

Beyond the Science of the Wall Street Journal

WeScience: Registration now Open — RSVP!

Stephan Oepen

Universitetet i Oslo & CSLI Stanford

oe@ifi.uio.no

Advance Warning: A Note of Caution

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 . Rudolph Agnew , 55 years old and former chairman of Consolidated Gold Fields PLC , was named a nonexecutive director of this British industrial conglomerate . A form of asbestos once used to make Kent cigarette filters has caused a high percentage of cancer deaths among a group of workers exposed to it more than 30 years ago , researchers reported . The asbestos fiber , crocidolite , is unusually resilient once it enters the lungs , with even brief exposures to it causing symptoms that show up decades later , researchers said . Lorillard Inc. , the unit of New York-based Loews Corp. that makes Kent cigarettes , stopped using crocidolite in its Micronite cigarette filters in 1956 . Although preliminary findings were reported more than a year ago , the latest results appear in today 's New England Journal of Medicine , a forum likely to bring new attention to the problem . A Lorillard spokeswoman said , " This is an old story . We 're talking about years ago before anyone heard of asbestos having any questionable properties . There is no asbestos in our products now . " Neither Lorillard nor the researchers who studied the workers were aware of any research on smokers of the Kent cigarettes . " We have no useful information on whether users are at risk , " said James A. Talcott of Boston 's Dana-Farber Cancer Institute . Dr. Talcott led a team of researchers from the National Cancer Institute and the medical schools of Harvard University and Boston University . The Lorillard spokeswoman said asbestos was used in " very modest amounts " in making paper for the filters in the early 1950s and replaced with a different type of filter in 1956 . From 1953 to 1955 , 9.8 billion Kent cigarettes with the filters were sold , the company said . ~~Among 33 men who worked closely with the substance , 28 have died — more than three times the expected number .~~ Four of the five surviving workers have asbestos-related diseases , including three with recently diagnosed cancer . The total of 18 deaths from



Advance Warning: A Note of Caution

+ personal

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 . Rudolph Agnew , 55 years old and former chairman of Consolidated Gold Fields PLC , was named a nonexecutive director of this British industrial conglomerate . A form of asbestos once used to make Kent cigarette filters has caused a high percentage of cancer deaths among a group of workers exposed to it more than 30 years ago , researchers reported . The asbestos fiber , crocidolite , is unusually resilient once it enters the lungs , with even brief exposures to it causing symptoms that show up decades later , researchers said . Lorillard Inc. , the unit of New York-based Loews Corp. that makes Kent cigarettes , stopped using crocidolite in its Micronite cigarette filters in 1956 . Although preliminary findings were reported more than a year ago , the latest results appear in today 's New England Journal of Medicine , a forum likely to bring new attention to the problem . A Lorillard spokeswoman said , " This is an old story . We 're talking about years ago before anyone heard of asbestos having any questionable properties . There is no asbestos in our products now . " Neither Lorillard nor the researchers who studied the workers were aware of any research on smokers of the Kent cigarettes . " We have no useful information on whether users are at risk , " said James A. Talcott of Boston 's Dana-Farber Cancer Institute . Dr. Talcott led a team of researchers from the National Cancer Institute and the medical schools of Harvard University and Boston University . The Lorillard spokeswoman said asbestos was used in " very modest amounts " in making paper for the filters in the early 1950s and replaced with a different type of filter in 1956 . From 1953 to 1955 , 9.8 billion Kent cigarettes with the filters were sold , the company said . Among 33 men who worked closely with the substance , 28 have died — more than three times the expected number . Four of the five surviving workers have asbestos-related diseases , including three with recently diagnosed cancer . The total of 18 deaths from



Beyond the Science of the Wall Street Journal (2)

ULA 05-DEC-07 (oe011-110)

Advance Warning: A Note of Caution

+ personal

- objective



Beyond the Science of the Wall Street Journal (2)

UCLA 5-DEC-07 (oe011-110)

Advance Warning: A Note of Caution

+ personal

- objective

- science



Beyond the Science of the Wall Street Journal (2)

UCLA - 05-DEC-07 (oe011-110)

Advance Warning: A Note of Caution

+ personal

- objective

- science

+ politics



Beyond the Science of the Wall Street Journal (2)

UCLA - 5-DEC-07 (oe011-110)

Advance Warning: A Note of Caution

+ personal

-- objective

- science

+ politics



Beyond the Science of the Wall Street Journal (2)

UCLA 5-DEC-07 (oe011-110.110)

Advance Warning: A Note of Caution

+ personal

-- objective

-- science

+ politics



Beyond the Science of the Wall Street Journal (3)

ULA 5-DEC-07 (oe0111.110.110)

Advance Warning: A Note of Caution

+ personal

-- objective

- science

+ politics



Beyond the Science of the Wall Street Journal (3)

ULA -- 5-DEC-07 (060120.110.110)

Some Near-Authentic Quotes and Reflections

To me, the ultimate goal of our new field of Computational Linguistics is to build machines that, in a suitable interpretation of that term, 'understand' human language.

(Martin Kay, 1960s)



Some Near-Authentic Quotes and Reflections

To me, the ultimate goal of our new field of Computational Linguistics is to build machines that, in a suitable interpretation of that term, 'understand' human language.

(Martin Kay, 1960s)

20 Years of Progress in Statistical Parsing

- Parsing into PTB-style trees has been a crisp task for many years;
- great advances: representations, algorithms, probabilistic models;
- F_1 : 0.84 (Magerman, 1994) \rightarrow 0.91 (Charniak & Johnson, 2005).



Some Near-Authentic Quotes and Reflections

To me, the ultimate goal of our new field of Computational Linguistics is to build machines that, in a suitable interpretation of that term, 'understand' human language.

(Martin Kay, 1960s)

20 Years of Progress in Statistical Parsing

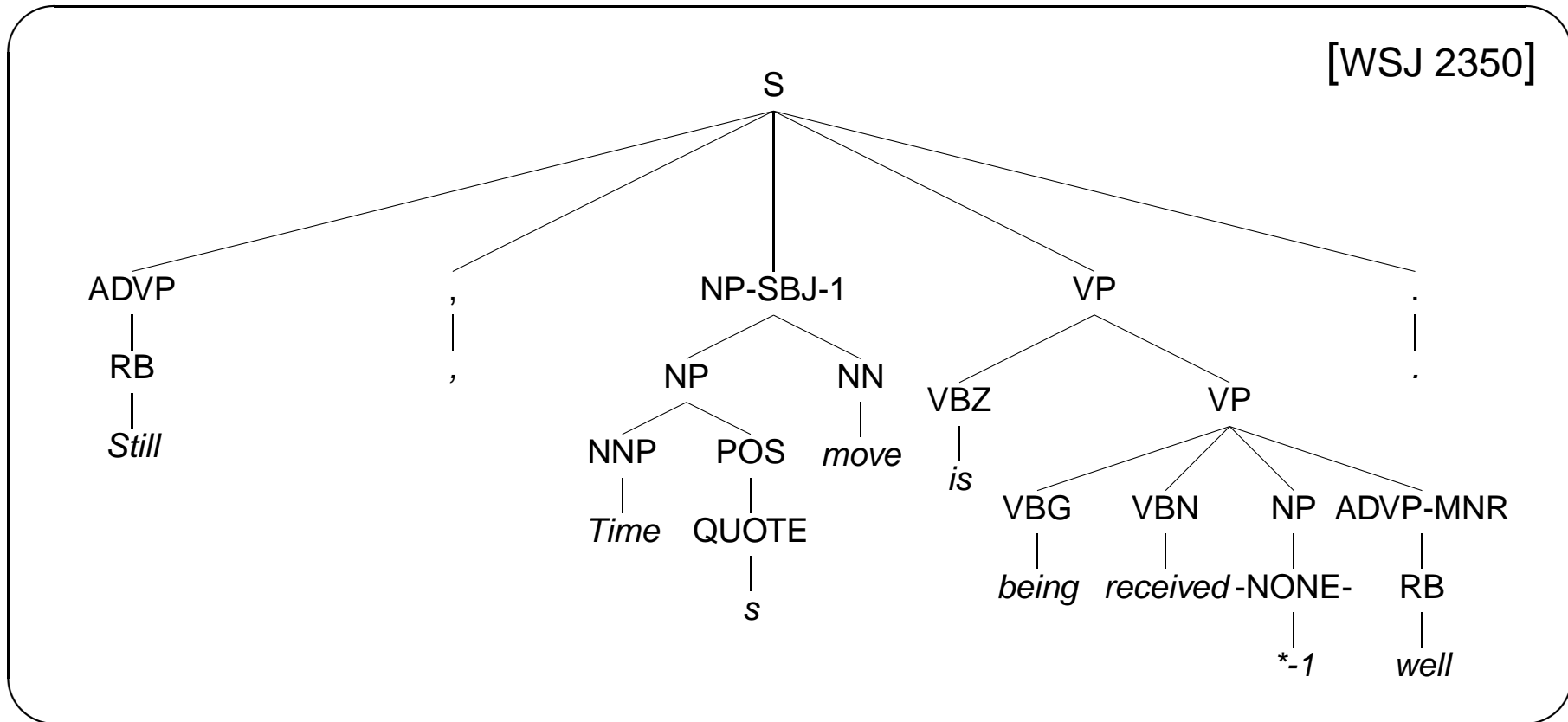
- Parsing into PTB-style trees has been a crisp task for many years;
- great advances: representations, algorithms, probabilistic models;
- F_1 : 0.84 (Magerman, 1994) \rightarrow 0.91 (Charniak & Johnson, 2005).

In the past decade or so, Computational Linguistics has degenerated into the science of the Wall Street Journal.

(Ron Kaplan, 2007)



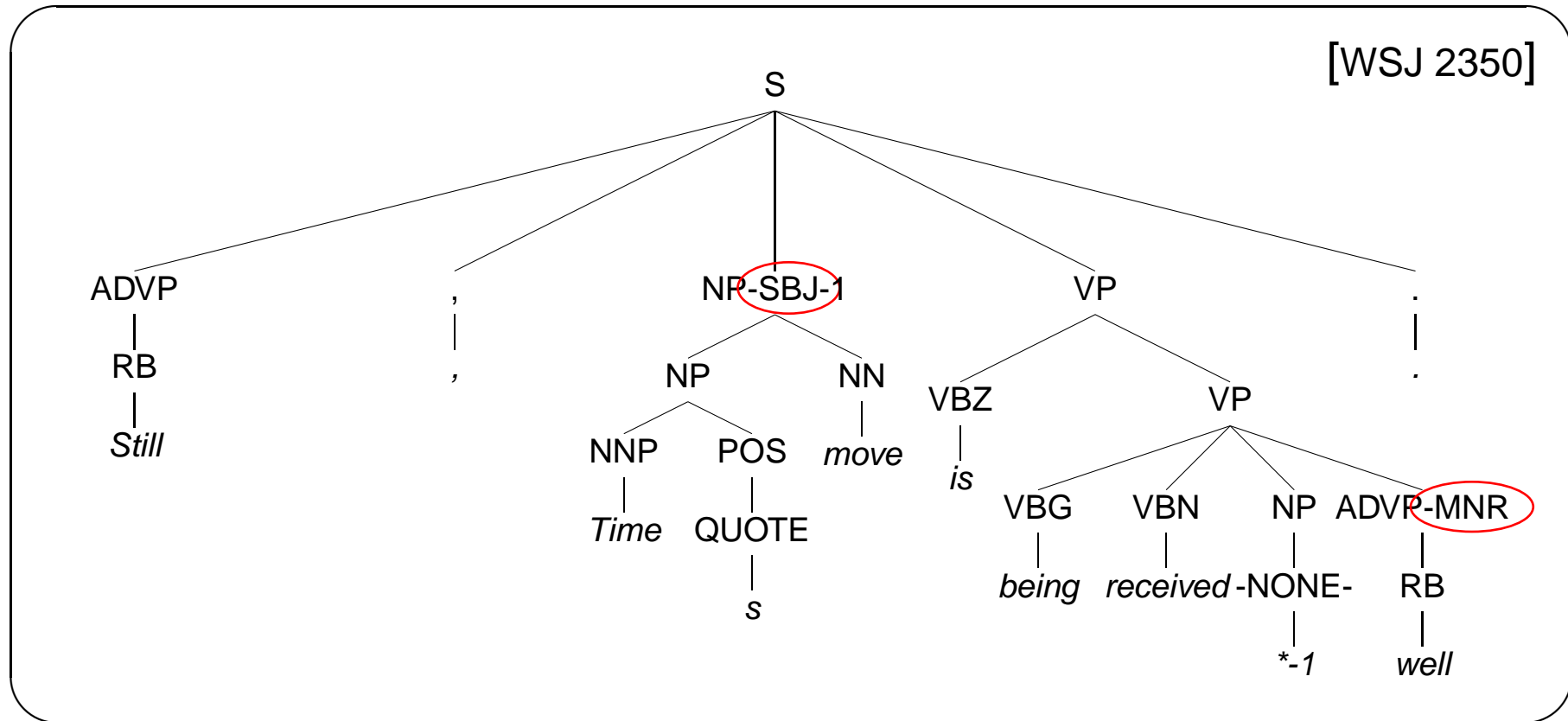
One Example from the Penn Treebank



Still, Time's move is being received well.



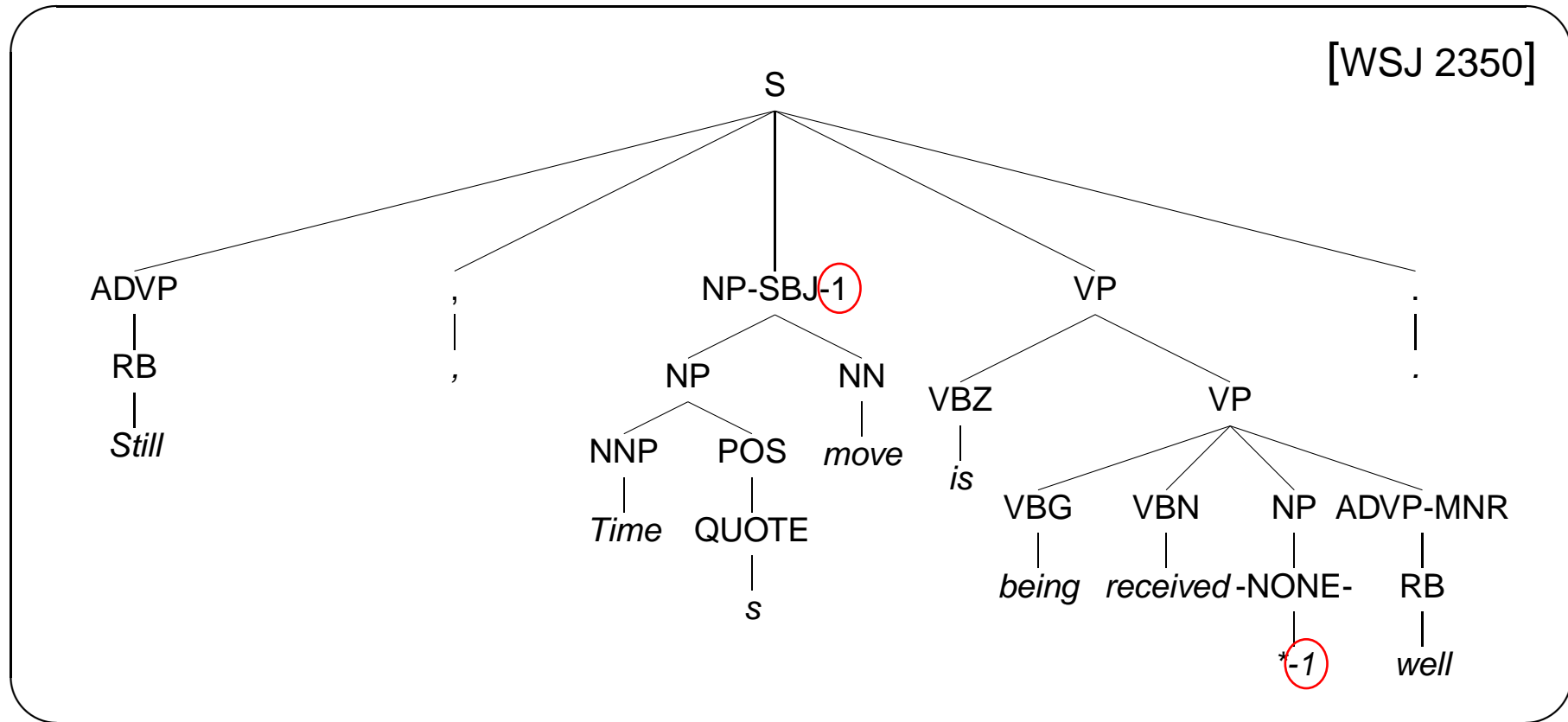
One Example from the Penn Treebank



Still, Time's move is being received well.



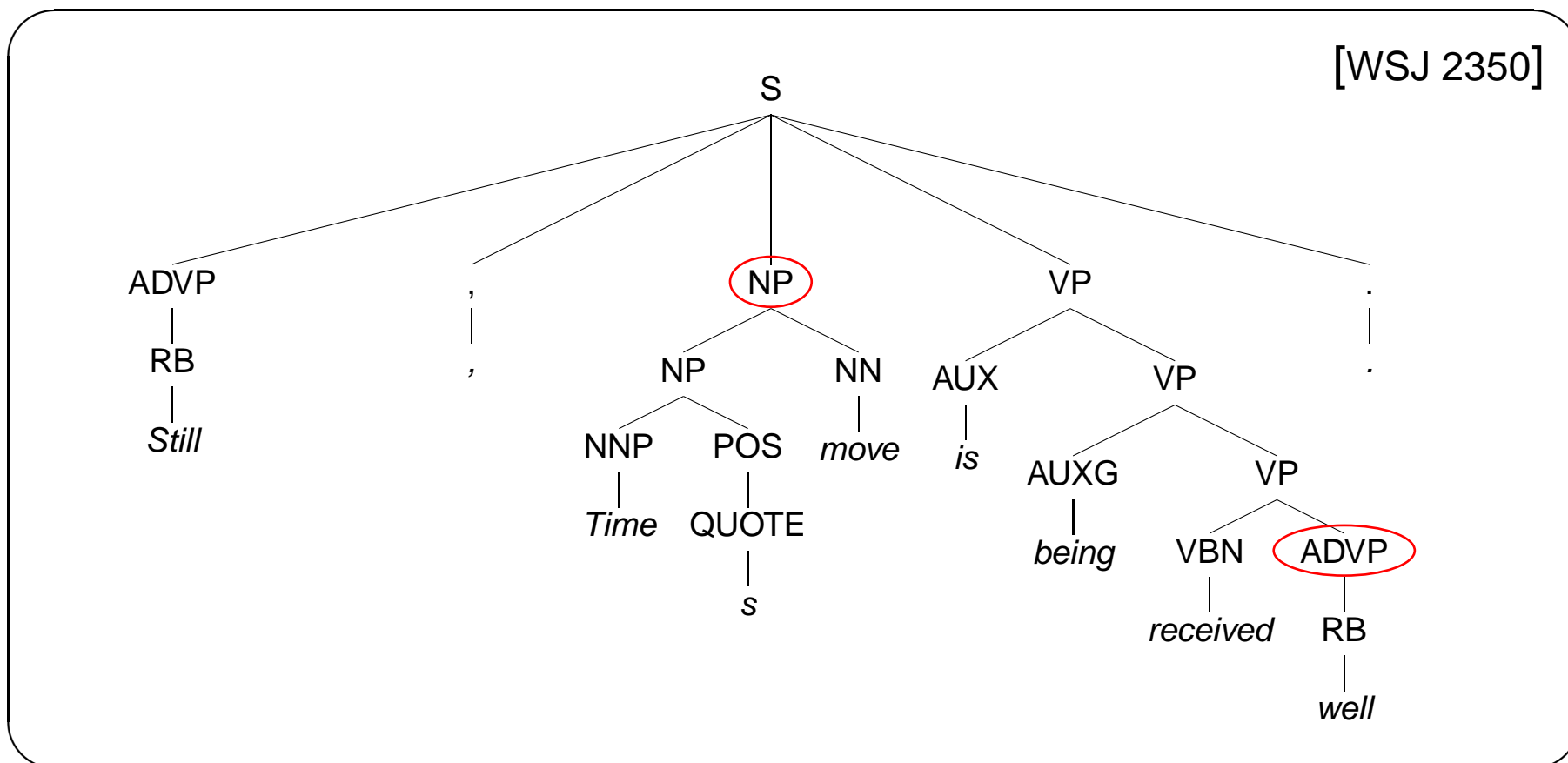
One Example from the Penn Treebank



Still, Time's move is being received well.



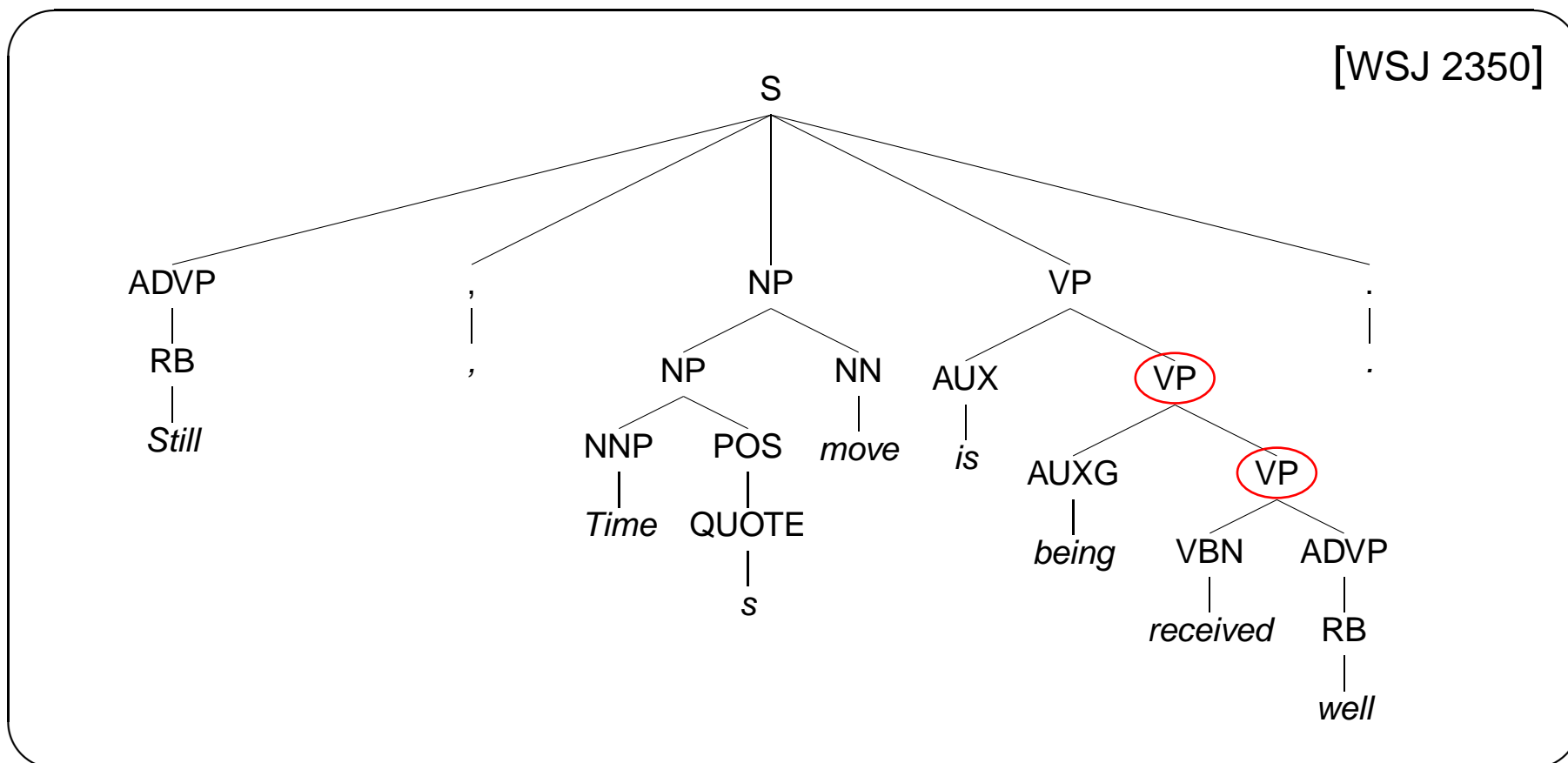
The Output from the Parser of Charniak (2000)



Still, Time's move is being received well.



The Output from the Parser of Charniak (2000)



Still, Time's move is being received well.



How to Evaluate (Syntactic) Parsing Accuracy?

ParsEval — Constituent Overlap (Black, et al., 1991)

- Break up tree into bracketing plus labelling, for example:

$\langle 0, 1, \text{ADVP} \rangle \langle 2, 5, \text{NP} \rangle \langle 5, 9, \text{VP} \rangle \langle 6, 9, \text{VP} \rangle \langle 0, 10, \text{S} \rangle$

- quantify precision (P) and recall (R) of labelled bracketings, when contrasting the gold-standard tree vs. the actual parser output;
- F Score, as the harmonic mean of precision and recall: $F_1 = \frac{2PR}{P+R}$;

→ combined with crossing brackets, dominant metric in PTB parsing.

Preliminary Reflections

- For our singleton example: $P = 0.8571$ (6/7), $R = 1$ — but so what?
- just one constituent error might obstruct correct analysis of passive.



Grumble, Grumble, Grumble — Known Issues

The Language

- WSJ is outdated, highly specialized, practically uninteresting data;
- WSJ parser on Brown Corpus: 0.80 instead of 0.86 F_1 (Gildea, 2001).

The Linguistics

- No (explicit) coding of heads; no argument vs. modifier distinction;
- no account of MWEs (e.g. verb – particles); flat NP structure; et al.

The Rules of the Game

- common parsers ignore a substantial part of the PTB annotations;
- Uninformative evaluation metric: high-frequency, ‘easy’ phenomena.



So, What Actually Characterizes 'Deep' Analysis?

- *Linguists write many papers.* vs. *They consume a lot of paper.*
- *the [annual [progress report]]* vs. *the [[annual meeting] report]*
- *Kim promised Sandy to sing.* vs. *Kim persuaded Sandy to sing.*
- *The window broke.* vs. *Kim broke the window.*
- *the belief that I will present to you today* — *pictures of Leonardo.*



So, What Actually Characterizes ‘Deep’ Analysis?

- *Linguists write many papers.* vs. *They consume a lot of paper.*
- *the [annual [progress report]]* vs. *the [[annual meeting] report]*
- *Kim promised Sandy to sing.* vs. *Kim persuaded Sandy to sing.*
- *The window broke.* vs. *Kim broke the window.*
- *the belief that I will present to you today* — *pictures of Leonardo.*

[...] The WSJ presents archaic, highly idiosyncratic data. WSJ parsing cannot inform us about many relevant linguistic phenomena. We should all stop using it.

(Adam Kilgariff, 2007)



So, What Actually Characterizes ‘Deep’ Analysis?

- *Linguists write many papers.* vs. *They consume a lot of paper.*
- *the [annual [progress report]]* vs. *the [[annual meeting] report]*
- *Kim promised Sandy to sing.* vs. *Kim persuaded Sandy to sing.*
- *The window broke.* vs. *Kim broke the window.*
- *the belief that I will present to you today* — *pictures of Leonardo.*

[...] The WSJ presents archaic, highly idiosyncratic data. WSJ parsing cannot inform us about many relevant linguistic phenomena. We should all stop using it.

(Adam Kilgarriff, 2007)

- Friendly competition towards theoretical and technological advances;
- think ‘deep’(er) vs. ‘shallow’(er) rather than CCG vs. LFG vs. HPSG.



Dependency Metrics (Carroll, Briscoe, Sanfilippo 1998)

Meridian will pay a premium of \$30.5 million to assume a deposit of \$2 billion.

(aux _ pay will)
(subj pay Meridian _)
(detmod _ premium a)
(mod _ million 30.5)
(mod _ \$ million)
(mod of premium \$)
(dobj pay premium _)
(mod _ billion 2)
(mod _ \$ billion)
(mod in \$ deposit)
(dobj assume \$ _)
(mod to pay assume)



Dependency Metrics (Carroll, Briscoe, Sanfilippo 1998)

Meridian will pay a premium of \$30.5 million to assume a deposit of \$2 billion.

(aux _ pay will)
(subj pay Meridian _)
(detmod _ premium a)
(mod _ million 30.5)
(mod _ \$ million)
(mod of premium \$)
(dobj pay premium _)
(mod _ billion 2)
(mod _ \$ billion)
(mod in \$ deposit)
(dobj assume \$ _)
(mod to pay assume)

GR: Some Advantages and Disadvantages

- + Annotated 'gold' targets freely available;
- + hierarchical inventory of relation labels;
- + resembles dependency graph → control;
- + 'common-sense' grammatical relations;
- + a handful of careful studies published;
- missing information (e.g. number, tense);
- at times at odds with semantic structure;
- GR mapping often all but straightforward;
- very little data; few annotation guidelines.



Beyond the Wall Street Journal?

High-Level Goals

- WSJ competition not a level playing field for general-purpose parsers;
 - tighten relationship between parsing and 'text understanding' tasks;
- among interested parties, establish an alternate parsing benchmark.



Beyond the Wall Street Journal?

High-Level Goals

- WSJ competition not a level playing field for general-purpose parsers;
 - tighten relationship between parsing and 'text understanding' tasks;
- among interested parties, establish an alternate parsing benchmark.

WeScience: Scholarly Literature in Our Own Field

- Wikipedia provides thousands of articles on linguistics, NLP, et al.
 - candidate applications: question answering, knowledge acquisition;
- + compile high-quality, scholarly corpus; what size needed for whom?
- + parallel PTB, GR, CCG, LFG, MRS, PropBank, et al. annotations;
- joint search towards semantic consensus through *community effort*.



Wikipedia — A Few Appetizers

Machine translation, sometimes referred to by the acronym MT, is a sub-field of computational linguistics that investigates the use of computer software to translate text or speech from one natural language to another. [...] The translation process may be stated as:

1. Decoding the meaning of the source text; and
2. Re-encoding this meaning in the target language.

Behind this ostensibly simple procedure lies a complex cognitive operation.

Martin Kay is a computer scientist known especially for his work in computational linguistics. [...] His achievements include the development of chart parsing and functional unification grammar and major contributions to the application of finite state automata in computational phonology and morphology. He is also regarded as a leading authority on machine translation.

+ Besides content relevance, Wikipedia is a rich, *unadulterated* hyper-text.



A Quick Glimpse at PropBank (Palmer, et al., 2005)

Georgia Gulf received a new takeover bid from investor Harold Simmons of \$50 million a share.

ARG0	:	<i>Georgia Gulf</i>
REL	:	<i>received</i>
ARG1	:	<i>a new takeover bid</i>
ARG2-from	:	<i>investor Harold Simmons</i>
ARG4-of	:	<i>\$50 million a share</i>



A Quick Glimpse at PropBank (Palmer, et al., 2005)

Georgia Gulf received a new takeover bid from investor Harold Simmons of \$50 million a share.

ARG0	:	<i>Georgia Gulf</i>
REL	:	<i>received</i>
ARG1	:	<i>a new takeover bid</i>
ARG2-from	:	<i>investor Harold Simmons</i>
ARG4-of	:	<i>\$50 million a share</i>

The attribute argument is necessary because of the way Treebank divides the constituent. Really, the ARG1 of the sentence above is ‘a new takeover bid of \$50 million a share’, but the ‘giver’ argument intervenes between the two, necessitating the additional argument.



Summing Up — No Conclusions

Some Lessons Learned

- PTB and WSJ ‘parsing game’ have been vital scientific catalysts;
- many core statistical NLP techniques out of PTB-based research;
- (dependency) parsing (for other languages) yields a similar picture;
- linguistically advanced broad-coverage ‘semantic’ parsers now becoming available for a variety of languages (e.g. CCG, LFG, HPSG);
 - lack of adequate gold-standard resources for training and evaluation;
 - PTB idiosyncrasies perpetuate into ongoing resource development.



Really, The Gist of This Talk

Open Questions

- ? *selection* synchronize collection of *relevant & high-quality* corpora;
- ? *representations* informative & fair for *both* syntax *and* semantics;
- ? *annotation* manual (linguistic undergraduates) vs. semi-automated;
- ? *analysis* interplay of grammar, WSD, co-reference resolution, et al.
- ? *applications* resource development & evaluation \leftrightarrow NLP needs;
- ? re-energize friendly dialogue with syntacticians, semanticists, et al.
- ?
- ?

