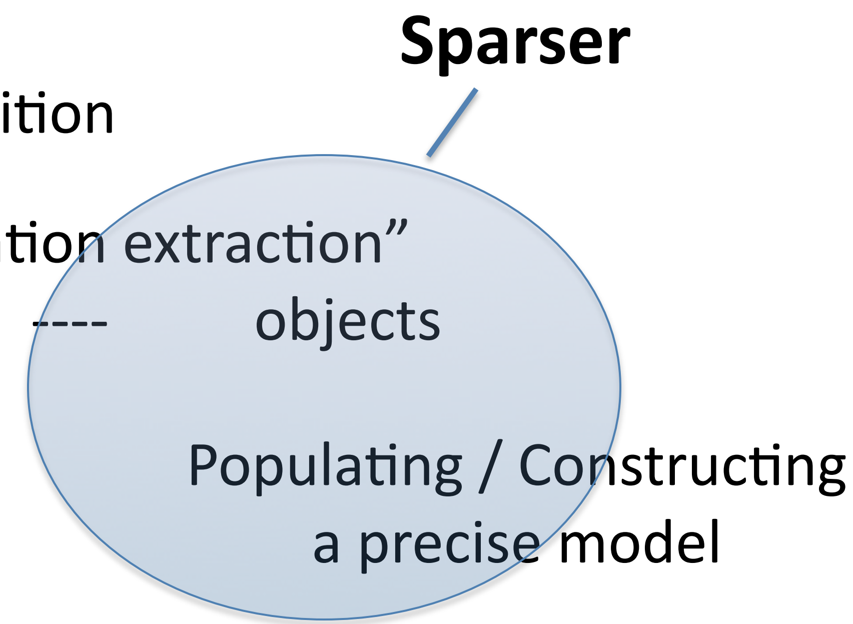
How much can you do with what reliability?

← **shallower techniques**

semantically-informed deeper techniques →

Topic identification

Named-entity recognition



Sparser

“information extraction”

strings ---- objects

Populating / Constructing
a precise model

The task is ‘text to tuples’

person-company-title/event

Economist Newspaper Ltd. (London) --- Pierre Vinken, 61 years old, will join the board as a nonexecutive director Nov. 29. Mr. Vinken is chairman of Elsevier N.V., the Dutch publishing group.

Wall Street Journal, 11/2/89
via the Data Collection Initiative

```
PCT/E
  person: Pierre Vinken
  company: Economist Newspaper
  title: board-member/director
  event: go-to-company
```

```
PCT/E
  person: Pierre Vinken
  company: Elsevier
  title: chairman
  event: at-company
```

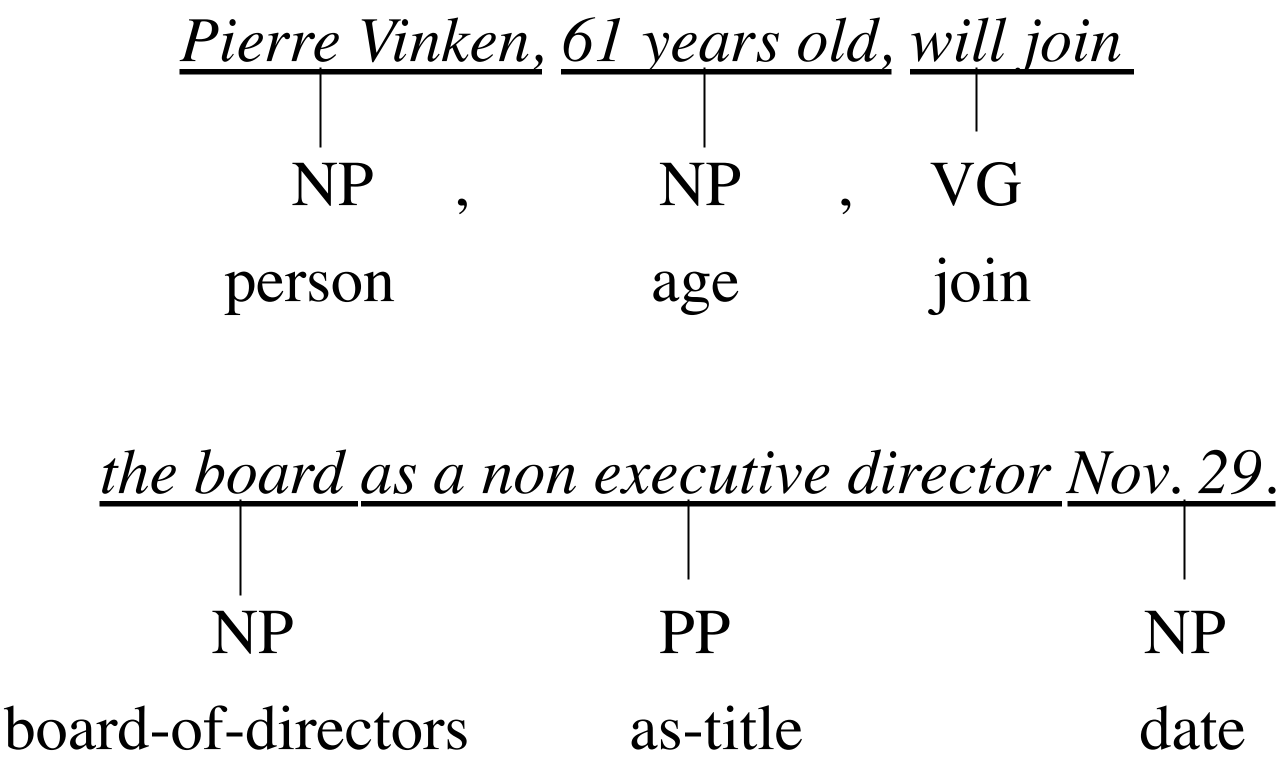
```
#<person p-37>
   name:      →    #<name-of-a-person n-12>
                     first: “Pierre”
                     last: “Vinken”
```

Focus has been
Domain-specific sublanguages

Xxxxxxxxx Xxxxxxxxx Ltd. (London) --- Xxxxxx Vvvvvv, 61 years old, will join the board as a xxxxxxxxxxx director Nov. 29. Mr. Vvvvvv is chairman of Xxxxxxxx N.V., the Dutch xxxxxxxxxx group.

Named entities: people, companies, …
Age
Dates
Who's News: retired, promoted, …
Unknown words / 'Debris Analysis'

All constituents have Semantic Labels



Pierre Vinken, 61 years old, will join

NP , NP , VG

person age join

the board as a non executive director Nov. 29.

NP PP NP

board-of-directors as-title date

101: Interpretation as typed structured objects

Economist Newspaper Ltd. (London) --- Pierre Vinken, 61 years old, will join the board as a nonexecutive director Nov. 29. Mr. Vinken is chairman of Elsevier N.V., the Dutch publishing group.

Wall Street Journal, 11/2/89
via the Data Collection Initiative

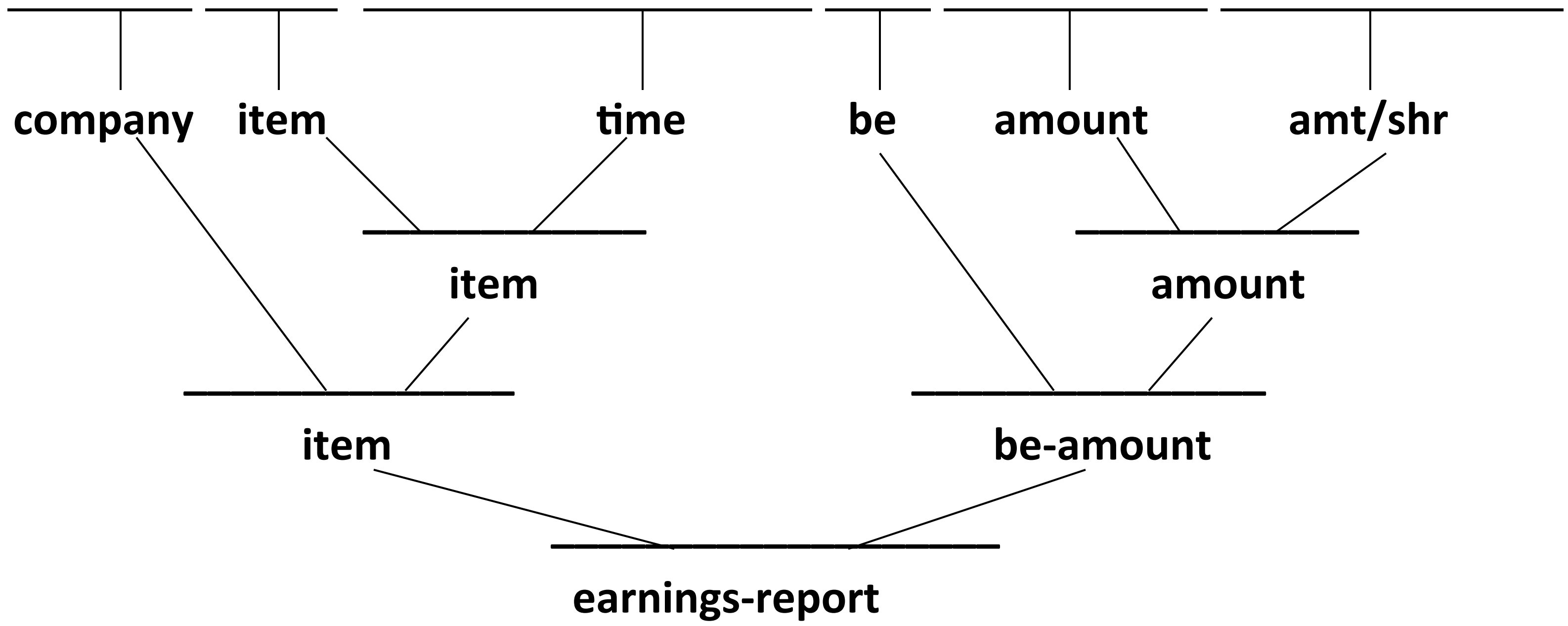
```
PCT/E
  person: Pierre Vinken
  company: Economist Newspaper
  title: board-member/director
  event: go-to-company
```

```
PCT/E
  person: Pierre Vinken
  company: Elsevier
  title: chairman
  event: at-company
```

```
#<person p-37>
  name:    →    #<name-of-a-person n-12>
                  first: “Pierre”
                  last: “Vinken”
```

101: Constituent Structure

“Gateway profits for its second quarter were $122 million or $0.37 a share”



company item time be amount amt/shr

item amount

item be-amount

earnings-report

Labels from Sparser’s semantic grammar of quarterly earnings reports.

Basic Data Structures

- **Words** – print form + case and morphology
- a **Chart** – sequence of numbered **Positions**. Words go between the positions
- **Edges** – span positions. Represent the completion of rules

```
sparser> (d (position# 4))
#<position4 4 "in"> is a
   structure of type position.
   It has these slots:
 array-index          4
 character-index      7
 display-char-index   nil
 token-index          4
 ends-here
     #<edges ending at 4>
 starts-here
     #<edges starting at 4>
 terminal
     #<word "in">
 preceding-whitespace
     #<word one-space>
 capitalization :lower-case
 assessed? :edge-fsas-done
```

Rules: from the Semantic Model

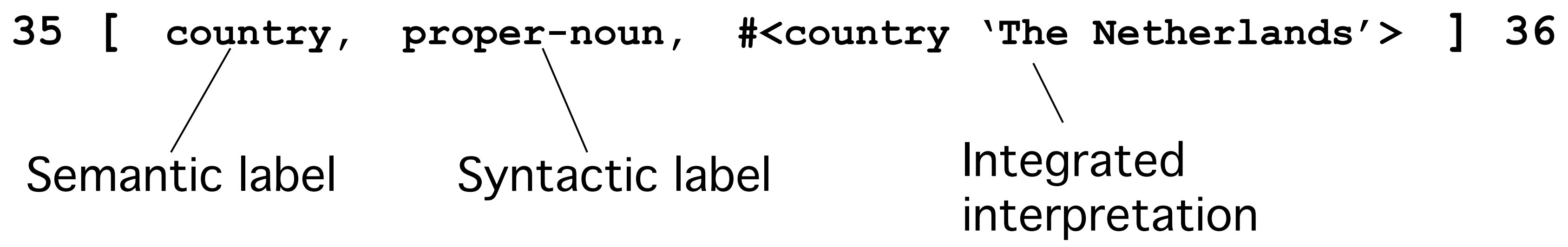
```
(define-country  "The Netherlands"
  :aliases ("Holland")
  :adjective-form ("Dutch"))
```

Concept definition

```
country -> "The" + "Netherlands"
country -> "Holland"
country -> "Dutch"
```

Custom rules written by define-country

"… Elsevier N.V., the Dutch publishing company"



```
35 [  country,  proper-noun,  #<country 'The Netherlands'>  ] 36
```

Semantic label

Syntactic label

Integrated interpretation

Rules: schema-based definition

“sales”, “income”, “profit”, “loss”, …

```
 (define-category  financial
   :specializes  object
   :slots ((name . :primitive word))
   :realization (:common-noun name))

(define-individual financial :name “income”)


(def-cfr financial (“income”)
   :form common-noun
   :referent  #<financial income>)
```

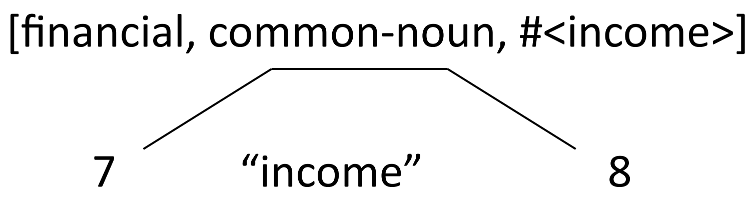
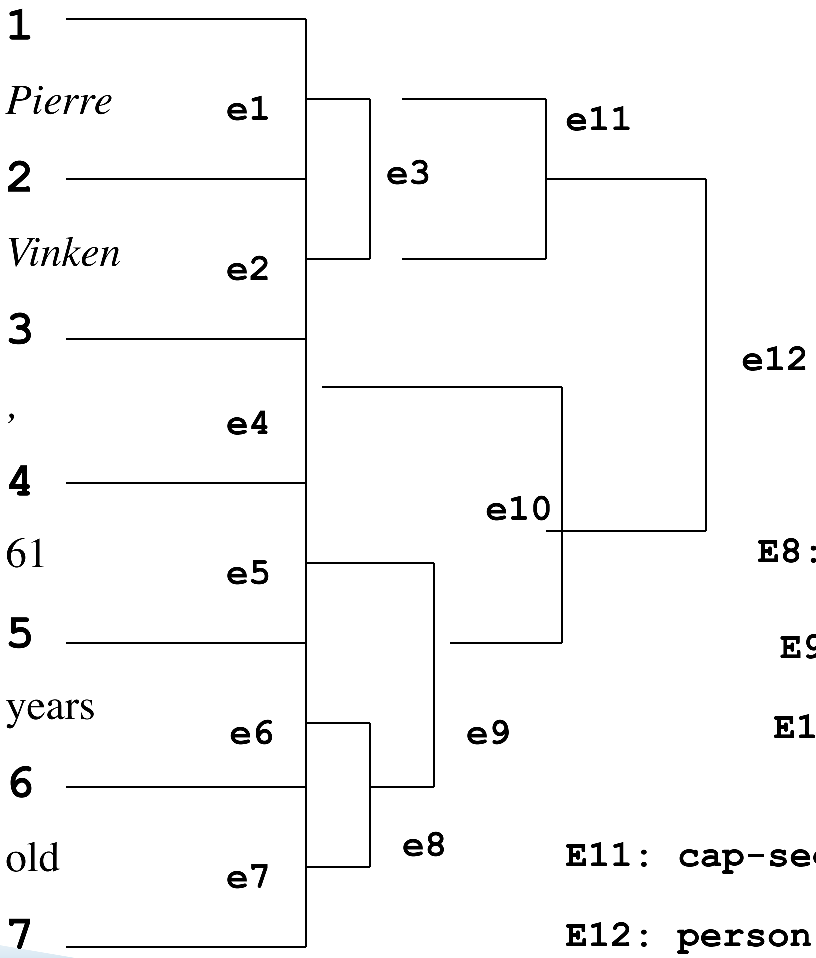


Chart Parsing: span the text with “edges”



E1,E2: capitalized-word

E3: cap-sequence

E4: comma

E5: number

E6: unit-of-time

E7: “old”

E8: age => unit-of-time + “old”

E9: age => number + age

E10: comma-age => comma + age

E11: cap-sequence => person / ___ comma-age

E12: person => person + comma-age

Real parsers have many kinds of "rules"

- Fixed phrases (polywords)
  - "M1A1"
- Finite state analyzers (regex)
  - "617-873-8002" http://alum.mit.org/www/dmcdonald/
- Rewrite-rules
  - Context free (unary, binary, n-ary)
    - "jul" => month, number + "x" => resource-quantity-prefix
  - Context sensitive
    - name => person / military-rank ___
  - Syntactic Form rules
    - "is" + <verb> => <verb>
- Heuristic 'rules'
  - Morphology: "<aaa>ing" +ing => verb "was" / ____

Summary

- Swiss-Army knife of parsers
- “No Presentation without Representation”
  - The best way to use Sparser is to start with a conceptual model (close to the language) and have it write the rules
  - But ad-hoc rules (fsa’s, cfrs, etc.) are often a necessary crutch
- Sublanguages can be parsed with very high precision

Partial Parsing with GATE

- GATE allows you to create a pipeline of NLP processes to run across data
  - Tokenization
  - Gazeteer (name lookup)
  - POS Tagging
  - Morphological analysis
  - FS Rules