

*Uni Konstanz: Machine Language Processing*

# **Dependency Parsing & Applications**

Gerold Schneider

Institute of Computational Linguistics,  
English Department,  
Zürcher Kompetenzzentrum Linguistik  
University of Zurich  
`gschneid@ifi.uzh.ch`

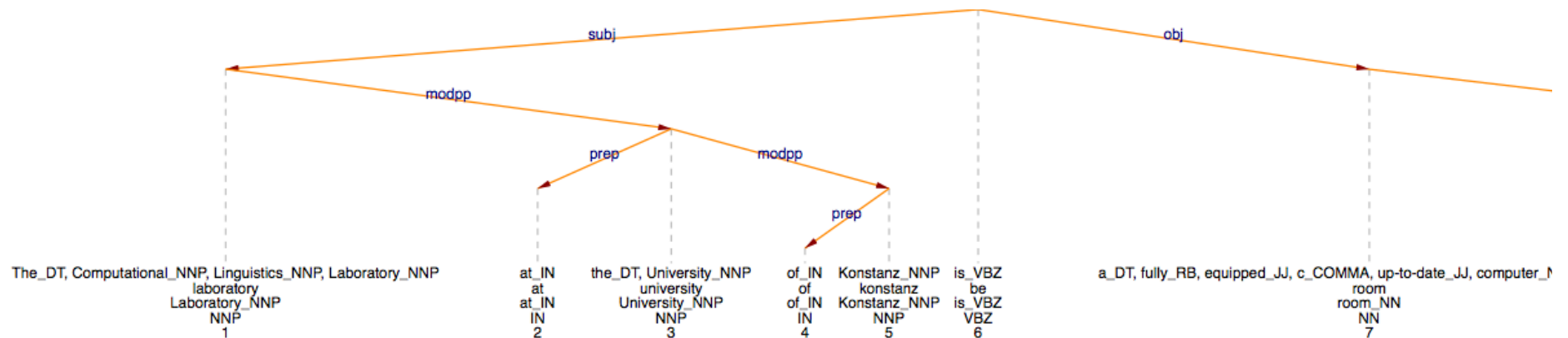
November 2014

## Contents

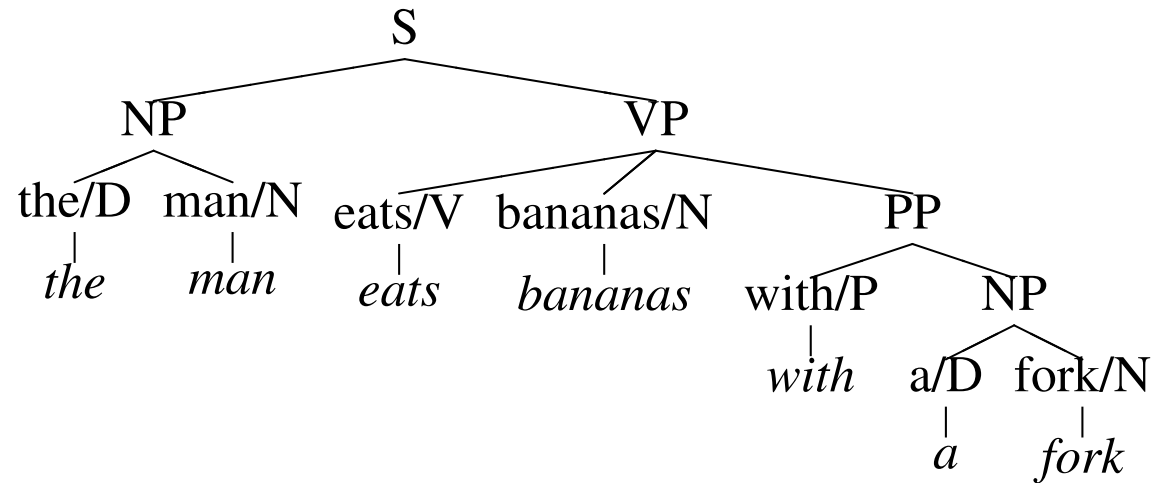
- Dependency Parsing
  1. Dependency Grammar
  2. Parsing Algorithms, in particular CYK
  3. Statistics for Disambiguation: PP attachment
- Selected Applications
  4. Data-Driven Models for Descriptive Linguistics
  5. Parser as Human Processing Model
  6. Syntax and Discourse for Text Mining

# 1 Dependency Grammar

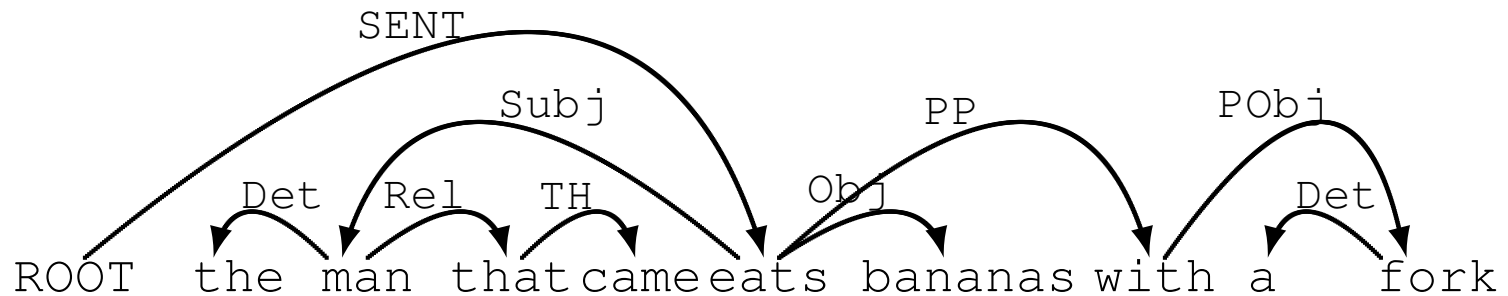
1. DG as Extended Valency Grammar
2. DG as Government Grammar
3. DG as Terminal Node Context-Free Grammar
4. DG as a Version of X-bar Theory
5. Long-distance Dependencies



Constituency focuses on what a phrase consists of



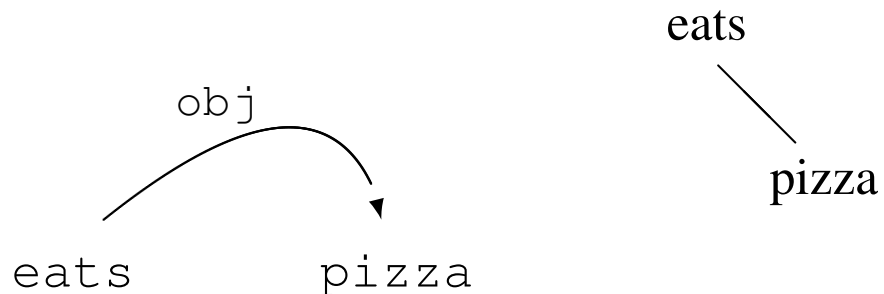
DG (Tesnière, 1959) focuses on the dependencies between words



## 1.1 DG as Extended Valency Grammar

- Intuitively, DG is a valency grammar. The term *valency* is borrowed from chemistry. Different verbs take a specific number and type of complement. Valencies also place selectoinal restrictions.
- Valency is not restricted to verbs only. Relational nouns and some predicative adjectives also open valencies.
  - (1) I am afraid of action.
  - (2) \*I am afraid for action.
  - (3) \*I am ready of action.
  - (4) I am ready for action.
- Valencies have been extended from verbs to many other word classes, and from syntax into semantics (Helbig, 1992, 108).
- Valency theory was also influenced by (Fillmore, 1968)'s Case Grammar and by collocation analysis.

- The word opening a valency is defined as *governor* or equivalently *head* in DG, the word filling the valency is called *dependent*. DG leaves the distinction between mother node, governor and head underspecified.
- Arrow notation vs. Stemma notation



- Valency was extended from argument to adjuncts and even to function words (to build up complete dependency structures). Some dependencies are not strictly valencies or grammatical functions (e.g. subordinate verb – complementizer).
- Tesnière's conception of *nucleus* is used to alleviate the need to create too many dependencies for which valency cannot account. A nucleus is a content word plus its attributed function words.
- Only nuclei have dependency relations among each other. For a verb, typical function words are auxiliaries. For a noun, typical function words are determiners.
- Valency background of DG is a major reason why the definition of heads in DG is often diametrically opposed to the definition of heads in GB or in Montague grammar.

## 1.2 Government Grammar

DG leaves the distinction between governor and head underspecified.

Government, a relatively complex constituent relation in GB, is a DG primitive.

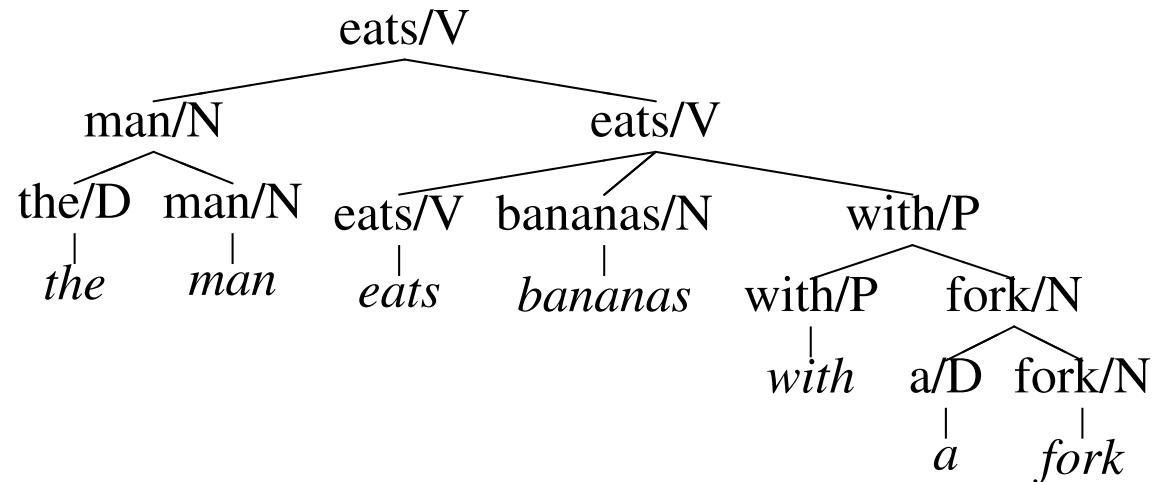
Covington (1992, 4) concludes that since only lexical items can govern in DG, immediate dependency and government coincide.

... it is clear that if only lexical items can govern, then the definition of government is: **A governs B iff B is an immediate dependent of A.** One can hardly ask for this to be simpler. (Covington, 1992, 4)



### 1.3 Terminal Node Context-Free Grammar

DG is strictly lexicalist, non-terminal nodes only exist as a derived concept, *endocentricity* is naturally enforced  $\rightarrow$  (Chomsky, 1995): Bare Phrase Structure



The *head* of a phrase and its *projection* are isomorphic.

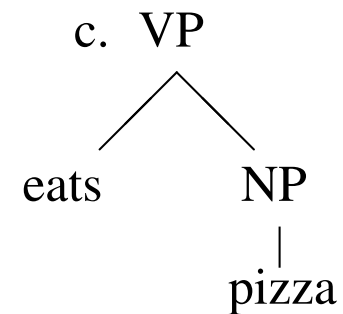
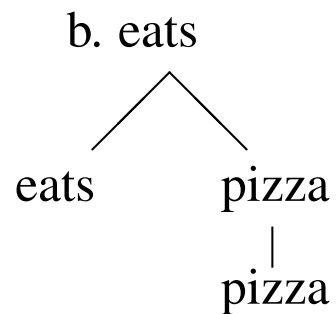
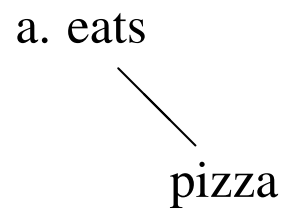
---

DG is a CFG that only knows terminal nodes. Stemma DG notation

(a.) stemma DG notation

(b.) redundant stemma DG notation (one daughter and the head are identical)

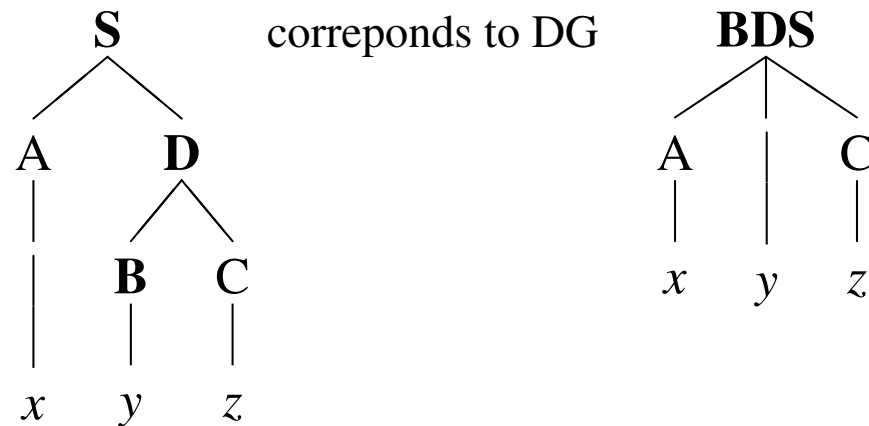
(c.) constituency representation in which the governor is a phrasal category.



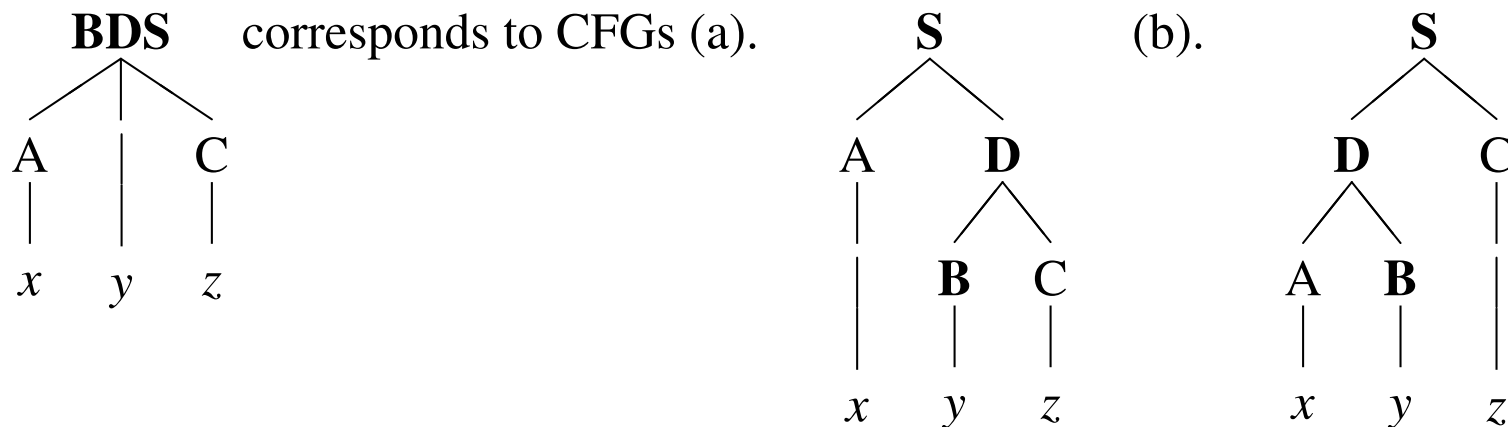
Is dependency more or less expressive than constituency?

- potentially more: dependencies are typically labelled
- potentially less: mapping of constituency to DG trees:  
Head projection line *BDS*: B,D, and S are all equivalent in DG

(5)



One can map a headed tree to a unique dependency tree. The projection dependency tree of a headed tree is unique, but several headed trees may have the same projection dependency tree.

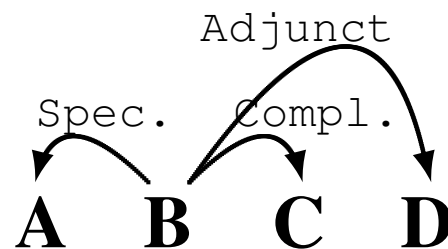


Projection-dependency trees abstract away from the order in which dependents are combined with their governor.

Such an unlabelled DG fails to express bar-level distinctions (X-bar).

## 1.4 DG as a Version of X-Bar Theory

X-bar theory uses three types of dependencies: *specifier*, the non-head dependent of  $X''$ ; *adjunct*, a non-head dependent of  $X'$  with  $X'$  as sister; and *argument*, a non-head dependent of  $X'$  with  $X^0$  as sister.



If one uses a *labelled* DG that knows these three types or can map to them unambiguously, then DG and X-bar are equivalent (Covington, 1994).

## 1.5 Dependency Labels

A list of important DG labels looks as follows (my Pro3Gres set, (Schneider, 2008)).

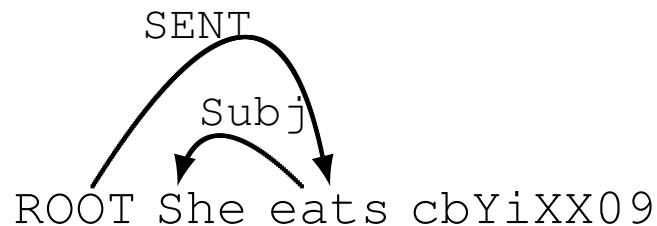
RELATION	LABEL	EXAMPLE
verb–subject	<i>subj</i>	<i>he sleeps</i>
verb–direct object	<i>obj</i>	<i>sees it</i>
verb–second object	<i>obj2</i>	<i>gave (her) kisses</i>
verb–adjunct	<i>adj</i>	<i>ate yesterday</i>
verb–subord. clause	<i>sentobj</i>	<i>saw (they) came</i>
verb–pred. adjective	<i>predadj</i>	<i>is ready</i>
verb–prep. phrase	<i>pobj</i>	<i>slept in bed</i>
noun–prep. phrase	<i>modpp</i>	<i>draft of paper</i>
noun–participle	<i>modpart</i>	<i>report written</i>
verb–complementizer	<i>compl</i>	<i>to eat apples</i>
noun–preposition	<i>prep</i>	<i>to the house</i>

Influential label sets:

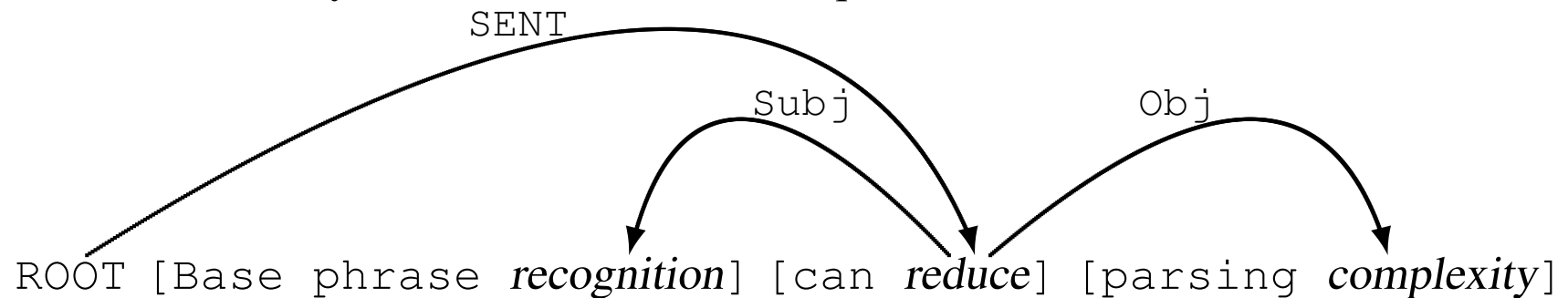
- GREVAL (Carroll, Minnen, and Briscoe, 2003): similar to my set, can be mapped.
- Stanford scheme (de Marneffe and Manning, 2008): similar to my set, based on GREVAL. Takes up apposition relation.
- CONLL set (Nivre et al., 2007): influential, easy to map from Penn Treebank as described in (Nivre, 2006), but relatively surface-oriented (e.g. main as dependent of aux).
- Constraint-Dependenz-Grammatik (Foth, 2005). The most popular German label set. Used e.g. by ParZu (Pro3GresDE) (Sennrich et al., 2009)
- Universal Dependencies (<http://universaldependencies.github.io/docs/>): current research endeavour to combine Google universal tags and Stanford scheme, for as many languages as possible.

## 1.6 Robustness

- Isomorphism of Words and Projections  $\longrightarrow$  Building the max. projection always succeeds



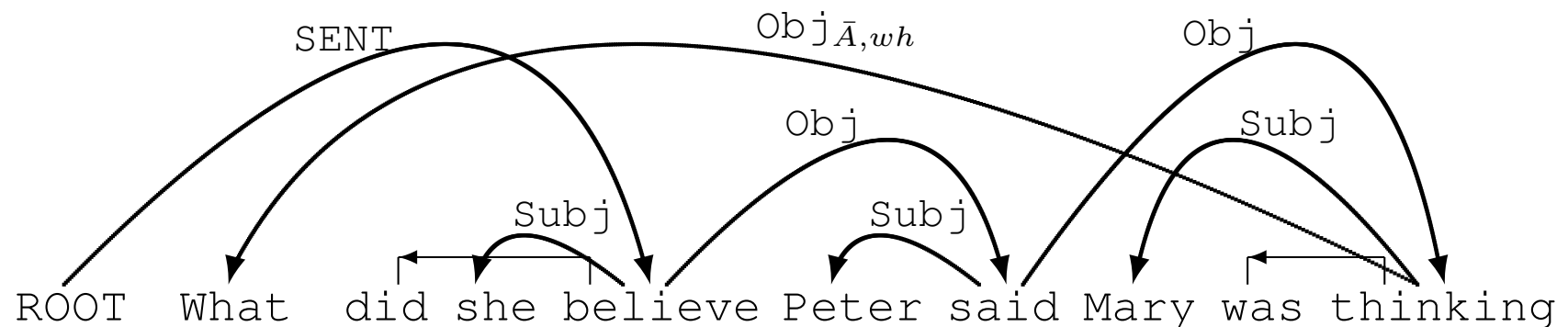
- A chunker with head-extraction offers the same head/phrase isomorphism as Tesnière (Abney, 1995)  $\longrightarrow$  *divide & conquer*





## 1.7 Long-Distance Dependencies (LDD)

... where context-free grammar ends

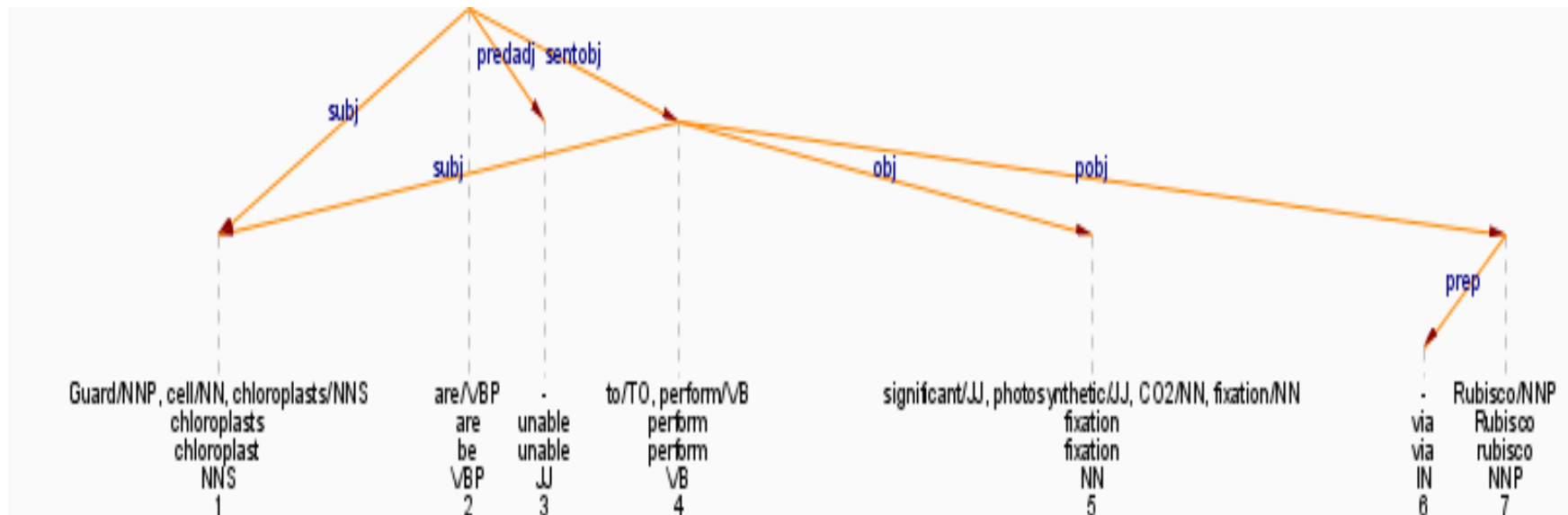


There are considerably fewer LDDs in dependency grammar than in constituency, as there is no distinction between inner and outer arguments (every clausal element attaches to the main verb), and as the direction (linear precedence) can be underspecified.

---

Most remaining LDDs are either or several of

- secondary links, e.g. control structures which can be added with patterns



- AUX-MAIN in the verb nucleus (HPSG argument composition)
- easy to spot (wh, sentence-initial,  $\bar{A}$ -movement)
- easy to treat with a SLASH-feature (e.g. HPSG) or extra stack

## 2 Parsing Algorithms, in particular CYK

### 2.1 Top-Down Algorithms

s --> np, vp.

pp --> p, np.

vp --> v, np.

vp --> vp, pp.

np --> [astronomers].

np --> [stars].

np --> [planets].

np --> [telescopes].

np --> [ears].

np --> np, pp.

p --> [with].

v --> [saw].

v --> [sleep].

The Top-down algorithm *recursive descent* naturally follows from CFG rewrite rules.

This is an actual Prolog DCG.

The algorithm is simple, but it is target-driven and not robust.

It is also inefficient with ambiguous grammars, in the same ways as shift-reduce is.

## 2.2 Shift-Reduce

Shift-reduce is a very simple algorithm with

- two stacks:  
input buffer (text to read) and reduce stack (where CFG rules are applied)
- two operations:  
SHIFT: move from input buffer to stack REDUCE: apply a grammar rule

The algorithm for an unambiguous grammar in pseudo-code:

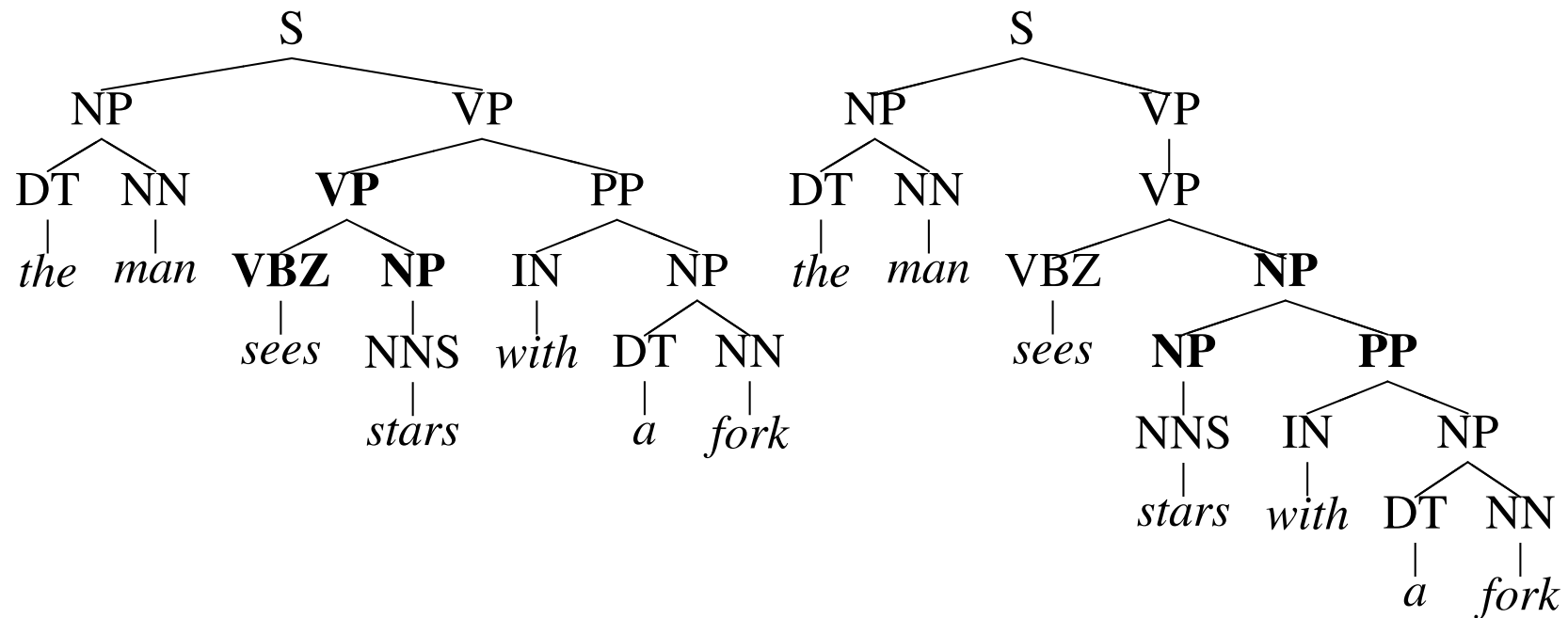
```
REPEAT UNTIL (only 1 head on reduce-stack)  $\wedge$  (input buffer empty) {  
    IF (grammar rule for 2 topmost words/tags on reduce-stack) THEN  
        {REDUCE}  
    ELSE  
        {SHIFT}  
}
```

State	Action	Reduce-Stack	Input Buffer	Applied Rule
1	Start	[]	[the_DT,dog_NN,eats_VBZ,a_DT,bone_NN]	
2	Shift	[the_DT]	[dog_NN,eats_VBZ,a_DT,bone_NN]	
3	Shift	[the_DT,dog_NN]	[eats_VBZ,a_DT,bone_NN]	
4	Reduce	[NP]	[eats_VBZ,a_DT,bone_NN]	NP → DT NN
5	Shift	[NP,eats_VBZ]	[a_DT,bone_NN]	
6	Shift	[NP,eats_VBZ,a_DT]	[bone_NN]	
7	Shift	[NP,eats_VBZs,a_DT,bone_NN]	[]	
8	Reduce	[NP,eats_VBZ,NP]	[]	NP → DT NN
9	Reduce	[NP,VP]	[]	VP → VBZ NP
10	Reduce	[S]	[]	S → NP VP

Table 1: Actions and Data Structures in a Shift-Reduce derivation

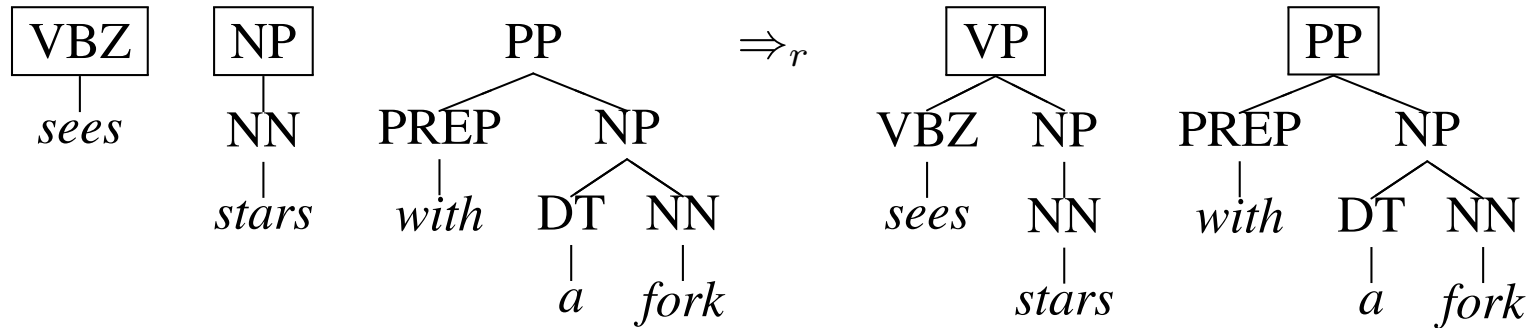
## 2.3 Treatment of Ambiguity

Problem: CFGs, and natural language, are ambiguous. E.g. PP-attachment:

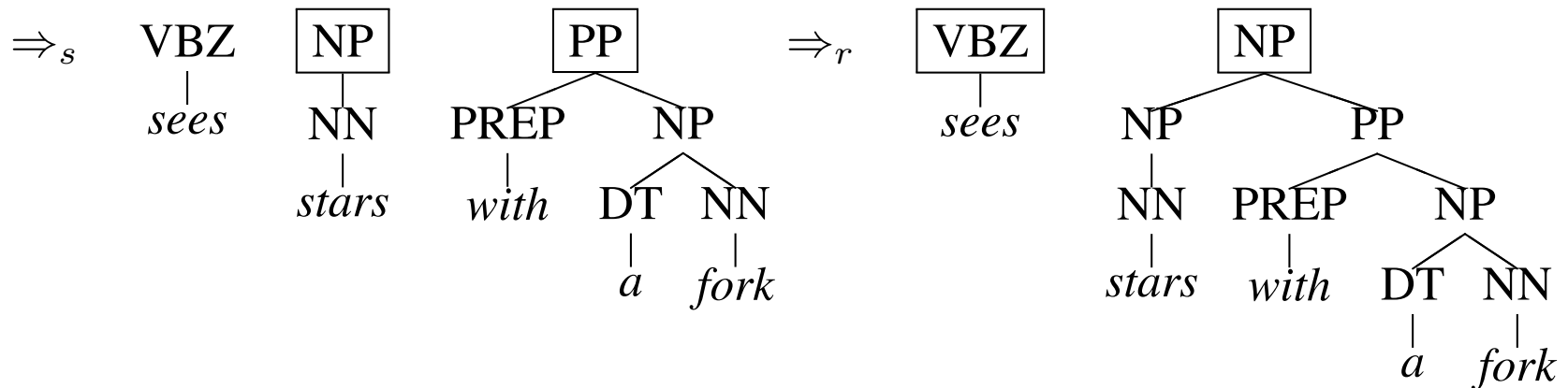


The items on the reduce stack are boxed.  $\Rightarrow_s$  = shift;  $\Rightarrow_r$  = reduce.

- Reduce has precedence  $\rightarrow$  *NP:stars* is attached to *sees\_VBZ*, and later *PP:with a fork* is also attached to *VP:sees*.



- Shift has precedence  $\rightarrow$  *PP:with a fork* is attached to *NP:stars*, and later *stars* is attached to *sees\_VBZ*.



So the shift-reduce code actually is:

```
REPEAT UNTIL (only 1 head on reduce-stack)  $\wedge$  (input buffer empty) {  
  IF (grammar rule for 2 topmost words/tags on reduce-stack) THEN  
    {REDUCE}  
  IF  
    {SHIFT}  
}
```

- Every ambiguity doubles the search space.
- Partial structures that were already calculated are not kept (e.g. the PP itself)  $\rightarrow O(2^n)$
- Depth-first search



## 2.4 CYK

### **CYK (Cocke, Younger, Kasami) informally**

... is a breadth-first, parallelized version of shift-reduce

- Let's first find the shortest possible **combinations**, i.e. of length  $j = 2 = 1 + 1$ , then in the next step  $j = 3$ , etc.
- we traverse the sentence from left to right: positions: 1+2, 2+3, 3+4, etc.
- if we can reduce something, we store it in a chart for later use.

**CYK Characteristics: 3 loops**  $\rightarrow O(n^3)$

- Structure length is monotonically increasing: every result of a combination is longer than the combined elements

This means that if we start at the shortest edges (the lexical items) and at every repeated step search for edges that are 1 word longer we never have to backtrack.

for  $j = 2$  to  $n$  # length of edge

- An edge can start anywhere, at the latest at  $n - j$ .

for  $i = 1$  to  $n - j + 1$  # beginning of edge

- Binary non-empty rules mean that every edge  $i..(i + j)$  can be separated in  $j - 1$  ways, e.g. 1..4:

1 2 3 4  $\rightarrow$  1 | 2 3 4  
                   1 2 | 3 4  
                   1 2 3 | 4

for  $k = i + 1$  to  $i + j - 1$  # separator position

---

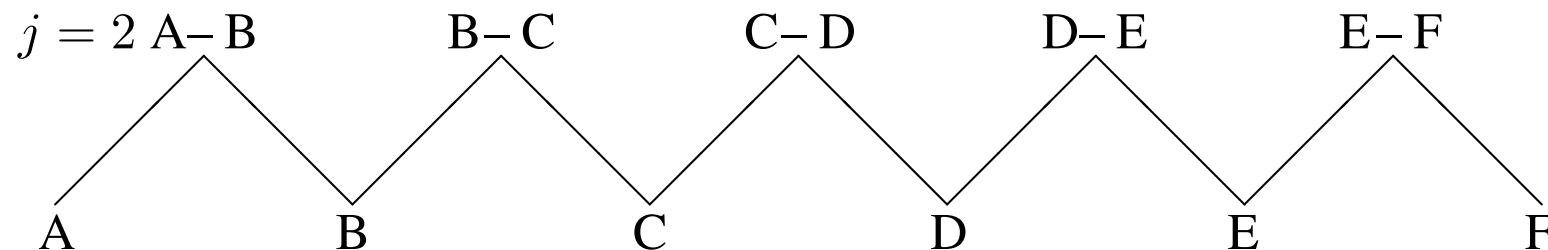
## The CYK algorithm, step 1: $j=2$

- CYK Parsing: bottom-up parallel processing,  $c = \text{chart}$

```

for  $j = 2$  to  $N$            # length of span
  for  $i = 1$  to  $N - j + 1$    # begin of span
    for  $k = i + 1$  to  $i + j - 1$  # separator position
      if  $Z \rightarrow XY$  and  $X \in c[i \text{ TO } k], Y \in c[(k + 1) \text{ TO } (i + j)]$ 
        and  $Z \notin c[i \text{ TO } (i + j)]$ 
        then insert  $Z$  at  $c[i \text{ TO } (i + j)]$ 
  
```

$i$  increases  $\longrightarrow$

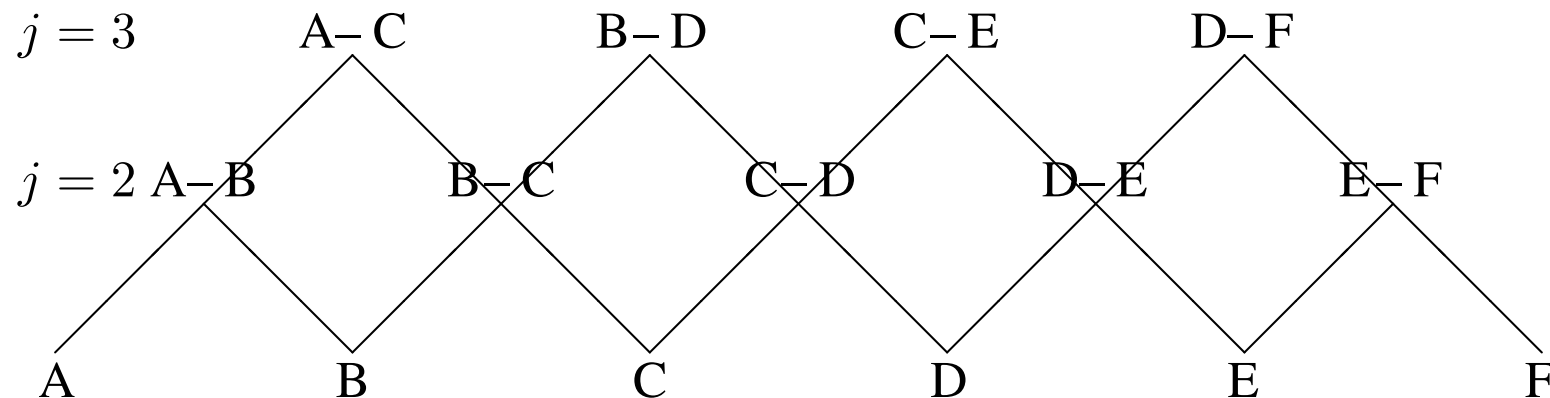


### The CYK algorithm, step 2: $j=3$

- CYK Parsing: bottom-up parallel processing,  $c = \text{chart}$

```

for  $j = 2$  to  $N$            # length of span
  for  $i = 1$  to  $N - j + 1$    # begin of span
    for  $k = i + 1$  to  $i + j - 1$  # separator position
      if  $Z \rightarrow XY$  and  $X \in c[i \text{ TO } k], Y \in c[(k + 1) \text{ TO } (i + j)]$ 
        and  $Z \notin c[i \text{ TO } (i + j)]$ 
        then insert  $Z$  at  $c[i \text{ TO } (i + j)]$ 
  
```

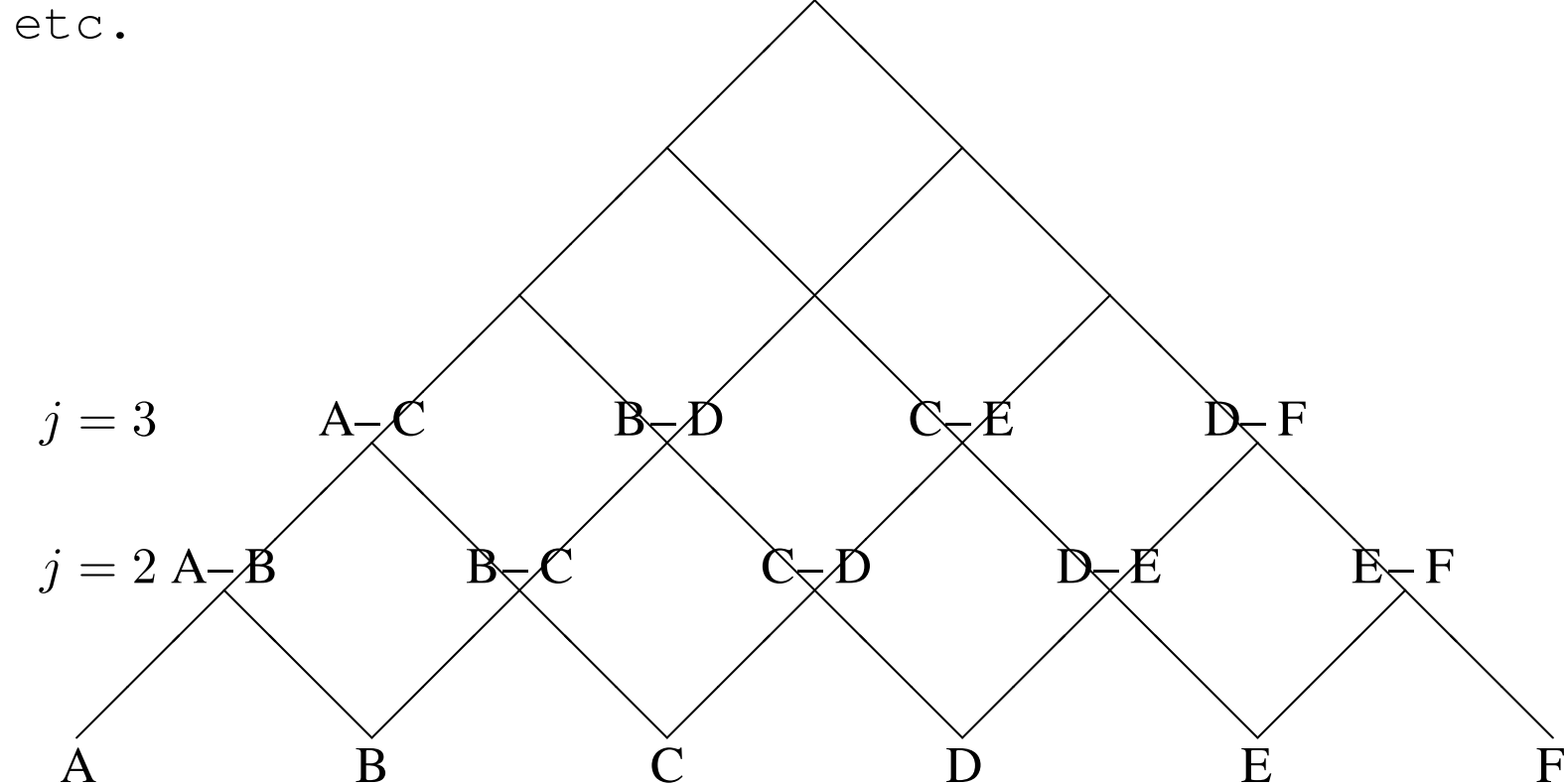


## The CYK algorithm, continued

- CYK Parsing: The analysis matrix

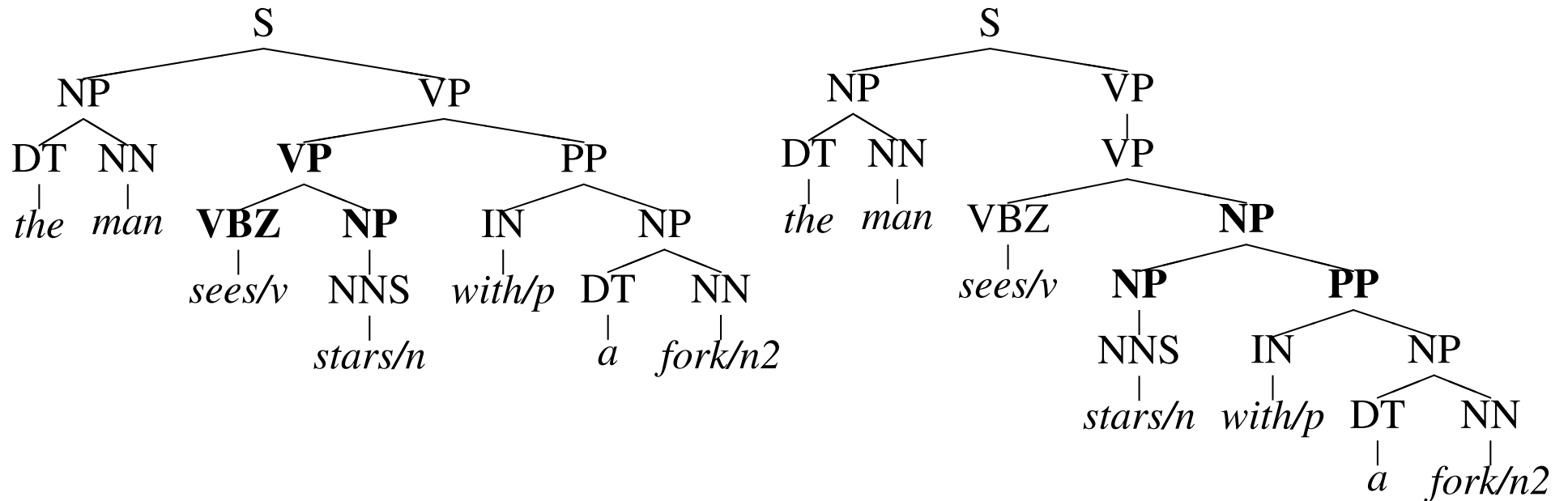
for  $j = 2$  to  $N$

# length of span



### 3 Ambiguity: the example of PP-attachment

The problem



Collins and Brooks (1995) train on the Penn Treebank.

- addresses PP-attachment, but can be used for other ambiguities (Collins, 1999)
- 1 mio. words is relatively small, dealing with unseen combinations is essential:

they use back-off. E.g. for verb-attachment (VPP):

$$p(VPP|v, n, p, n2) = \frac{f(VPP, v, n, p, n2)}{f(v, n, p, n2)} \quad \text{if } > 0, \text{ else} \quad (1)$$

$$p(VPP|v, n, p, n2) = \frac{f(VPP, v, n, p) + f(VPP, v, p, n2) + f(VPP, n, p, n2)}{f(v, n, p) + f(v, p, n2) + f(n, p, n2)} \quad \text{if } > 0, \text{ else} \quad (2)$$

$$p(VPP|v, n, p, n2) = \frac{f(VPP, v, p) + f(VPP, p, n2) + f(VPP, n, p)}{f(v, p) + f(p, n2) + f(n, p)} \quad \text{if } > 0, \text{ else} \quad (3)$$

$$p(VPP|v, n, p, n2) = \frac{f(VPP, p)}{f(p)} \quad \text{if } > 0, \text{ else } P(VPP = 0) \quad (4)$$

**Parser Score** for a sentence = Summed probabilities over all parsing steps for an entire derivation. Each parsing step is an attachment decision  $p(Rel|words)$  leading to a node in the ensuing syntax tree.

## 4 Data-Driven Models for Descriptive Linguistics

### 4.1 Regional syntactic innovations of Indian English

Regional differences are very subtle and intricate.

[D]istinctive phenomena tend to concentrate at the interface between grammar and lexicon, concerning structural preferences of certain words (like the **complementation patterns that verbs allow**), **co-occurrence and collocational** tendencies of words in phrases, and also patterns of word formation.

(Schneider, 2004, 229, boldface added)

- I. Verb Formation: ditransitive complementation is particularly restricted and depends on verb semantics. See Mukherjee (2009) for Indian English
- II. Particularly frequent word-sequences (lexical bundles) can be detected by using statistical distribution measures such as mutual information, Z-score or Observed/Expected (O/E). We compare O/E from ICE-India divided by O/E from British English (BNC).
- III. We also investigated which frequent ICE-India trigrams are absent in the BNC.



I. Double object verb counts:

I. ICE-India	Count	ICE-GB	Count
give	89	give	104
send	37	send	27
<b>provide</b>	12	offer	12
offer	9	tell	10
<b>grant</b>	6	call	8
show	6	do	8
call	4	show	7
<b>develop</b>	4	cost	6
<b>hand</b>	4	pay	6
pay	4	bring	5
bring	3	ask	4
do	3	allow	3
owe	3	earn	3
ask	2	teach	3
consider	2	consider	2
deny	2	deliver	2
earn	2	find	2
find	2	grant	2
promise	2	hand	2
tell	2	promise	2

## Dependency Parsing & Applications

II. O/E ratio	Trigram	O(BNC)	O(ICE-India)
1575	this_DT court_NN that_IN	3	21
975	the_DT blood_NN group_NN	3	13
810	do_VBP not_RB recollect_VB	5	18
750	the_DT household_NN sector_NN	3	10
731.25	as_RB to_TO why_WRB	4	13
675	statement_NN before_IN the_DT	4	12
675	state_NN government_NN has_VBZ	3	9
675	is_VBZ known_VBN as_RB	3	9
675	in_IN the_DT hostel_NN	8	24
630	proviso_NN to_TO section_NN	5	14
610.7	the_DT best_JJS feature_NN	7	19
600	were_VBD produced_VBN with_IN	3	8
600	the_DT twentieth_NN of_IN	3	8
600	the_DT election_NN commission_NN	3	8
600	submitted_VBD a_DT memorandum_NN	3	8
562.5	in_IN the_DT in_IN	6	15
534.3	a_DT very_RB very_RB	8	19
525	things_NNS are_VBP there_RB	3	7
525	over_IN medium_JJ heat_NN	6	14
525	not_RB to_TO venture_NN	3	7
506.25	on_IN and_CC so_RB	4	9
506.25	both_CC the_DT parties_NNS	4	9
487.5	the_DT rain_NN water_NN	6	13
487.5	<b>for_IN number_NN of_IN</b>	6	13

III. Trigram Absent in BNC	O(ICE-India)
now_RB a_DT days_NNS	42
special_JJ P_NN P_NN	35
canvassed_VBN before_IN this_DT	32
statement_NN was_VBD recorded_VBN	28
learned_VBN special_JJ P_NN	28
<b>is_VBZ called_VBN as_IN</b>	27
scene_NN of_IN offence_NN	26
the_DT honourable_JJ minister_NN	23
for_IN grain_NN yield_NN	22
the_DT learned_VBN special_JJ	21
in_IN the_DT cyclone_NN	19
delay_NN in_IN reply_NN	18
best_JJS feature_NN film_NN	18
avoid_VB delay_NN in_IN	18
small_JJ circle_NN to_TO	17
of_IN solid_JJ wastes_NNS	17
general_JJ body_NN meeting_NN	17
evidence_NN of_IN P_NN	17
feature_NN film_NN in_IN	16
crores_NNS of_IN rupees_NNS	16
in_IN the_DT nodules_NNS	15
has_VBZ also_RB canvassed_VBN	15

I. Ditransitive Complementation example:

(6) I am enclosing herewith a detailed resume of my professional career and feel that I can *provide you the best possible services* in the areas required. (ICE-India W1b-024)

II. The majority of the hits in II. the O/E ratio table arise from text selection criteria, e.g. many legal texts in ICE-India (*proviso to section, statement before the*), many medical texts (*the blood group*), and the spoken data percentage is larger, showing hesitations etc. (*a very very, in the in*). But we also see zero articles (*for number of*)

(7) And *for number of* years following the Nehruvian outlook this society has built itself. (ICE-India S1b-054)

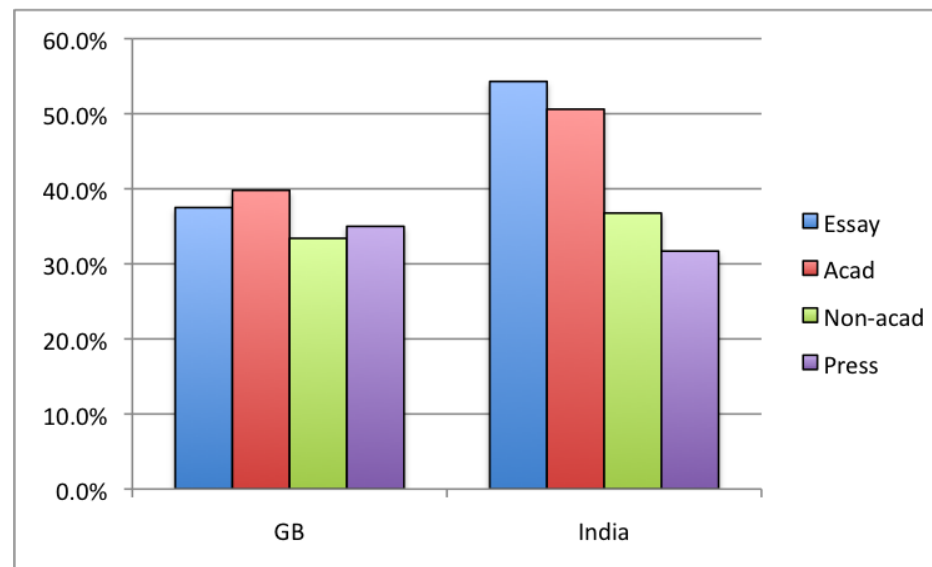
III. Besides text selection, Indian features like archaic spellings (*now a days*), formal language (*the honourable minister*), unusual verb complementation with prepositional phrases (*is called as*) appear in (3) the list of BNC absence.

(8) A substance which is helping in chemical reaction *is called as* a reagent. (ICE-India S1b-004)

Ad II. Zero Articles (see e.g. (Sand, 2004)):

We have tested a large subset, consisting of two thirds of the written part of the ICE corpora. In ICE-GB, 10,034 of the 27,360 singular common nouns, or 36.7%, have no article. In ICE-India, 12,633 of the 29,032 singular common nouns, or 43.5% have no article. The difference is statistically highly significant (chi-square contingency test,  $p < 0.01\%$ ).

Zero articles by genre: while the percentage is spread quite homogenously across genres in ICE-GB, ICE-India shows a peak in the least edited genre, student essays, and a tendency towards over-correction in the most edited genre, press.



## Ad III. Verb Complementation, verb-PP (see e.g. (Sedlatschek, 2009))

$$O/E \text{ ratio} = \frac{O/E(\text{India})}{O/E(\text{BNC})} = \frac{\frac{O(\text{India})}{E(\text{India})}}{\frac{O(\text{BNC})}{E(\text{BNC})}} = \frac{\frac{O_{\text{India}}(R, w_1, w_2) \cdot N_{\text{India}}}{O_{\text{India}}(R, w_1) \cdot O_{\text{India}}(R, w_2)}}{\frac{O_{\text{BNC}}(R, w_1, w_2) \cdot N_{\text{BNC}}}{O_{\text{BNC}}(R, w_1) \cdot O_{\text{BNC}}(R, w_2)}} \quad (5)$$

where  $N$  is corpus size,  $R$  is the relation (*pobj* or *modpp*),  $w_1$  the head (verb or noun),  $w_2$  the preposition.

O/E ratio	Head	Prep	f(India)	O/E (India)	O/E(BNC)	manual inspection comment
80.6962	discuss	about	10	148.012	1.83419	You come we will <i>discuss about</i> it.
51.3664	study	about	7	67.7127	1.31823	Today we are <i>studying about</i> rotation and revolution of the earth.
705.33	advise	into	7	279.731	0.396597	no, consistent parsing error
39.8306	result	into	5	55.3685	1.3901	This <i>resulted into</i> a deep sense of growing loneliness
78.7867	burst	of	5	234.214	2.97276	no
53.0517	arrest	from	5	59.374	1.11917	five more terrorists <i>were arrested from</i> his home
93.5978	etch	at	3	147.232	1.57303	no
67.2343	withstand	to	2	139.353	2.07265	no
46.6381	significant	on	2	33.1642	0.711096	no
45.8399	nice	on	2	70.0133	1.52734	no
84.4974	line	of	2	120.453	1.42552	no
47.4123	land	into	2	102.124	2.15396	Atul's tendency of worrying too much ... <i>landed him into</i> trouble
107.968	exciting	on	2	315.06	2.9181	no
214.685	benefit	out	2	128.156	0.596949	yes: So they'll <i>benefit out of the faculty teaching</i>

Language models are imperfect and produce a certain level of errors.

Even

- when applied on out-of-domain texts,
- without choice context
- without reaching statistical significance

can pick up a signal, e.g. in verb-preposition constructions variation,

based on checking the hits, against lexical entries, and using rationalist intuition.

“Corpus as bicycle of the linguistic mind”

## Same on ICE-Fiji

O/E ratio	Head	Prep	f(Fiji)	O/E (Fiji)	O/E(BNC)	manual inspection comment
14.4021	regard	to	7	41.9521	2.91292	partly: he or she will be reading in regards to a bigger picture
14.616	cause	on	3	34.3407	2.34952	yes: The thought of how much anxiety he had caused on his pare
19.7136	stick	as	2	42.1458	2.1379	no
10.9451	pick	to	2	11.5253	1.05301	yes: allow me to pick my team to the world cup
33.9525	join	into	2	52.5526	1.54783	yes: Women by joining into these organisation benefit a lot
11.1615	involve	into	2	24.255	2.17311	yes: women involving themselves into prostitution
33.3689	include	into	2	65.2377	1.95505	yes: they have included rare ... species ... into the displays
22.3632	implicate	for	2	46.4807	2.07845	no
472.801	gather	upon	2	895.141	1.89327	*: upon evaluating the ... Education Act, it was gathered that ...
15.2663	explain	from	2	40.2206	2.6346	no, consistent parsing error
81.3601	engage	through	2	167.625	2.06028	no
31.246	concentrate	from	2	54.5852	1.74695	no
48.866	capable	in	2	14.2045	0.290684	partly: are capable in committing themselves to work
61.3927	arrive	into	2	43.9975	0.716656	yes: Megan Simpson is expected to arrive into the country

See (Schneider, 2013; Schneider and Zipp, 2013)

---



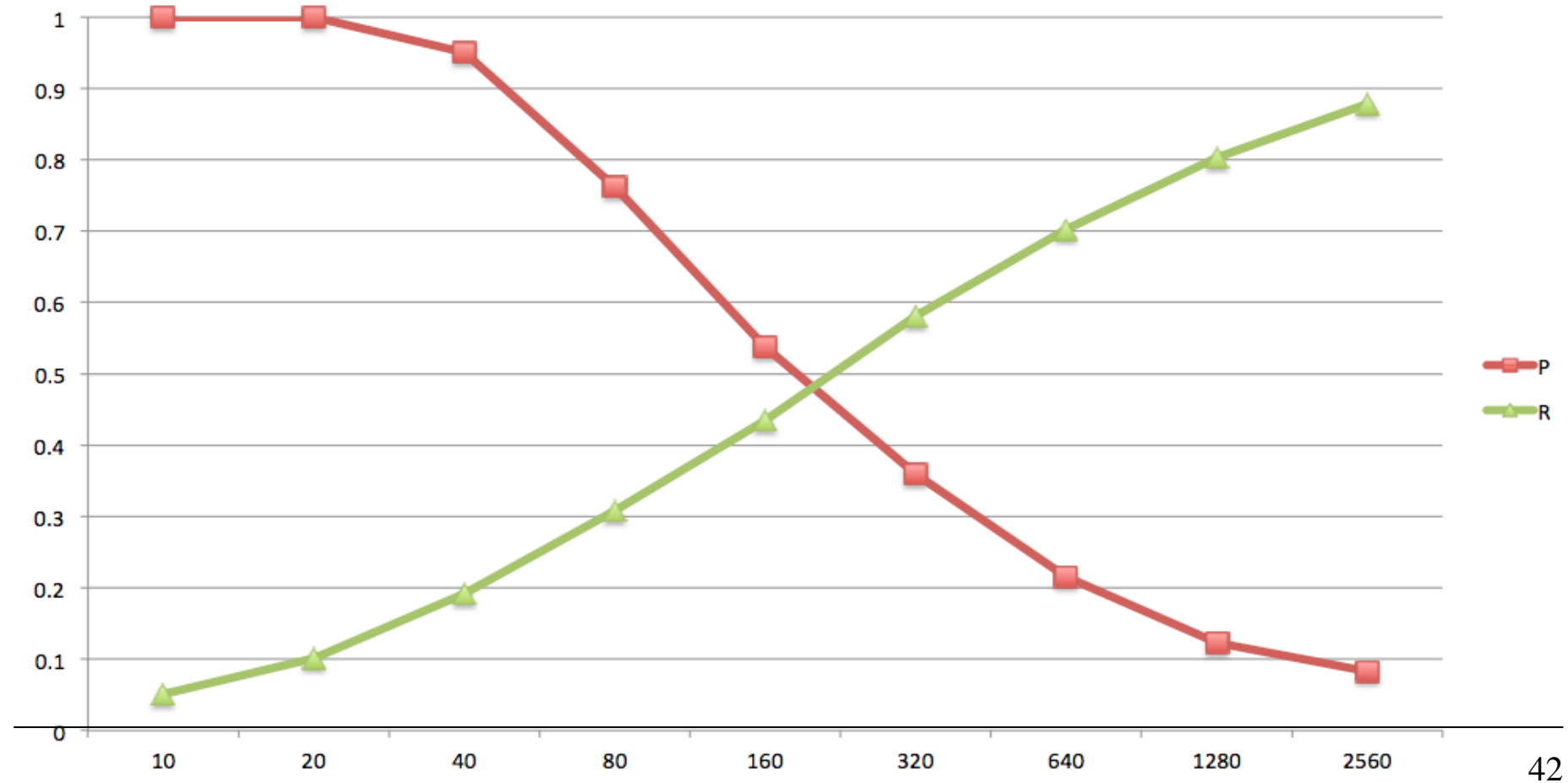
## 4.2 Light Verb Constructions

### T-Score on BNC, correct ones marked

117249	OE	T	Chi	V	Obj	f	f(V)	f(N)	manu
bncx	22.4051	97.0768	219641.0	take	place	10325	128201	21501	+
bncx	8.4774	64.0072	52584.1	have	effect	5266	303211	12254	+
bncx	272.553	59.5303	968964.0	shake	head	3570	6221	12594	
bncx	52.7729	55.3838	167118.0	see	pp	3187	112461	3212	
bncx	5.2405	55.036	22132.7	do	thing	4626	157530	33518	
bncx	41.2167	53.5146	119208.0	ask	question	3008	30365	14376	+
bncx	92.0118	49.4467	226112.0	play	role	2499	19223	8451	+
bncx	32.7848	47.9877	76299.1	play	part	2450	19223	23253	+
bncx	6.3164	47.2243	18094.1	take	part	3148	128201	23253	+
bncx	7.0943	47.083	21023.8	do	anything	3004	157530	16078	
bncx	31.4713	46.4546	68682.6	go	home	2302	26840	16301	
bncx	4.6561	46.3483	15831.8	do	something	3484	157530	28412	
bncx	12.7681	46.2312	31919.8	make	sense	2516	147869	7971	+
bncx	7.8569	46.1637	21934.2	do	job	2798	157530	13522	+
bncx	12.7301	45.6558	31139.6	make	decision	2455	147869	7801	+
bncx	142.241	45.319	293114.0	open	door	2083	11317	7740	
bncx	4.5442	44.5113	25899.6	have	idea	3257	303211	14139	+
bncx	159.397	43.1556	297871.0	answer	question	1886	4923	14376	
bncx	6.3766	43.035	28310.3	have	look	2605	303211	8059	+
bncx	10.938	42.4899	24204.3	make	use	2187	147869	8088	+

Precision = 12 / 20

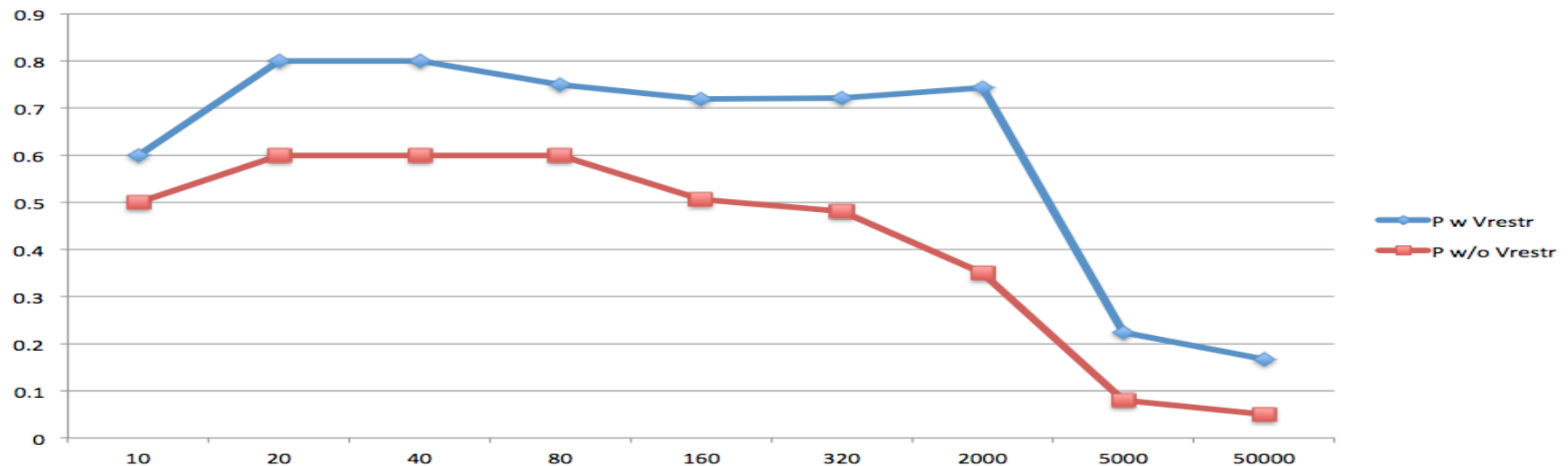
# Evaluation: *give* Precision & Recall on BNC, using T-Score & simple filter



## Extending to all verbs

BNC, top 320 complete manual filter, then stratified slices @100 types

- Vrestr = [take, have, do, ask, play, make, give, provide, draw, tell, pay, meet, change, keep, form, attend, raise, become, reach, cause, live, sing, turn, catch, perform, adopt, put, cover, lead, cover, send, focus, show, receive, suffer, issue, exercise, pay, form, set, feel, ring, issue, suffer, commit]



- Here T-score works best: frequency of LVC is a factor
- LVCs are an open list, and gradient
- some regional variation: e.g. take vs. make decision

see Ronan and Schneider (accepted for publication)

## 5 Parser as Human Processing Model

Statistical disambiguation models necessary, e.g.  $p(Rel|words)$ , lexical priming for syntax.

E.g.  $P(\text{object}(\text{eat}, \text{pizza})) > P(\text{adjunct}(\text{eat}, \text{pizza}))$

$P(\text{verbal\_pobj}(\text{take}, \text{report}, \text{into}, \text{consideration})) > P(\text{nominal\_ppmod}(\text{take}, \text{report}, \text{into}, \text{consideration}))$

Pro3Gres (Schneider, 2008) MLE <sup>a</sup>:  $P(R, dist|a, b) \cong p(R|a, b) \cdot p(dist|R) \cong$

$$\frac{f(R, a, b)}{\sum_{i=1}^n f(R_i, a, b)} \cdot \frac{f(R, dist)}{fR} \quad (6)$$

Disambiguation not Generation model: Reader not Writer

**Parser Score** for a sentence = Summed probabilities over all parsing steps for an entire derivation. Each parsing step is an attachment decision  $p(Rel|words)$  leading to a node in the ensuing syntax tree. Added and adding many other factors, e.g. semantic expectations (Schneider, 2012)

Keller (2010) suggests to use broad-coverage parsers as psycholinguistic language models.

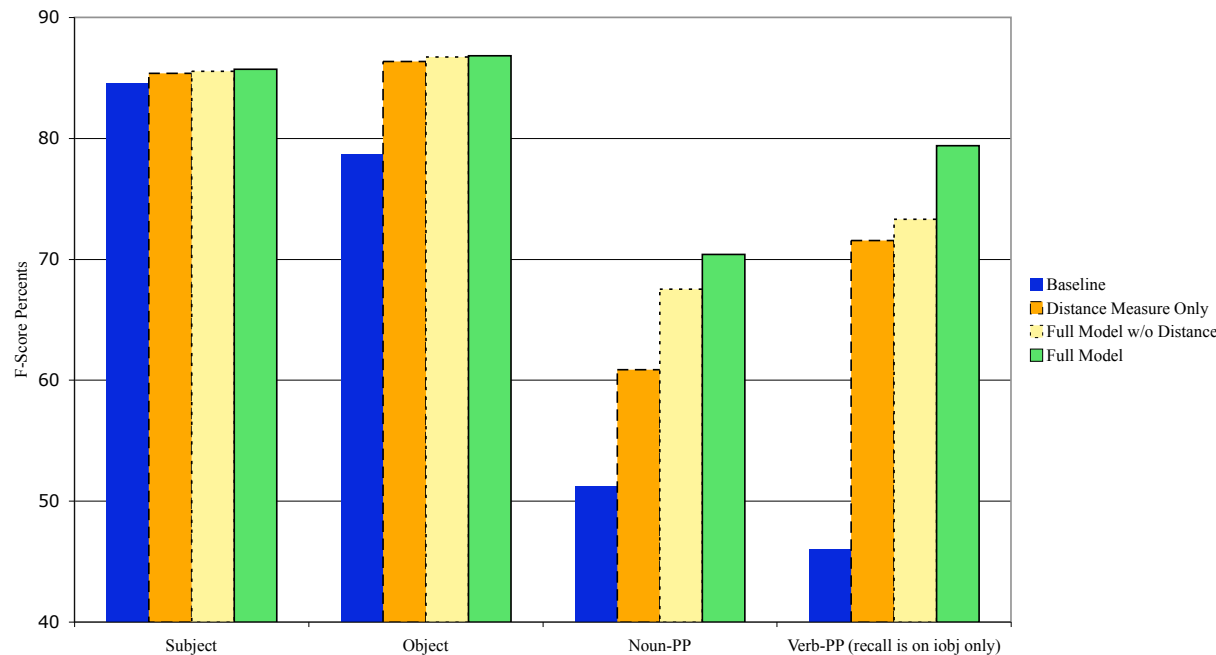
---

<sup>a</sup>  $R$ = Gram. Relation;  $a, b$  = head lemmas,  $dist$  = distance (chunks)

## 5.1 Varying model parameters

Evaluation on 500 sentence GREVAL corpus (Carroll, Minnen, and Briscoe, 2003)

Effects of the Statistical Model: F-Scores from Baseline to Full Model

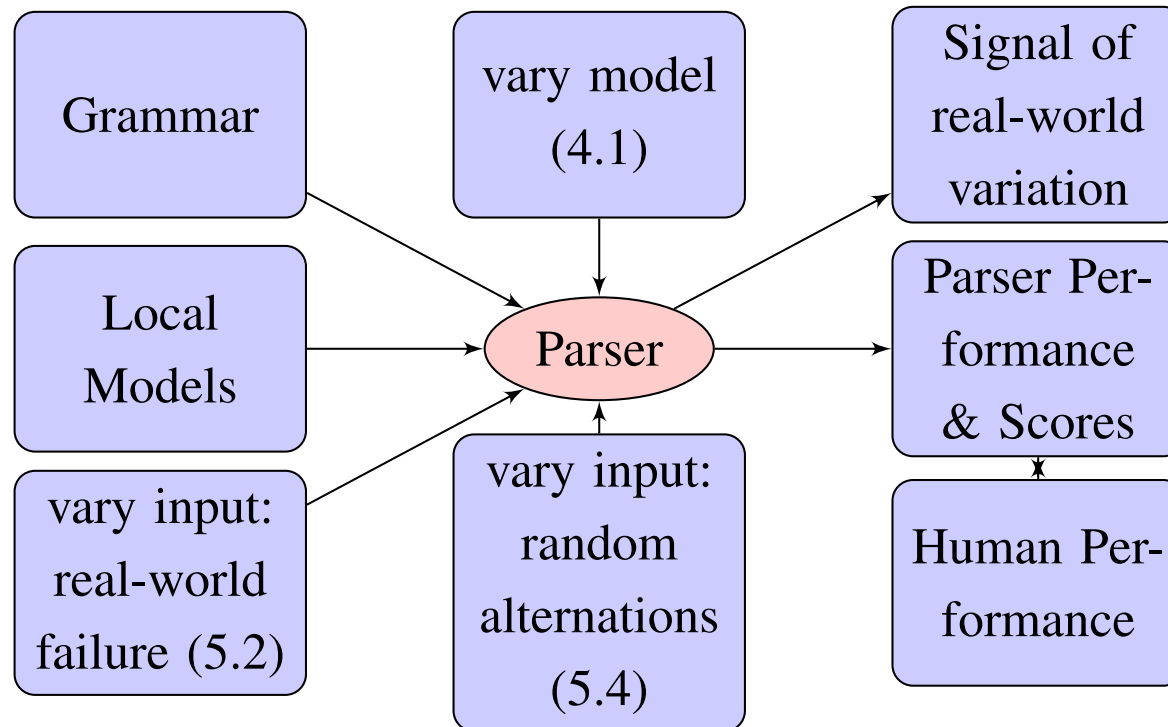


- Baseline: only syntax rules: w/o bi-lexical probabilities (a,b)
- Distance measure: recency (dist)
- Full Model w/o distance: bi-lexical preferences (a,b)
- Full Model: recency + bi-lexical probabilities (dist,a,b)

This increase tells us how much (both for humans and automatic parsers) lexical preferences and recency (idiom principle) help for the syntactic interpretation (syntax principle).

Language model of syntactic parser: Programme:

- use a syntactic parser as global model on large corpora
- compare human and machine parses
- vary the model: with/out lexical priming
- use real-world data with real-world failures, e.g. learner English (L2)
- randomly manipulate input using permitted syntactic operations



## 5.2 Varying input: real-world learner corpus

original L2 vs. error-corrected utterance. L2 utterances fit the parsing model less well.

We have manually annotated  $> 500$  syntactic relations from random sentences from the NICT corpus. Error rate is almost halved on the corrected text.

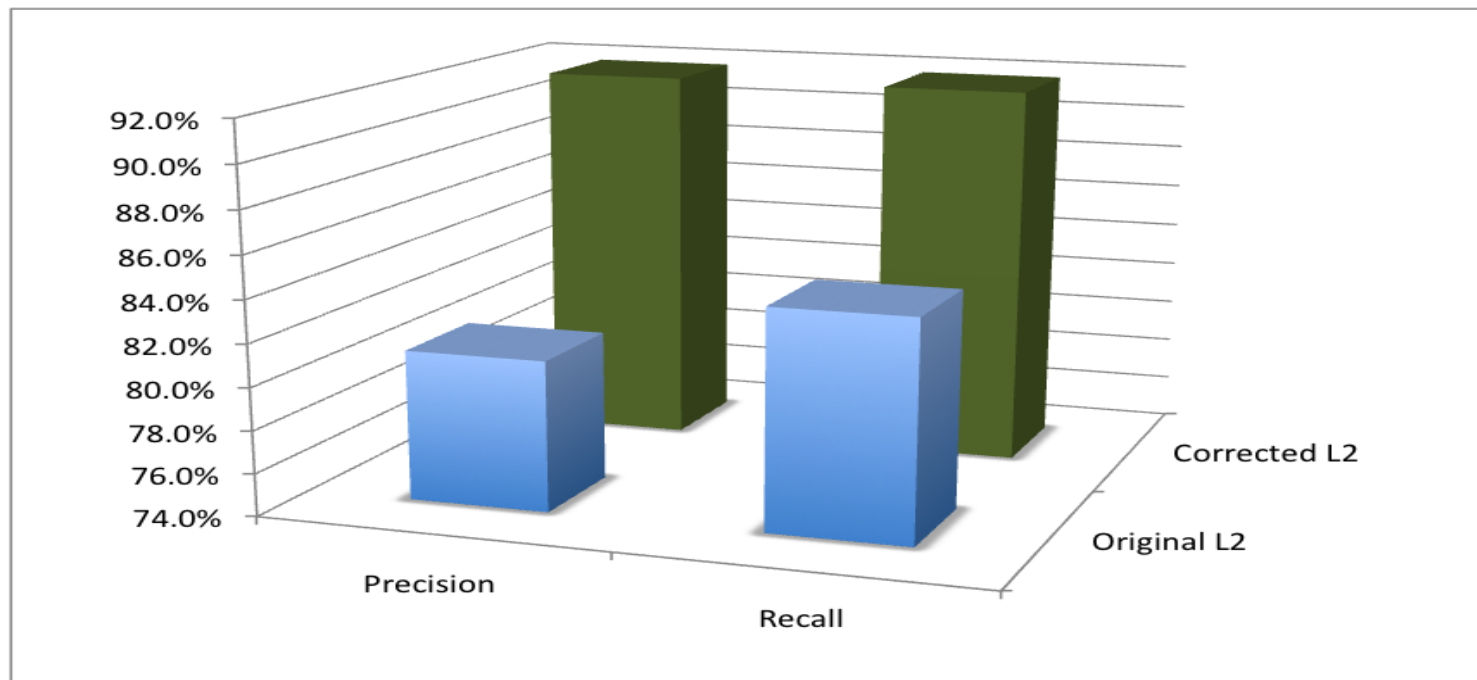
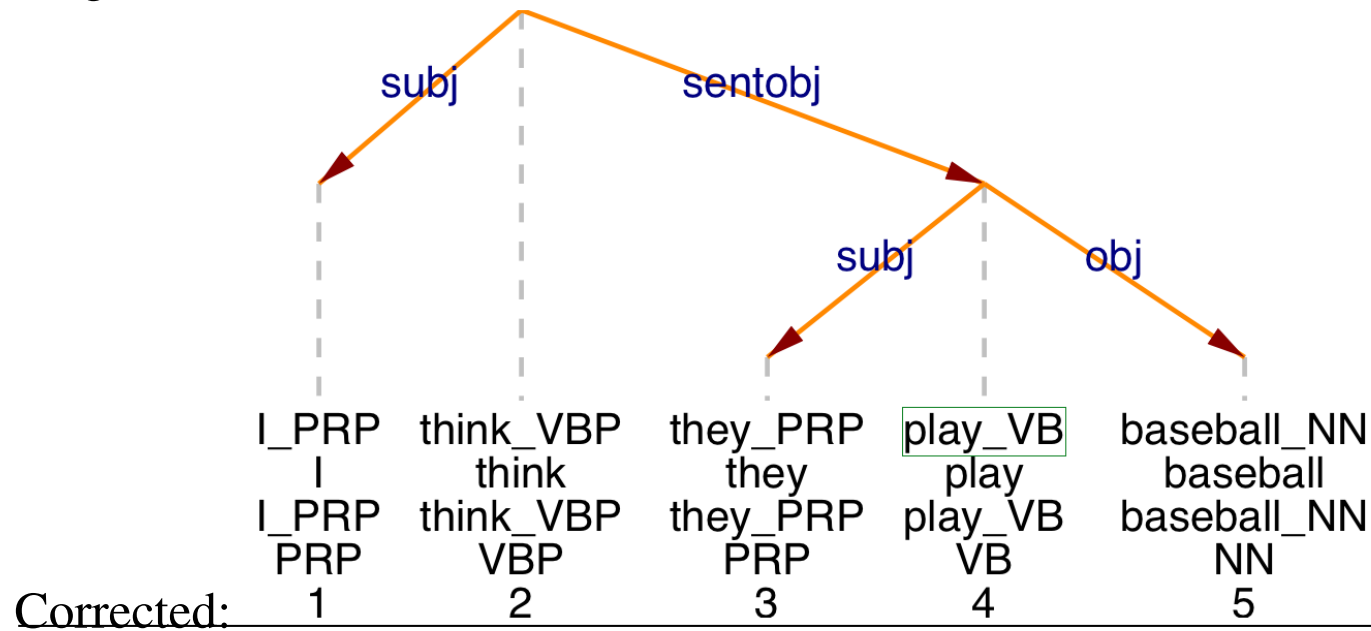
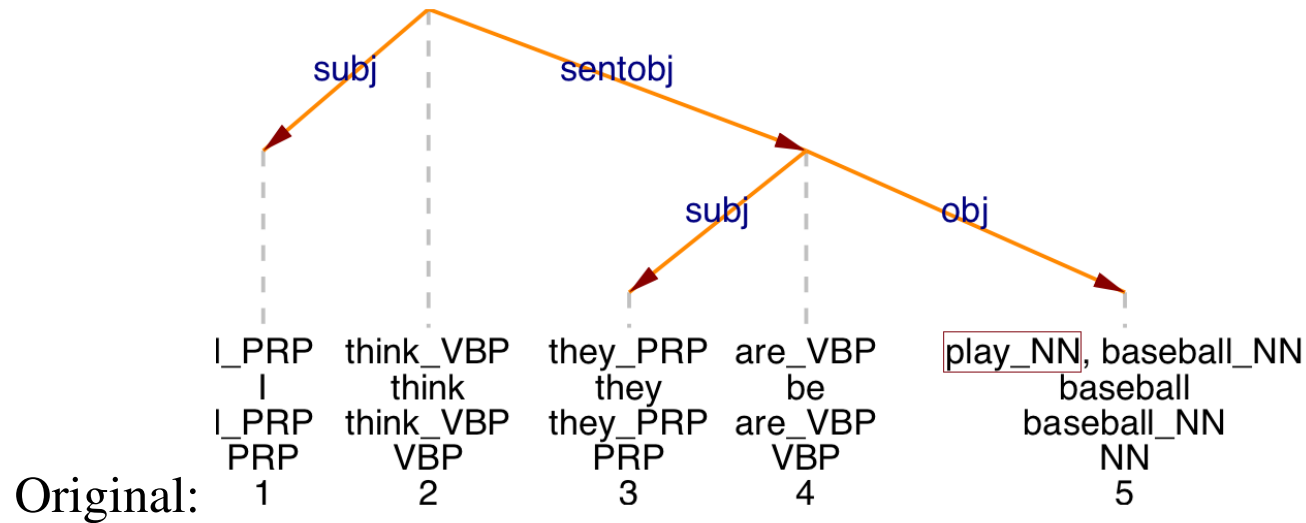


Figure 1: Precision and recall of subset of parsed NICT learner corpus



# Dependency Parsing & Applications

---



### 5.3 Parser Scores

Each derivation of a sentence gets a summation of its  $p(SynRel|words)$ . Model of listener expectations.

- probability-based scores for disambiguation and ranking
- high parser score:
  - utterance matches the expectation / a syntactic parse, ie e.g. entrenched
  - lexical items in combination strongly point to a certain analysis
- low parser score:
  - unexpected input
  - the parser cannot map it well to a syntactic analysis

V	Sentence	Score
ORIG	Usually , I go to the library , and I rent these books .	5054.31
CORR	Usually , I go to the library , and I borrow these books .	8956.83
ORIG	For example , at summer , I can enjoy the sea and breeze .	7186.86
CORR	For example , in summer , I can enjoy the sea and breeze .	8965.99
ORIG	The computer game is very violence in today , but I do n't like it .	6570.44
CORR	Computer games are very violent today , but I do n't like them .	161.753

NB. Depends on sentence length

---

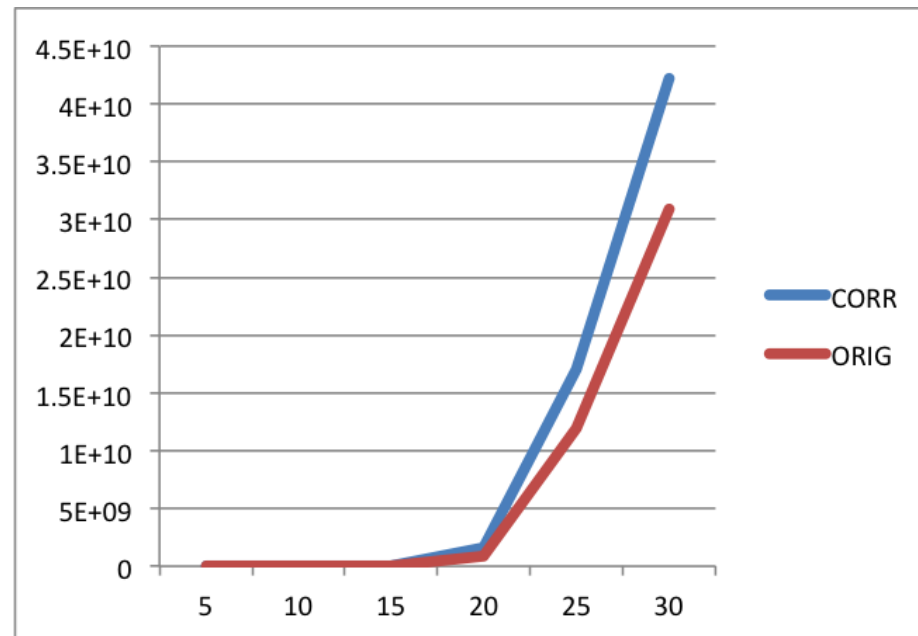


Figure 2: Parser scores on NICT learner corpus, by sentence length in chunks

Parser model fit is a measure of syntactic (un)expectedness (surprise). Green (2014): Parser score development is syntactic surprisal.

Pawley and Syder (1983, 193): “native speakers do not exercise the creative potential of syntactic rules to anything like their full extent, and that, indeed, if they did do so they would not be accepted as exhibiting natively like control of the language.”

---

## 5.4 Ambiguity

Prototypical ambiguity: garden path sentences: discrepancy between a *local maximum* and a *global maximum*. A locally most plausible interpretation needs to be revised due to subsequent text data (Schneider et al. 2005)

typically avoided. Zero-relative clause

*I saw the flower I like*

(?) *I saw the flower pots like*

(1) the collocation between *flower* and *pot* triggers a garden-path noun phrase to be constructed

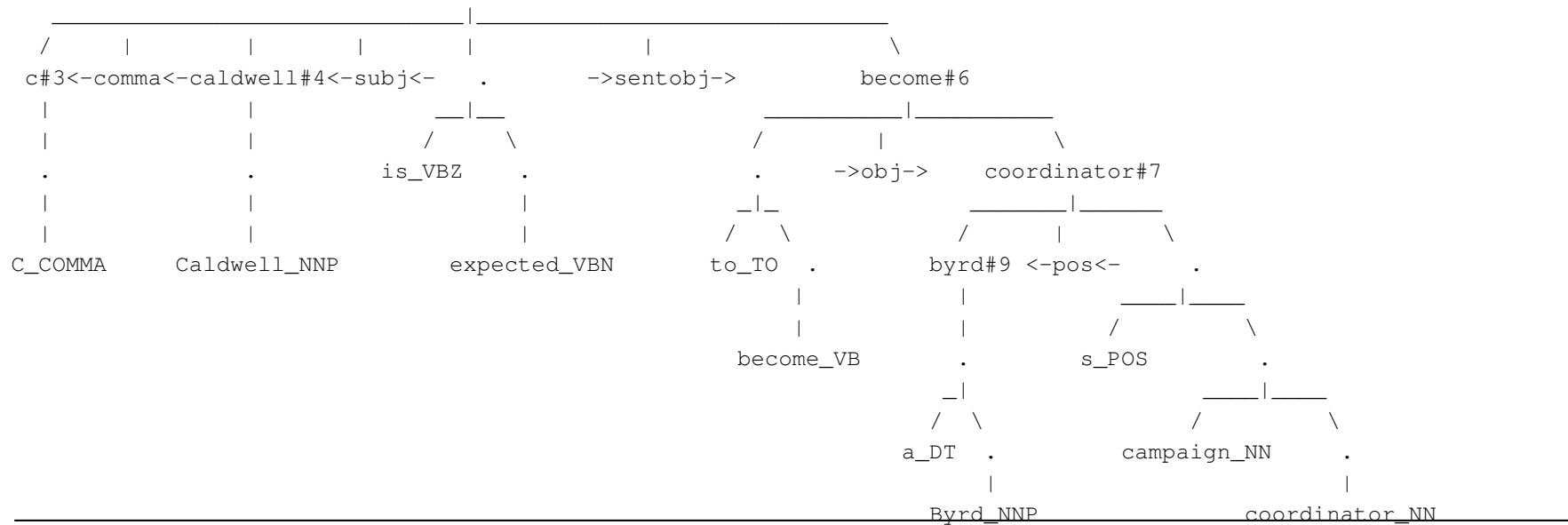
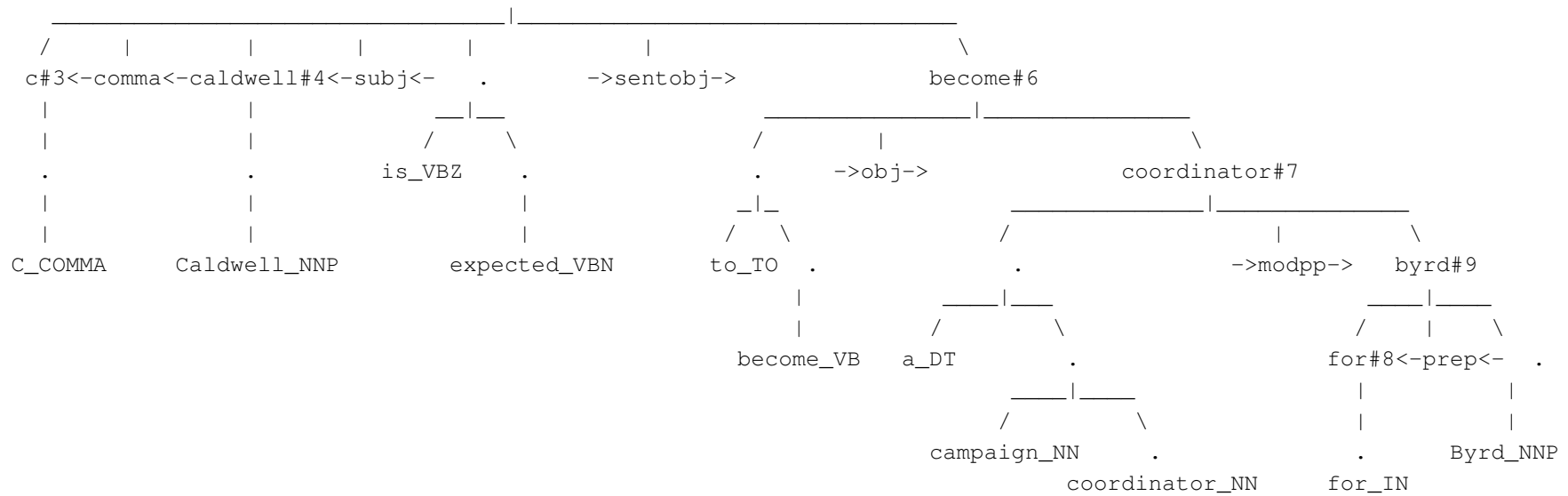
(2) pronouns cannot be pre-modified, which means that no such ambiguity can exist – indeed zero relatives in front of pronouns are particularly frequent.

(3) inanimates such as pots are rarely active verb subjects

So such examples are rare, we cannot find them → we need **to force breaking** the collaboration between syntax and idiom principle → measure human parser surprise, and parser error rate.

# Dependency Parsing & Applications

---

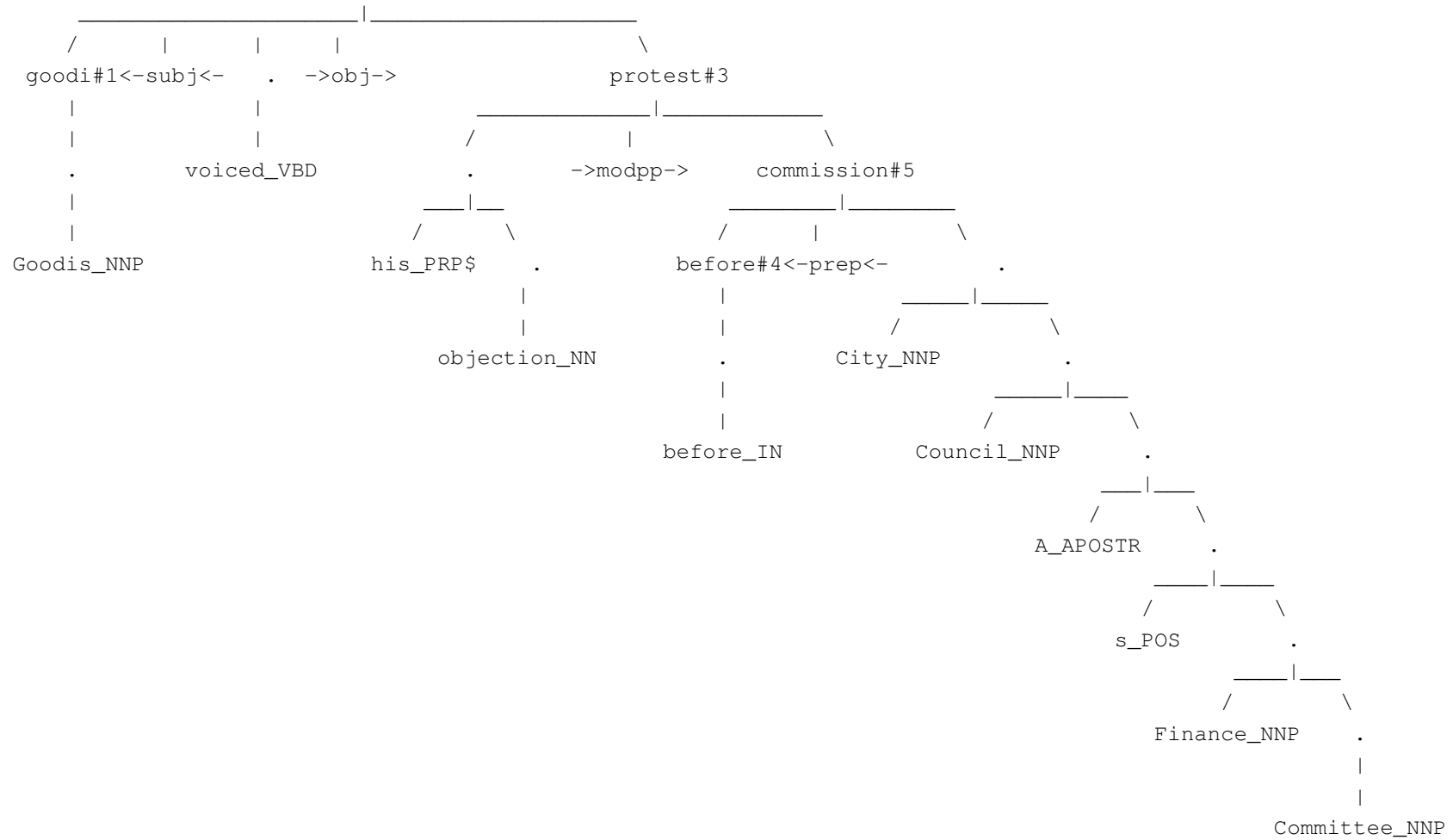




# Dependency Parsing & Applications

---

```
1 +> 230.48533467615135 :: []  
      sound#2
```

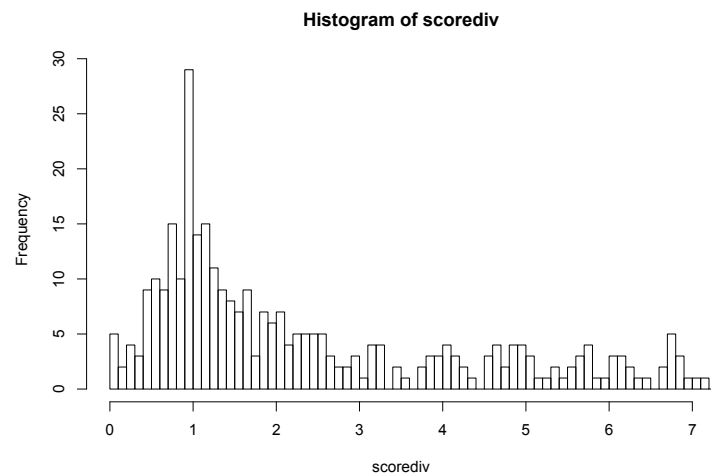


After alternator:

- sentences almost always judged as worse by humans
- surprisal increases (from mean of 13.5 to 13.7, newspaper text)

After synonymizer:

- sentences always judged as worse by humans
- parser scores on texts with synonyms are much lower (4 times, measured by median)





## 6 Syntax and Discourse for Text Mining

- PharmGKB and in the CTD corpus parsed with a dependency parser (Schneider, 2008)
  - Lingpipe: token and sentence segmentation
  - Term recognition: a dictionary-based tool which delivers annotated document spans (terms)
- All entities appearing in same sentence are potentially interacting: *candidate path*.
- If gold standard states that both entities interact in the document → *relevant path*

Assumption: connecting paths between relevant entities are relevant

→ weakly supervised approach learning syntactic features from document-level annotation

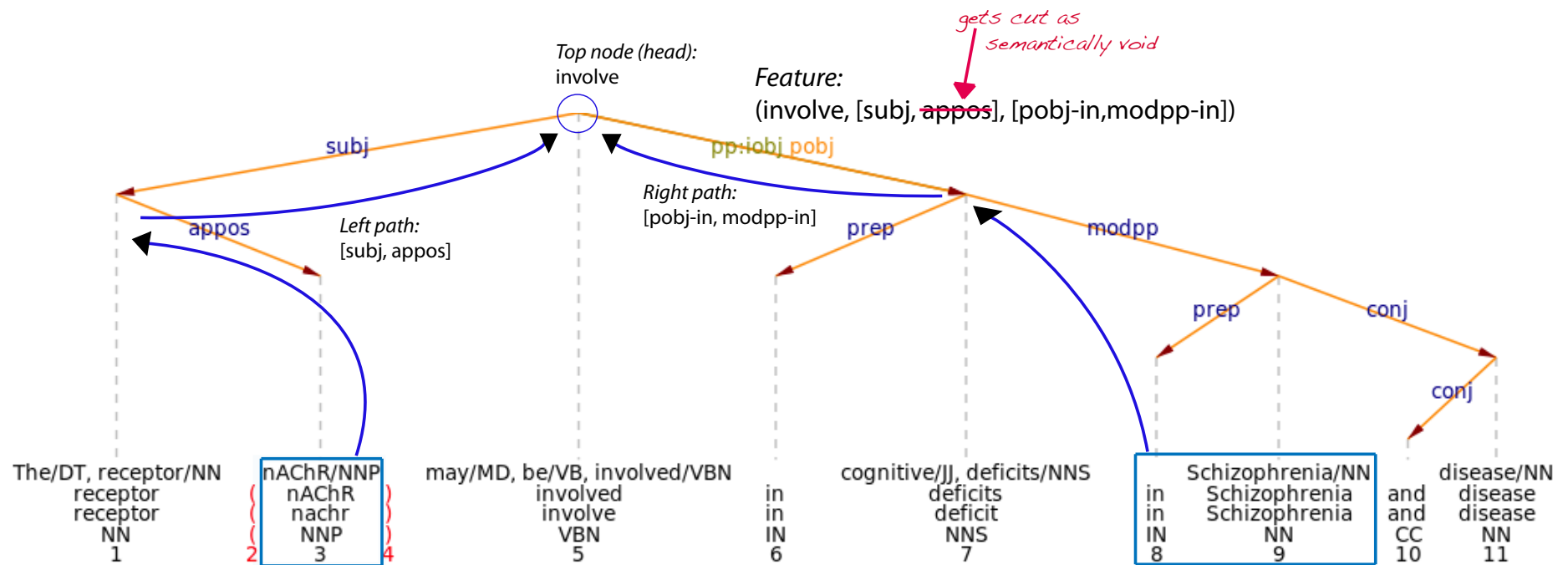


Figure 3: Simplified internal syntactic representation of the sentence “The **neuronal nicotinic acetylcholine receptor alpha7 (nAChR alpha7)** may be involved in cognitive deficits in Schizophrenia and Alzheimer’s disease.” from PubMed abstract 15695160. The curved arrows and dark red notes are aimed at illustrating the path features.

## 6.1 Learning Syntactic Patterns

The calculation of the number of *relevant paths* divided by the number of *candidate paths* gives us the Maximum-Likelihood probability that a path is relevant:

$$p(\text{relevant}) = \frac{\text{freq}(\text{relevant path})}{\text{freq}(\text{candidate path})}$$

p(relevant)	Head	Path1	Path2	TP	Count
13.62%	associate	subj	pobj-with	53	389
17.82%	associate	subj modpp-in	pobj-with	31	174
18.92%	effect	modpp-of	modpp-on modpp-of	21	111
20.65%	association	modpp-of	modpp-with	19	92
6.29%	be	obj modpp-of	subj	19	302
17.82%	metabolize	pobj-by	subj	18	101
29.63%	inhibit	pobj-by	subj	16	54
23.81%	cause	subj modpp-in	obj	15	63
100.00%	analyze	subj modpp-in	pobj-in modpart pobj-with	14	14

Some of the most frequent path types in the PharmGKB training set

---

Apply  $p(\textit{relevant})$  directly  $\rightarrow$  sparse data problem

- using half-paths (MLE model)
- Maximum-Entropy classifier
- Expand abbreviations (and filter)
- Transparent words (shortens paths)

We use deep-linguistic resources like discourse and transparent words (Meyers et al., 1998)

## 6.2 Linguistic Discourse

Discourse: “a unit of language larger than a sentence and which is firmly rooted in a specific context ” (Martin and Ringham, 2000, 51). Broad area of linguistics, partly overlaps with pragmatics and includes a wide range of aspects, for example anaphora resolution, text genre studies, cohesion, felicity, and community-wide background knowledge.

### **Transparent Words (Discourse Truth function preserving simplification)**

- Relations for appositions, conjunctions and hyphens are cut from the path feature
- parts of trees which are headed by a transparent word are cut
- transparent word: word that does not change the meaning of a sentence much if left out: *drug A affects groups of patients* → *drug A affects patients*
- frequency-based approach (Schneider, Kaljurand, and Rinaldi, 2009): words that occur particularly often inside paths are regarded as transparent.

## **Expanding introduced acronyms (Document as unit of language)**

(Schwartz and Hearst, 2003) introduce an algorithm for detecting acronyms in brackets. We use the syntactic relation *apposition*, and profit from concept references.

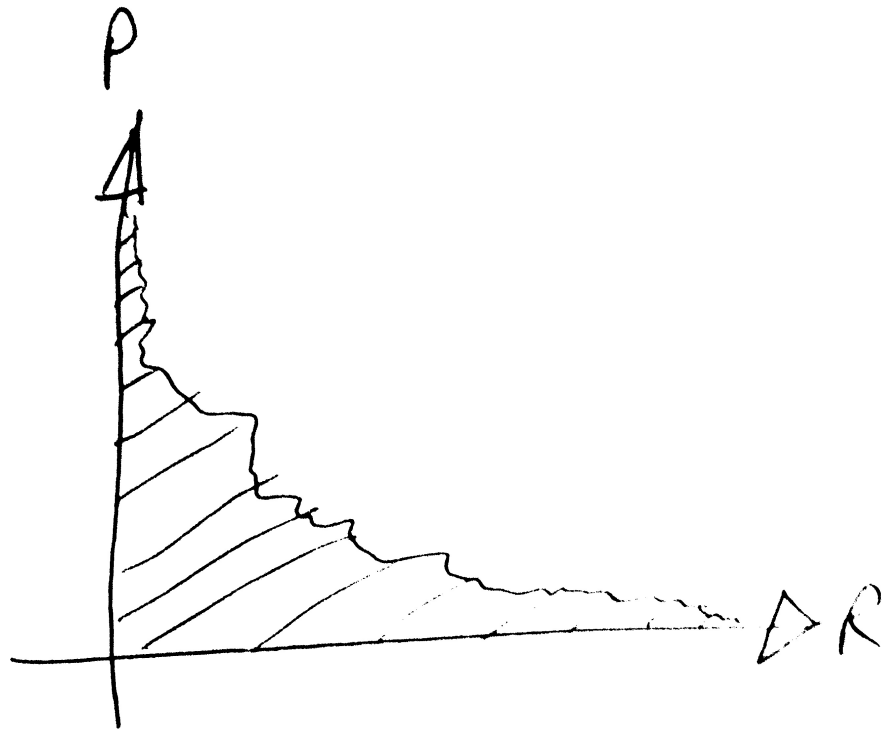
We add the expansion to all acronyms that are introduced in a document, if their concepts differ. This step increases recall at the cost of precision.

*The current studies were designed to examine if quinone intermediates are involved in the toxicity of hepatotoxic halobenzenes, bromobenzene (BB) and 1,2,4-trichlorobenzene (1,2,4-TCB). (CTD, pubmed 10092053)*

Acronym *BB* is given a gene concept by the term recognizer, while it is an acronym of the chemical substance *bromobenzene*, to which *BB* is connected via a syntactic apposition relation. All 5 occurrences of *BB* in the document are thus given the chemical concept of *bromobenzene*.

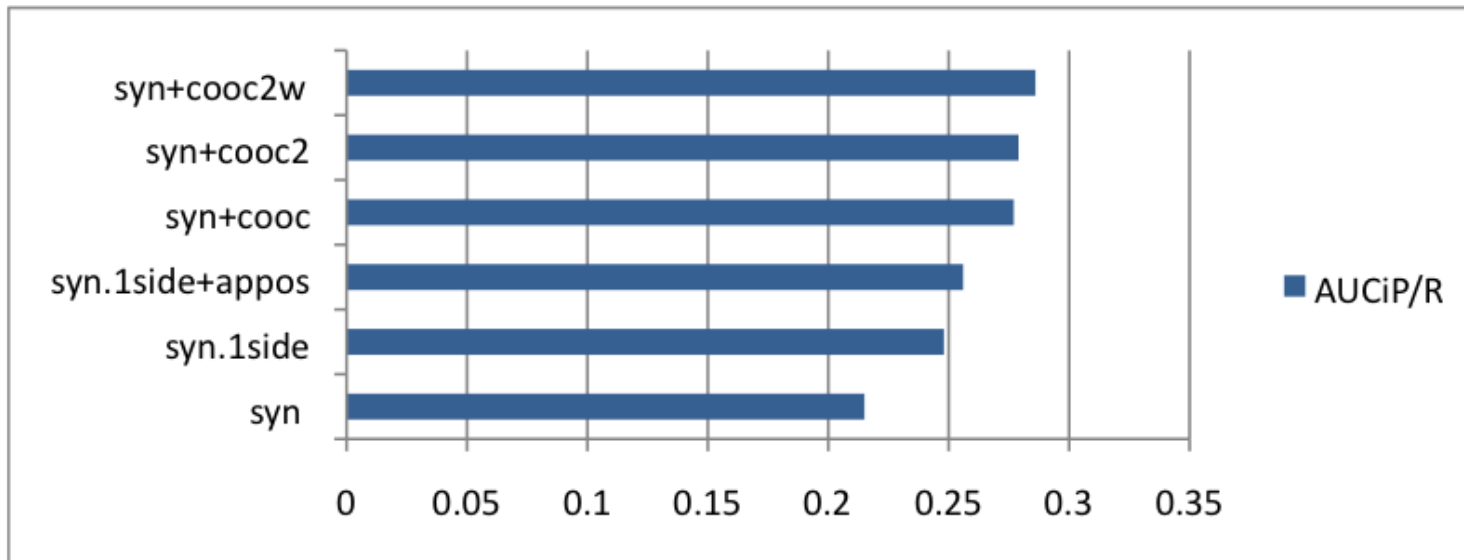
## Evaluation metrics: AUCiP/R

- AUCiP/R measures the area under the interpolated Precision/Recall curve.



This measure directly relates to the expected user's benefit in a curation scenario.

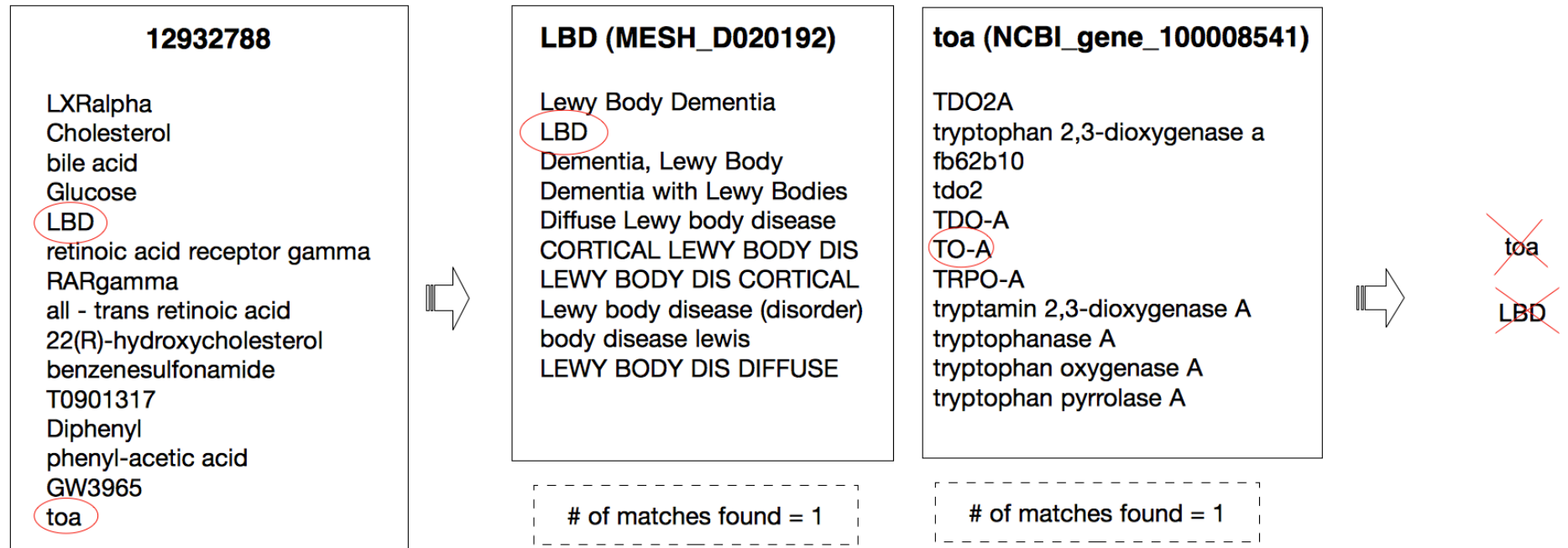
---



- **syn** is purely our syntactic method
- **syn.1side** uses half-path features as a backoff. If either left or right side from a term to the top node match to a decision from gold standard, the decision is reported
- **syn.1side+appos** additionally recognizes acronyms that were introduced by a syntactic apposition relation
- **syn+cooc**. Low recall of syntactic methods can be increased by including sentence-cooccurrence
- **syn+cooc2** sentence-cooccurrence score extended to including the neighbouring sentence. The increase in recall indicates that context of more than one sentence is often necessary.
- **syn+cooc2w** weighs the sentence-cooccurrence score by distance, giving higher scores to entities that appear closer



## Filtering acronyms without expansion candidates



*LBD* refers to ‘LXRalpha ligand-binding domain’, but it was recognized as the disease term ‘Lewy Body Dementia’.

Algorithm checks if any other variant listed under the concept *MESH:D020192* occurs in the text. In the case of *LBD* it is not → *LDB* referring to the disease is removed.

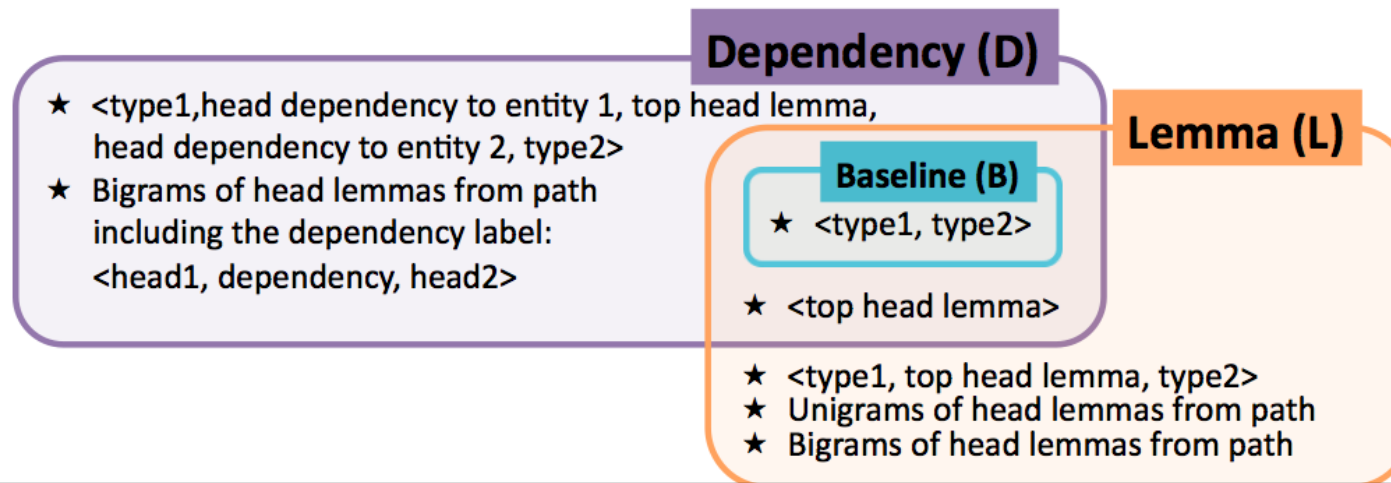
*toa*: mapped to gene ID *CTD:100008541* due to our aggressive candidate generation, but it actually refers to the sequence ‘to a’ in the text. No other variants of the concept *CTD:100008541* can be found in the text → also discarded.

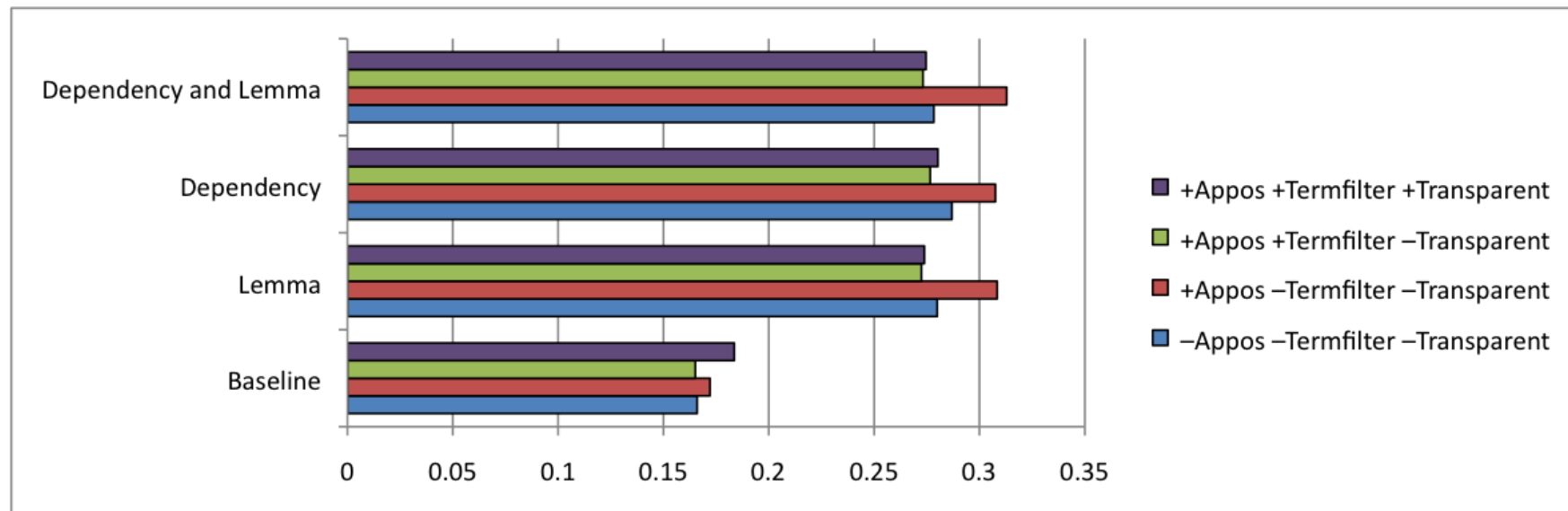
Concepts of short acronyms without promising expansion candidates in the document are filtered (increases precision at the cost of recall)

## 6.3 Maximum Entropy based estimation of path relevancy

- Experiment on CTD knowledge base
- PubMed articles with more than 12 curated relations were omitted
- CTD corpus contains about 24,000 PubMed abstracts; 72,000 relations
- Test (10%) & training data (90%) stratified by number of relations per article
- Maximum Entropy classification tool megam (Daumé, 2004)

Sets of features derived from the candidate paths:





Performance on CTD relation ranking measured by AUCiP/R

The dependency model (**D**), the lemma model (**L**), and the combined model (**DL**) perform substantially better than the baseline (**B**), improving relation ranking by 68%.

- **Appos** shows relative improvement for all eval metrics, **DL** improves by 12%.
- **Termfilter** leads to better precision & F-score, but AUCiP/R and TAP-k suffer.
- Cutting **transparent** words leads to a marginally higher performance, further investigations are needed here.

## 7 Conclusions

- Dependency Grammar as simple yet expressive formalism
- CYK algorithm as reasonable compromise
- Bilexical probabilities help to disambiguate
- Case study on regional variation: Indian English
- Collocations on syntactic structures to extract idioms and light verb constructions
- Parser as language model: higher score, performance and model fit of corrected L2
- Ambiguity: forced alternations lead to lower score and higher surprisal
- Syntax and discourse can help Text Mining

# References

- Abney, Steven. 1995. Chunks and dependencies: Bringing processing evidence to bear on syntax. In Jennifer Cole, Georgia Green, and Jerry Morgan, editors, *Computational Linguistics and the Foundations of Linguistic Theory*, pages 145–164. CSLI.
- Buyko, Ekaterina, Elena Beisswanger, and Udo Hahn. 2012. Extraction of pharmacogenetic and pharmacogenomic relations – a case study using pharmgkb. In *Proceedings of the Pacific Symposium on Biocomputing (PSB)*, pages 376–387, Hawaii.
- Carroll, Hyrum D, Maricel G Kann, Sergey L Sheetlin, and John L Spouge. 2010. Threshold Average Precision ( TAP-k ): A Measure of Retrieval Designed for Bioinformatics. *Methods*, pages 1–8.
- Carroll, John, Guido Minnen, and Edward Briscoe. 2003. Parser evaluation: using a grammatical relation annotation scheme. In Anne Abeillé, editor, *Trebanks: Building and Using Parsed Corpora*. Kluwer, Dordrecht, pages 299–316.
- Carroll, John, Guido Minnen, and Ted Briscoe. 1999. Corpus annotation for parser evaluation. In *Proceedings of the EACL-99 Post-Conference Workshop on Linguistically Interpreted Corpora*, Bergen, Norway.
- Chomsky, Noam. 1995. *The Minimalist Program*. The MIT Press, Cambridge, Massachusetts.
- Collins, Michael. 1999. *Head-Driven Statistical Models for Natural Language Parsing*. Ph.D. thesis, University of Pennsylvania, Philadelphia, PA.
- Collins, Michael and James Brooks. 1995. Prepositional attachment through a backed-off model. In *Proceedings of the Third Workshop on Very Large Corpora*, Cambridge, MA.
- Covington, Michael A. 1992. GB theory as Dependency Grammar. Technical Report AI1992-03, University of Georgia, Athens, Georgia.
- Covington, Michael A. 1994. An empirically motivated reinterpretation of Dependency Grammar. Technical Report AI1994-01, University of Georgia, Athens, Georgia.
- Daumé, Hal III. 2004. Notes on CG and LM-BFGS optimization of logistic regression. Paper available at <http://pub.hal3.name#daume04cg-bfgs>, implementation available at <http://hal3.name/megam/>.
- de Marneffe, Marie-Catherine and Christopher D. Manning. 2008. The stanford typed dependencies representation. In *COLING 2008 Workshop on Cross-framework and Cross-domain Parser Evaluation*, Manchester, UK.
- Fillmore, Charles J. 1968. The case for case. In Emmon Bach and Robert Harms, editors, *Universals in Linguistic Theory*. Holt, Rinehart and Winston, New York, pages 1–88.

- Foth, Killian A. 2005. *Eine umfassende Constraint-Dependenz-Grammatik des Deutschen*. University of Hamburg.
- Green, Matthew J. 2014. An eye-tracking evaluation of some parser complexity metrics. In *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR)*, pages 38–46, Gothenburg, Sweden, April. Association for Computational Linguistics.
- Helbig, Gerhard. 1992. *Probleme der Valenz- und Kasus-theorie*. Konzepte der Sprach- und Literaturwissenschaft. Niemeyer, Tübingen.
- Keller, Frank. 2010. Cognitively plausible models of human language processing. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics: Short Papers*, pages 60–67.
- Martin, Bronwen and Felizitas Ringham. 2000. *Dictionary of Semiotics*. Cassell, New York.
- Meyers, Adam, Catherine Macleod, Roman Yangarber, Ralph Grishman, Leslie Barrett, and Ruth Reeves. 1998. Using NOMLEX to produce nominalization patterns for information extraction. In *Coling-ACL98 workshop Proceedings: the Computational Treatment of Nominals*, Montreal, Canada.
- Mukherjee, Joybrato. 2009. The lexicogrammar of present-day Indian English. Corpus-based perspectives on structural nativisation. In Ute Römer and Rainer Schulze, editors, *Exploring the Lexis-Grammar Interface*. John Benjamins, Amsterdam, pages 117–135.
- Nivre, Joakim. 2006. *Inductive Dependency Parsing*. Text, Speech and Language Technology 34. Springer, Dordrecht, The Netherlands.
- Nivre, Joakim, Johan Hall, Sandra Kübler, Ryan McDonald, Jens Nilsson, Sebastian Riedel, and Deniz Yuret. 2007. The CoNLL 2007 shared task on dependency parsing. In *Proceedings of the CoNLL Shared Task Session of EMNLP-CoNLL 2007*, pages 915–932.
- Pawley, Andrew and Frances Hodgetts Syder. 1983. Two puzzles for linguistic theory: Native-like selection and native-like fluency. In J. C. Richards and R. W. Schmidt, editors, *Language and Communication*. Longman, London, pages 191–226.
- Ronan, Patricia and Gerold Schneider. accepted for publication. Determining light verb constructions in contemporary british and irish english. *International Journal of Corpus Linguistics*.
- Sand, Andrea. 2004. Shared morpho-syntactic features in contact varieties of English: Article use. *World Englishes*, 23:281–98.
- Schneider, Edgar. 2004. How to trace structural nativization: Particle verbs in World Englishes. *World Englishes*, 23:2:227–249.
- Schneider, Gerold. 2008. *Hybrid Long-Distance Functional Dependency Parsing*. Doctoral Thesis, Institute of Computational Linguistics, University of Zurich.
- Schneider, Gerold. 2012. Using semantic resources to improve a syntactic dependency parser. In Viktor Pekar Verginica Barbu Mititelu, Octavian Popescu, editor, *SEM-II workshop at LREC 2012*.
- Schneider, Gerold. 2013. Using automatically parsed corpora to discover lexico-grammatical features of English varieties. In Fryni Kakoyianni Doa, editor, *Penser le*
-

*Lexique-Grammaire, perspectives actuelles*, Colloques, Congrès et Conférences – Sciences du Langage, Histoire de la Langue et des Dictionnaires. Éditions Honoré Champion, Paris, pages 491–504.

Schneider, Gerold, Kaarel Kaljurand, and Fabio Rinaldi. 2009. Detecting Protein/Protein Interactions using a parser and linguistic resources. In *CICLing 2009, 10th International Conference on Intelligent Text Processing and Computational Linguistics*, pages 406–417, Mexico City, Mexico. Springer LNC 5449.

Schneider, Gerold, Fabio Rinaldi, Kaarel Kaljurand, and Michael Hess. 2005. Closing the gap: Cognitively adequate, fast broad-coverage grammatical role parsing. In *ICEIS Workshop on Natural Language Understanding and Cognitive Science (NLUCS 2005)*, Miami, FL, May 2005.

Schneider, Gerold and Lena Zipp. 2013. Discovering new verb-preposition combinations in New Englishes. In Joybrato Mukherjee and Magnus Huber, editors, *Studies in Variation, Contacts and Change in English, Volume 14 – Corpus Linguistics and Variation in English: Focus on non-native Englishes*. Varieng, Helsinki.

Schwartz, AS and MA Hearst. 2003. A simple algorithm for identifying abbreviation definitions in biomedical text. *Pac Symp Biocomput*, pages 451–462.

Sedlatschek, Andreas. 2009. *Contemporary Indian English: variation and change*. Varieties of English around the world. John Benjamins, Amsterdam / Philadelphia.

Sennrich, Rico, Gerold Schneider, Martin Volk, and Martin Warin Warin. 2009. A new hybrid dependency parser for German. In C. Chiarcos, R. E. de Castilho, and M. Stede, editors, *Von der Form zur Bedeutung: Texte automatisch verarbeiten / From Form to Meaning: Processing Texts Automatically. Proceedings of the Biennial GSCL Conference 2009*, pages 115–124, Tübingen, Germany.

Tesnière, Lucien. 1959. *Éléments de Syntaxe Structurale*. Librairie Klincksieck, Paris.