DerivTool 0.6

New Corpus | Corpus=f-arabic.plain (81 lines) 48 | Go to line | Load | Save | Print | Print to File | Redraw | Close

Derivation Tree | Freeform Notes

"وقال"

IN("that")
-> "ان"

S-C(x0:NP-C VP(x1:MD VP-C(x2:VB x3:NP-C)))
-> x0 x1 x2 x3

.(".")
-> "."

"ان"

English: the process of decision would take six months

NP-C(x0:NPB x1:PP)
-> x0 x1

MD("would")
-> "ان"

VB("take")
-> "تستغرق"

NP-C(NPB(CD("six") NNS("months")))
-> "اشهر" "ستة"

"."

NPB(DT("the") NN("process"))
"العملية" ->

PP(IN("of") NP-C(NPB(NN("decision"))))
-> "القرار" "من"

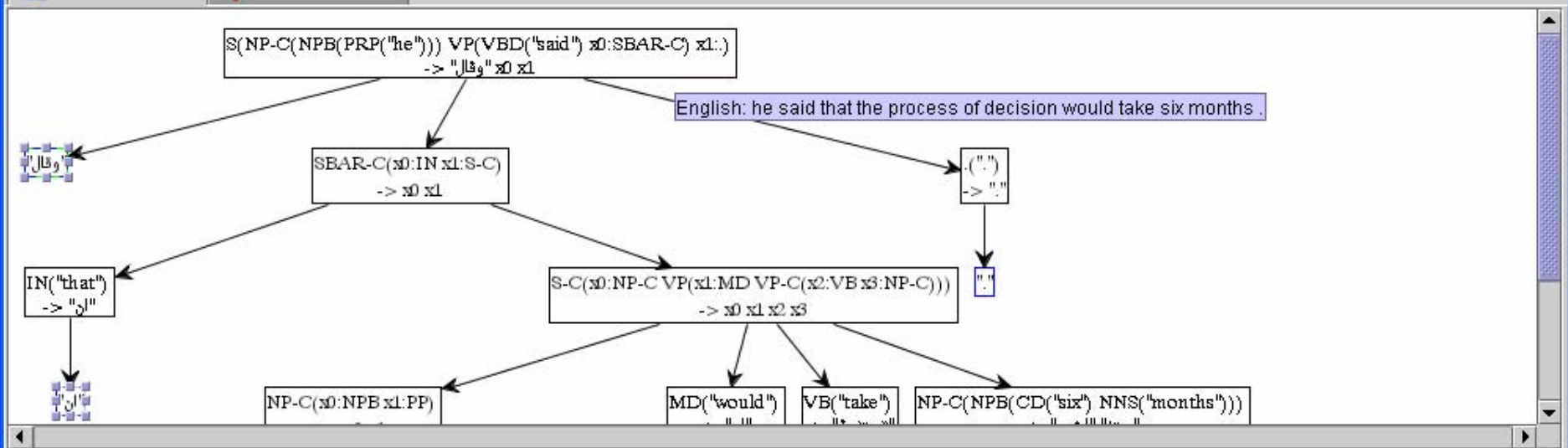"ان"

"تستغرق"

"ستة"

"اشهر"

"العملية"

"من"

"القرار"

Click, Ctrl-click, or drag above to select. For adding rules, please select a contiguous span of top level nodes. | Modify | Delete

Phrase-based MT: he said that the process of decision to take six months .

Searching | Manual | **Selecting Template**

Lines = 100 | Redraw | Click below to add rule. | (Red=no rules found, Blue=used by AT translation, Green=default translation, Purple=Red+Blue)

"وقال"	"ان"	"العملية"	"من"	"القرار"	"ان"	"تستغرق"	"ستة"	"اشهر"
he said	that the process is		which would			take		four months
he		practical		would		takes		six weeks
that		the process		is to be		last		six months ' time
he		the operation		were to		took		six months ,
he said		operation		is		lasting	of six		@-@ mon
was		exercise	from	scheduled to		would take		month
had		operational		planned that		span		six @-@ months '
his		an operation		be		last for		6 month
while		the practical		the rapporteur		taken		of six months from

start | derivtool arabic | DerivTool 0.6 | Microsoft PowerPoint ... | 7:55 PM

DerivTool 0.6

New Corpus | Corpus=f-arabic.plain (81 lines) | 48 | Go to line | Load | Save | Print | Print to File | Redraw | Close

Derivation Tree | Freeform Notes

S(NP-C(NPB(PRP("he"))) VP(VBD("said") x0:SBAR-C) x1:.)
-> "وقال" x0 x1

English: he said that the process of decision would take six months .

SBAR-C(x0:IN x1:S-C)
-> x0 x1

.(".")
-> "."

IN("that")
-> "ان"

S-C(x0:NP-C VP(x1:MD VP-C(x2:VB x3:NP-C)))
-> x0 x1 x2 x3

"."

"ان"

NP-C(x0:NPB x1:PP)

MD("would")

VB("take")

NP-C(NPB(CD("six") NNS("months")))

Click, Ctrl-click, or drag above to select. For adding rules, please select a contiguous span of top level nodes. | Modify | Delete

Phrase-based MT: he said that the process of decision to take six months .

Searching | Manual | Selecting Template

Lines = 100 | Redraw | Click below to add rule. (Red=no rules found, Blue=used by AT translation, Green=default translation, Purple=Red+Blue)

"وقال"	"ان"	"العملية"	"من"	"المقرر"	"ان"	"تستغرق"	"سعة"	"اشهر"
he said	that the process is			which would		take	four months	
he		practical		would		takes	six weeks	
that		the process		is to be		last	six months ' time	
he		the operation		were to		took	six months ,	
he said		operation		is		lasting	of six	@-@ mor
was		exercise	from	scheduled to		would take		month
had		operational		planned that		span	six @-@ months '	
his		an operation		be		last for	6 month	
while		the practical		the rapporteur		taken	of six months from	

start | derivtool arabic | DerivTool 0.6 | derivtool | 7:53 PM

input

dev-little (line 47) - dev-little

Input: 中 资 已 成为 澳门 最 大的 外来 投资者 。
Reference: the chinese enterprises have become the biggest outside investors in macao .
AlTemp-e: investment $_0$ | in $_1$ | macao $_2$ | has $_3$ | become $_4$ | the largest $_5$ | foreign $_6$ | investors $_7$ | . $_8$
AlTemp-f: 中 $_1$ | 资 $_0$ | 已 $_3$ | 成为 $_4$ | 澳门 $_2$ | 最 大的 $_5$ | 外来 $_6$ | 投资者 $_7$ | 。 $_8$
[dev-little] 1-Best: the chinese - funded enterprises have become the largest external investor in macao .

phrase-based
system output

[dev-little] 1-Best Tree
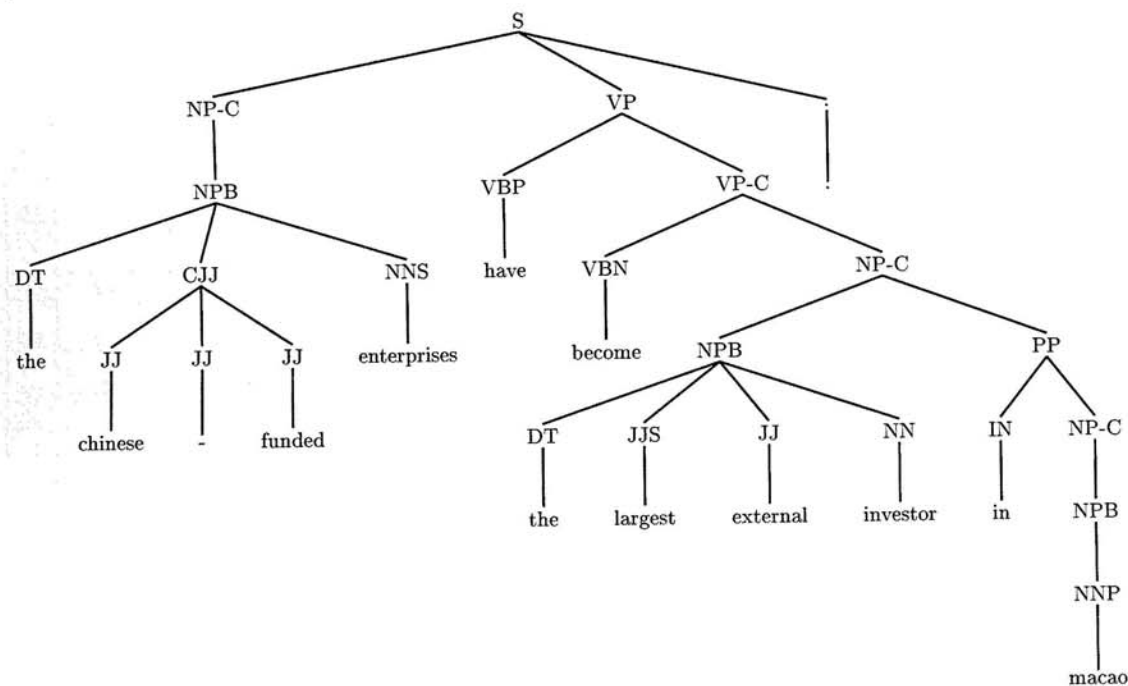
syntax-based
system output

input

dev-little (line 59) - dev-little

Input: 基 纳 纳 对 中国 过去 向 坦 桑 提供 的 大量 援助 表示 感谢 。

Reference: keenana expressed gratitude to china for its great assistance to tanzania in the past .

AlTemp-e: kinana $_0$ | to $_1$ | china $_2$ | in the past to $_3$ | tanzania $_4$ | expressed appreciation $_5$ | for the substantial $_6$ | assistance $_7$ | . $_8$

AlTemp-f: 基 纳 纳 $_0$ | 对 $_1$ | 中国 $_2$ | 过去 向 $_3$ | 坦 桑 $_4$ | 提供 的 大量 $_6$ | 援助 $_7$ | 表示 感谢 $_5$ | 。 $_8$

[dev-little] 1-Best: kinana thanked china for substantial assistance to tanzania in the past .
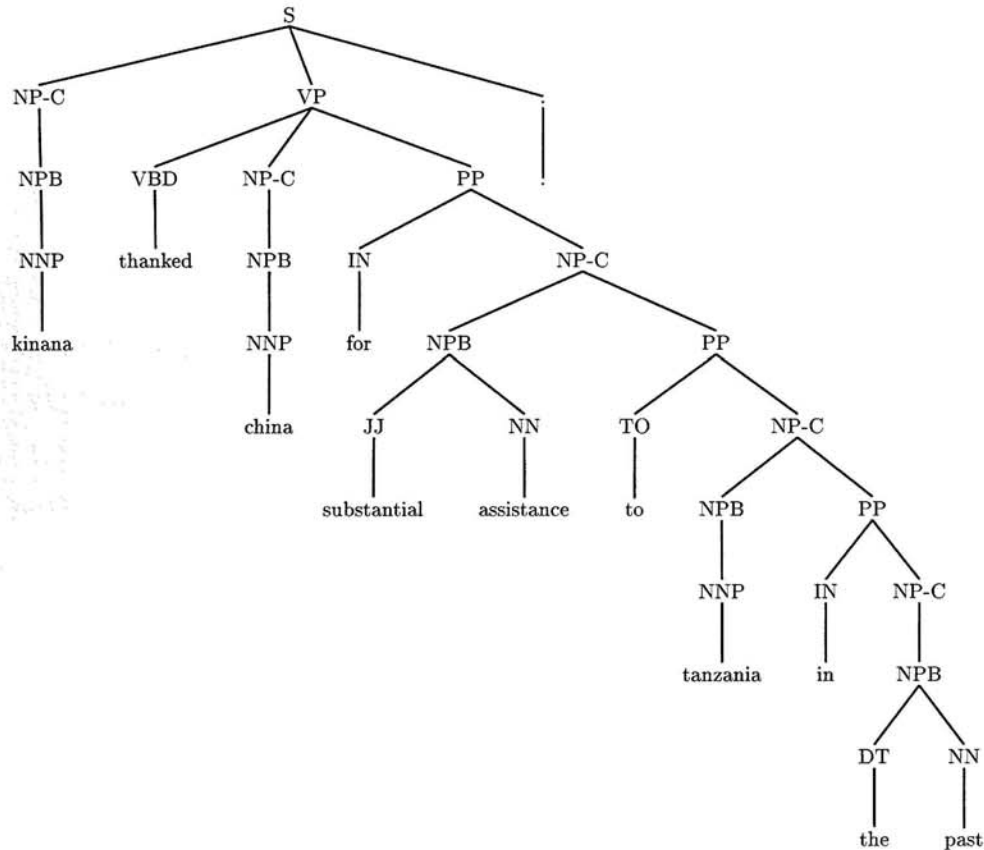
phrase-based system output

[dev-little] 1-Best Tree

syntax-based system output

dev-little (line 38) - dev-little

Input: 此次 为期 两天 的 研讨会 ， 由 世界贸易组织 上海 研究中心 与 上海市 对外 服务 有限公司 联合 举办 。

Reference: the two - day seminar is jointly sponsored by the wto shanghai research center and shanghai foreign service company limited .

AlTemp-e: the $_0$ | two - day $_1$ | seminar $_2$ | by the world trade organization $_3$ | , $_4$ | shanghai research center $_5$ | and $_6$ | shanghai foreign service $_7$ | co . , ltd . $_8$ | jointly $_9$ | . $_{10}$

AlTemp-f: 此次 $_0$ | 为期 两天 的 $_1$ | 研讨会 $_2$ | , $_4$ | 由 世界贸易组织 $_3$ | 上海 研究中心 $_5$ | 与 $_6$ | 上海市 对外 服务 $_7$ | 有限公司 $_8$ | 联合 举办 $_9$ | 。 $_{10}$

[dev-little] 1-Best: the two - day seminar , organized jointly by shanghai foreign services ltd . and shanghai research center of the world trade organization .
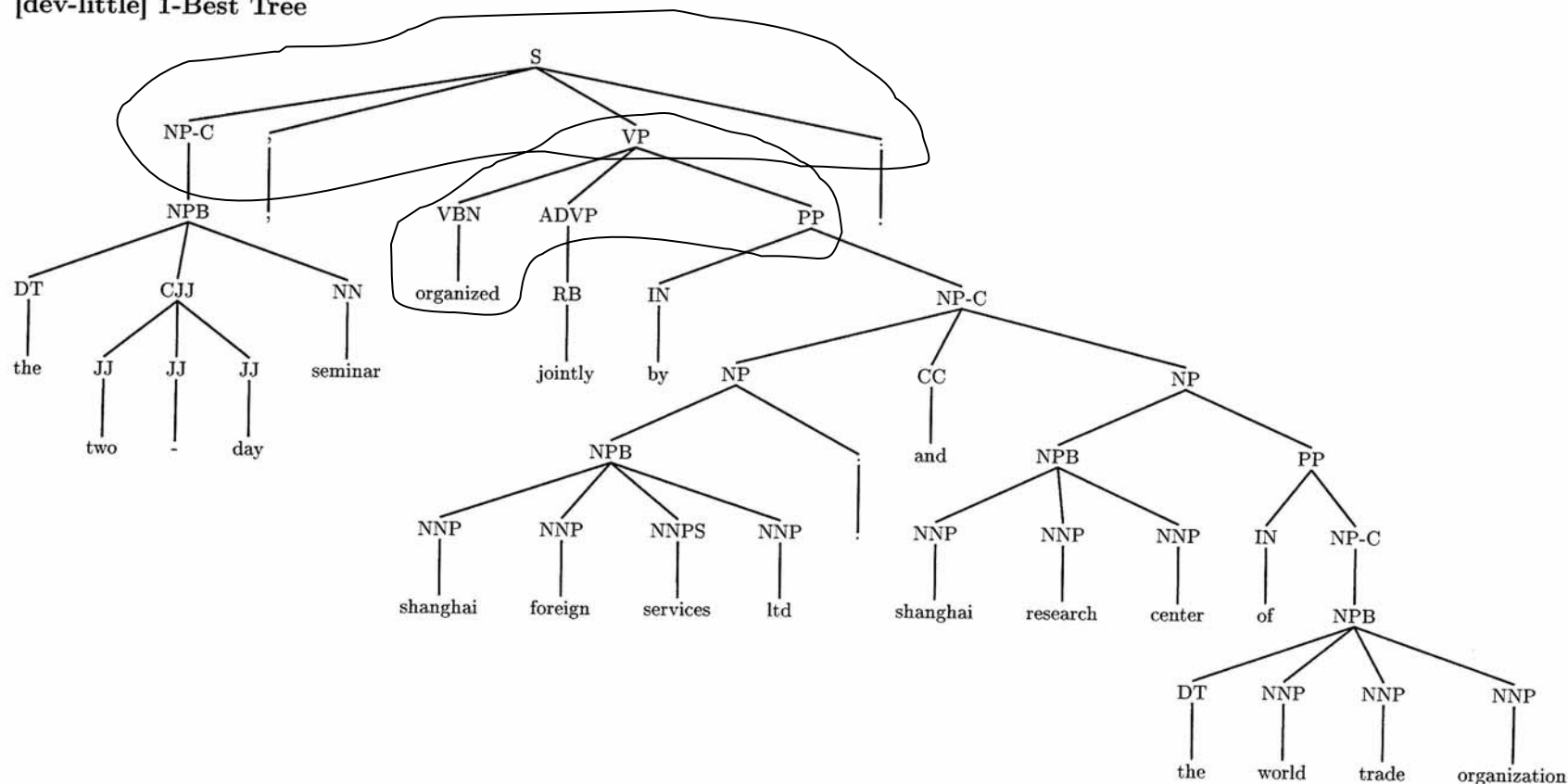
[dev-little] 1-Best Tree

**Input:** 丁 豪 在 儿童 福利 院 读 完 小学 ， 随后 进入 附近 乡里 一所 学校 上 初中 。

Reference: ding hao completed his primary school at the children welfare school , and then went to a nearby township middle school .

AlTemp-e: ding hao $_0$ | in $_1$ | children 's welfare institute $_2$ | complete primary school $_3$ | , $_4$ | then entered $_5$ | commune $_6$ | near $_7$ | a school $_8$ | , $_9$ | junior high school level . $_{10}$

AlTemp-f: 丁 豪 $_0$ | 在 $_1$ | 儿童 福利 院 $_2$ | 读 完 小学 $_3$ | ， $_4$ | 随后 进入 $_5$ | 附近 $_7$ | 乡里 $_6$ | 一所 学校 $_8$ | 上 $_9$ | 初中 。 $_{10}$

dev-little] 1-Best: ding hao in children 's homes finishing elementary school and was subsequently entered junior high school at a school near their home towns .

dev-little] 1-Best Tree
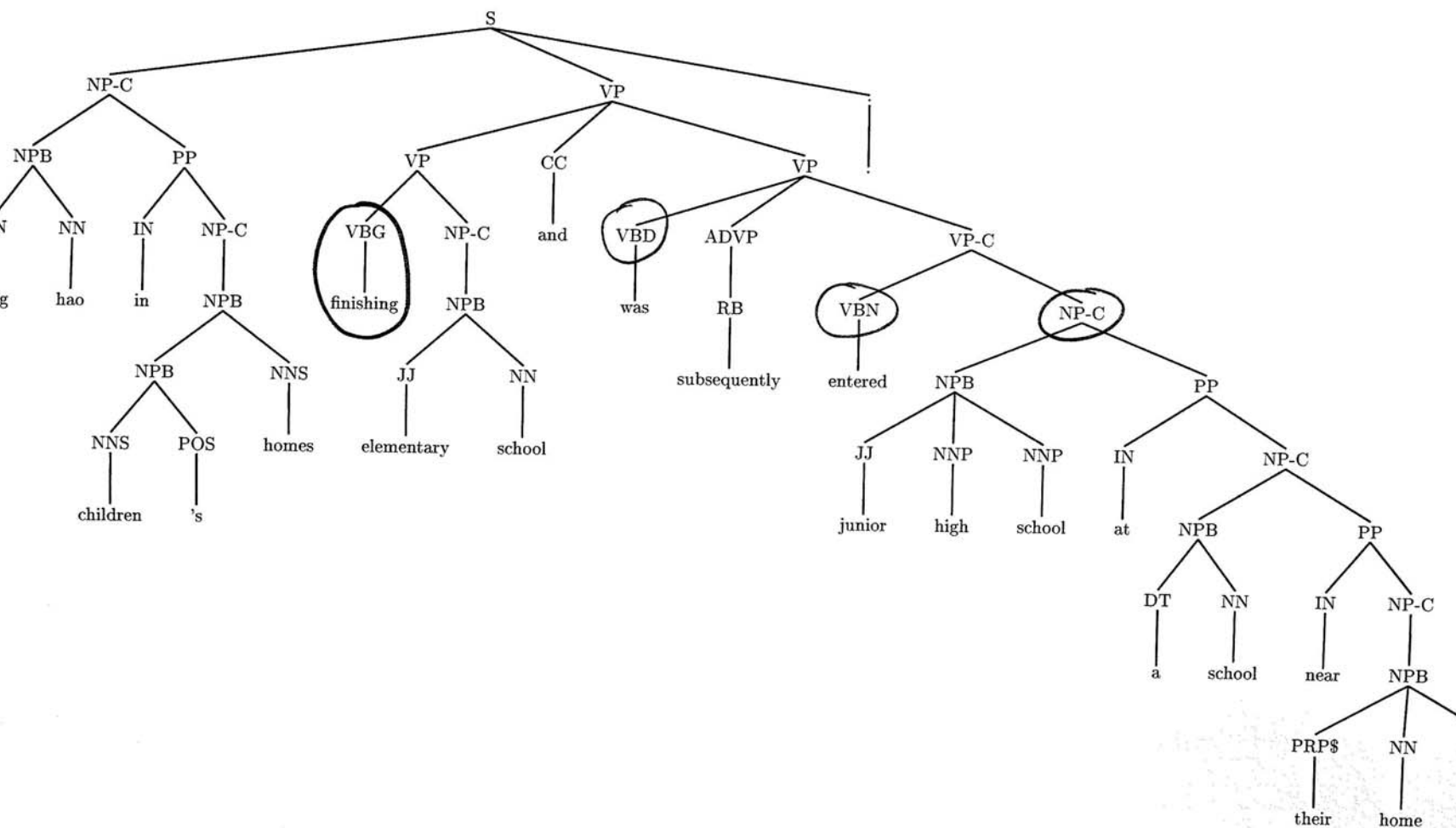
dev-little (line 64) - dev-little

con become very good partes

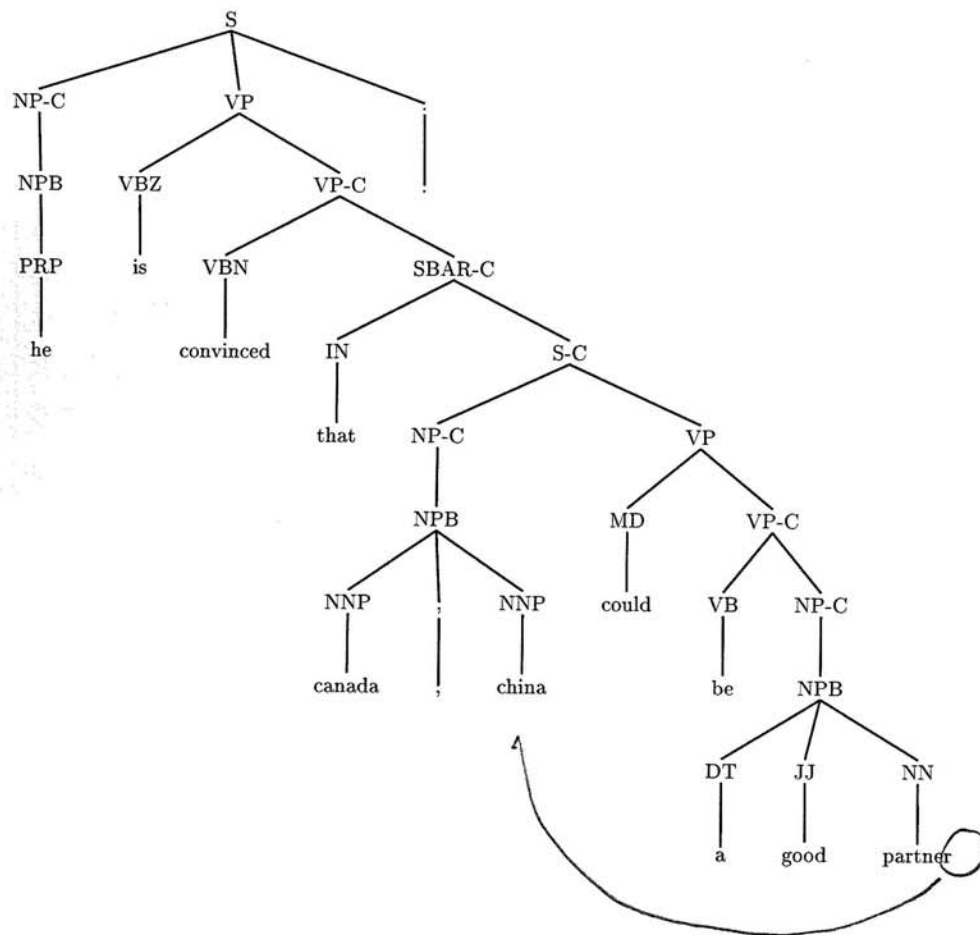**Input:** 他 确信 ， 加 、 中 两国 可以 成为 很 好的 合作 伙伴 。
Reference: he assured that canada and china can become very good partners .
AlTemp-e: he was convinced that $_0$ | the $_1$ | two countries $_2$ | , $_3$ | can $_4$ | become good $_5$ | partners $_6$ | . $_7$
AlTemp-f: 他 确信 ， $_0$ | 加 $_1$ | 、 中 $_3$ | 两国 $_2$ | 可以 $_4$ | 成为 很 好的 $_5$ | 合作 伙伴 $_6$ | 。 $_7$
[dev-little] 1-Best: he is convinced that canada , china could be a good partner .

[dev-little] 1-Best Tree

Subj - obj number agreement.

dev-little (line 51) - dev-little

**Input:** 法国 外长 昨天 是 在 法国 国民议会 外事 委员会 会议 上 发表 上述 声明 的 。

Reference: the french foreign minister made the above statement in a meeting of the foreign affairs commission of the french national congress .

AlTemp-e: french $_0$ | foreign minister $_1$ | in the french national assembly $_2$ | yesterday , $_3$ | the statement delivered by $_4$ | foreign affairs $_5$ | committee meeting $_6$ | . $_7$

AlTemp-f: 法国 $_0$ | 外长 $_1$ | 昨天 是 $_3$ | 在 法国 国民议会 $_2$ | 外事 $_5$ | 委员会 会议 上 $_6$ | 发表 上述 声明 $_4$ | 的 。 $_7$

[dev-little] 1-Best: the french foreign minister was the statement issued at the french national assembly foreign affairs committee meeting yesterday .

[dev-little] 1-Best Tree
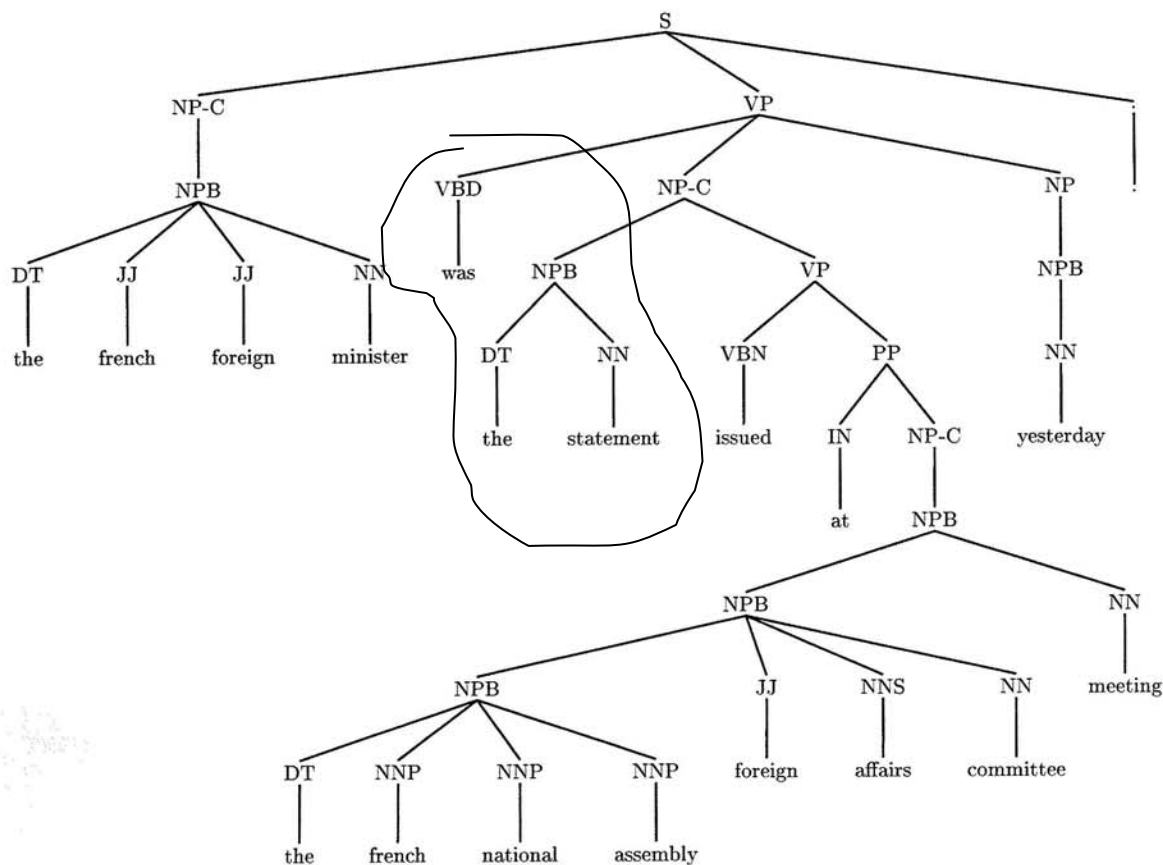
dev-little (line 248) - dev-little

|	
Input:	国防部 前天 对 导弹 发射 事故 作出 道歉 。
Reference:	the defense ministry apologized for the missile accident the day before yesterday .
AlTemp-e:	the day before yesterday $_0$ | the ministry of defense $_1$ | against $_2$ | missile $_3$ | launching accidents $_4$ | make an apology . $_5$
AlTemp-f:	国防部 $_1$ | 前天 $_0$ | 对 $_2$ | 导弹 $_3$ | 发射 事故 $_4$ | 作出 道歉 。 $_5$
[dev-little] 1-Best:	defence apologized to the missile launch accident the day before yesterday .

[dev-little] 1-Best Tree
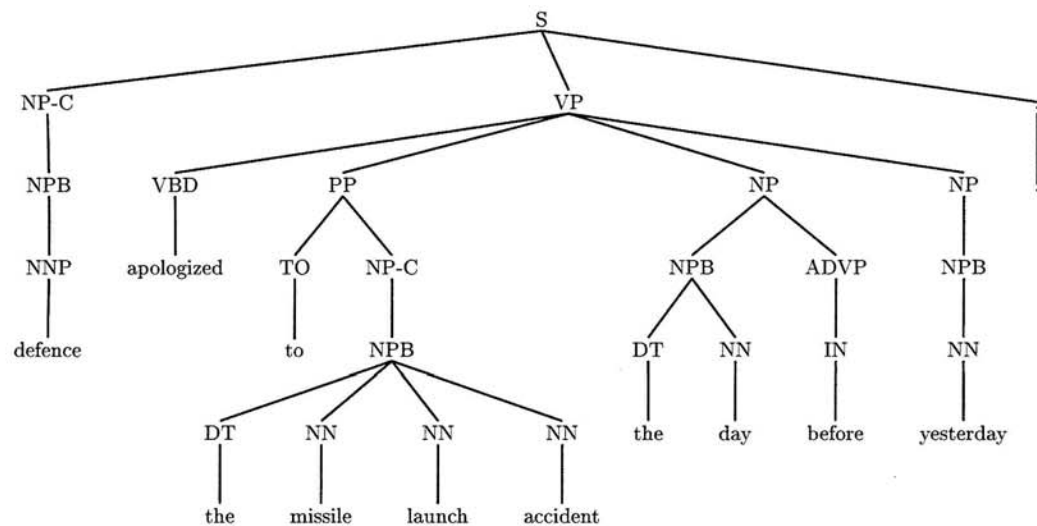
dev-little (line 125) - dev-little

rain

Input: 今年 在 加利福尼亚 州 和 南部 地区 (的) 豪 雨 都 归咎 于 厄尔尼诺 作 宠 。

Reference: the torrential rain this year in california and its southern part is attributed to the el nino .

AlTemp-e: this year $_0$ | in california $_1$ | and southern $_2$ | areas $_3$ | of heavy rains $_4$ | attributed to $_5$ | a $_6$ | favorite $_7$ | el nio $_8$ | . $_9$

AlTemp-f: 今年 $_0$ | 在 加利福尼亚 州 $_1$ | 和 南部 $_2$ | 地区 $_3$ | 的 豪 雨 $_4$ | 都 归咎 于 $_5$ | 厄尔尼诺 $_8$ | 作 $_6$ | 宠 $_7$ | 。 $_9$

[dev-little] 1-Best: this year made love in california and southern areas of torrential rains were blamed on el nino .

[dev-little] 1-Best Tree

Scope.
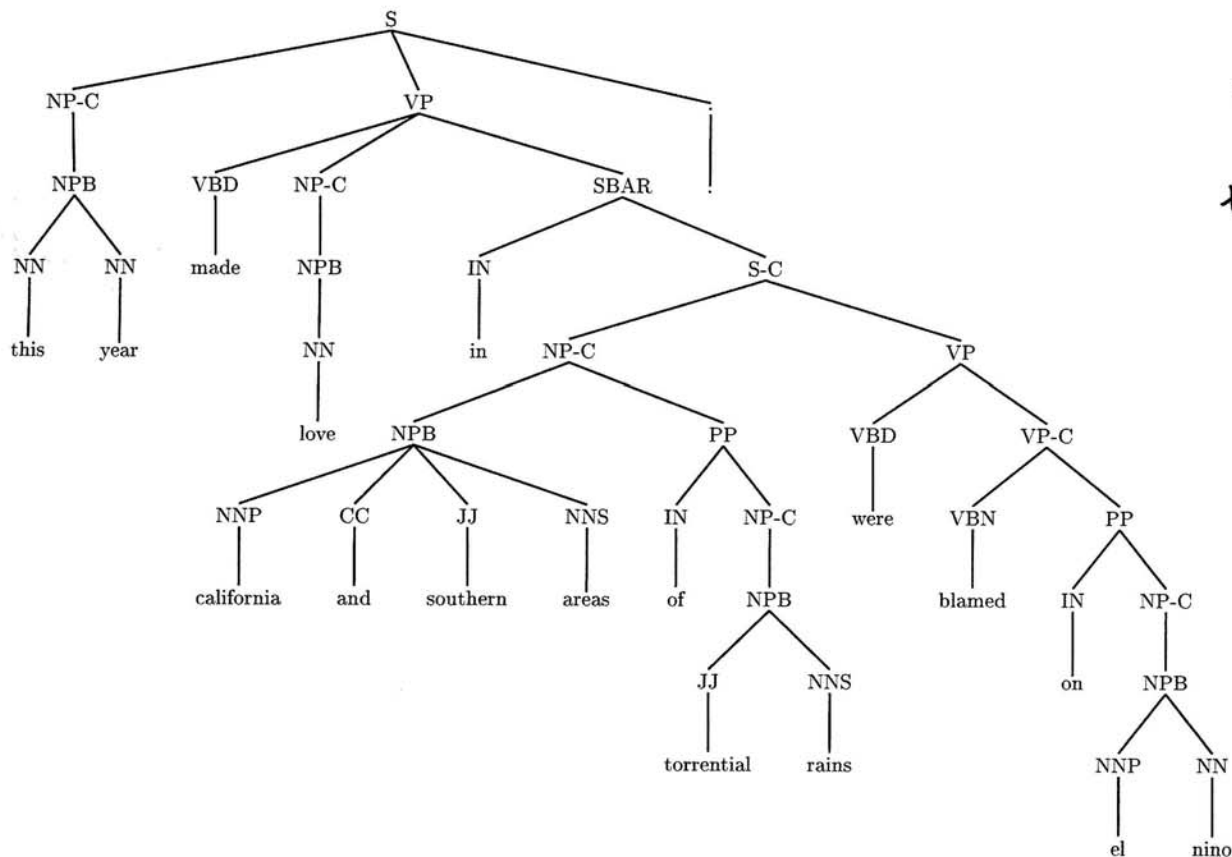
funny.

this year realy NP-C

Lots of Open Problems

- Choosing syntactic categories that are appropriate for translation

- Decoder search in runtime translation
  - Many search errors hurt accuracy
  - Much faster speed needed for experimentation

- More context for rule choice

- Modeling English fluency, using trees

Chomsky's Program [1957]

- **Algorithmically distinguish between grammatical and ungrammatical sentences:**
  - John thinks Sara hit the boy
  - * The hit thinks Sara John boy
  - John thinks the boy was hit by Sara
  - Who does John think Sara hit?
  - John thinks Sara hit the boy and the girl
  - * Who does John think Sara hit the boy and?
  - John thinks Sara hit the boy with the bat
  - What does John think Sara hit the boy with?
  - Colorless green ideas sleep furiously.
  - * Green sleep furiously ideas colorless.

This Research Program has Contributed Powerful Ideas

Context-free grammar

Formal language hierarchy

Syntax, Phonology…

This Research Program is Really Unfinished

Type in your English sentence here:

```
[                                                            ]
```

Is this grammatical?

Is this sensible?

immediate consequences for machine translation and life in general …
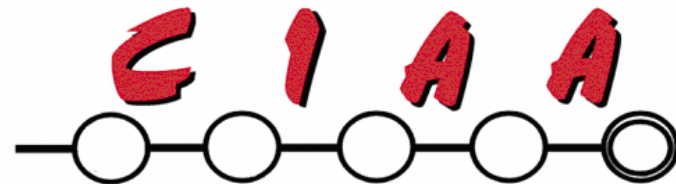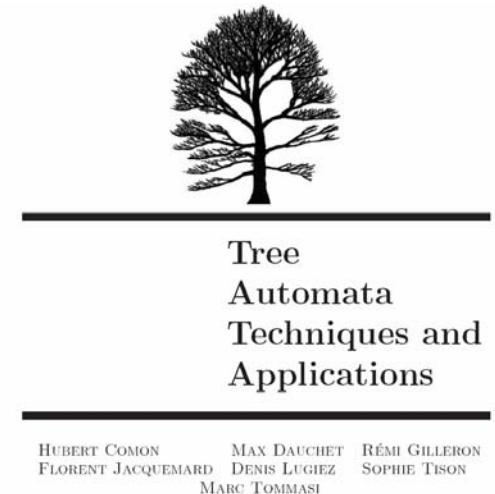
Big World of Tree Automata Theory

Doner (1968), Rounds (1970), Thatcher (1970), Engelfriet thesis (1975), Gecseg & Steinby textbook (1984), …

Workshop on Tree Automata

Magnus Steinby

Tree Automata Techniques and Applications

HUBERT COMON MAX DAUCHET RÉMI GILLERON
FLORENT JACQUEMARD DENIS LUGIEZ SOPHIE TISON
MARC TOMMASI

Heiko Vogler

CIAA
Conference on Implementation and Application of Automata

Tiburon: A Tree Automata Toolkit

- Developed by Jonathan May, ISI
- First version distributed in April 2006
- Inspired by string automata toolkits

- You cast your problem in terms of tree acceptors and transducers
  - doesn't have to be MT

- You get implemented algorithms for free
  - e.g., as Kumar/Byrne'03 and Pereira/Riley'96 use string automata toolkits for MT and ASR

- Wealth of tree automata literature to drawn on
- Lots of open problems in tree automata and in choosing formalisms for modeling NLP

Tiburon: A Tree Automata Toolkit

	String World	Tree World
N-best …	… paths through a lattice (Viterbi, 1967; Eppstein, 1998)	… trees in a forest (Huang & Chiang, 2005)
EM training	Forward-backward EM (Baum & Welch, 1971)	Tree transducer EM training (Graehl & Knight, 2004)
Determinization …	… of weighted string acceptors (Mohri, 1997)	… of weighted tree acceptors (May & Knight, 2005)
Intersection	WFSA intersection	Tree acceptor intersection (despite CFG not closed)
Applying transducers	string → WFST → WFSA	tree → TT → weighted tree acceptor
Transducer composition	WFST composition (Pereira & Riley, 1996)	Many tree transducers are not closed under composition! (Rounds 70; Engelfriet 75; Graehl, Hopkins, Knight, sub.)

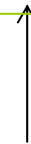# Classes of
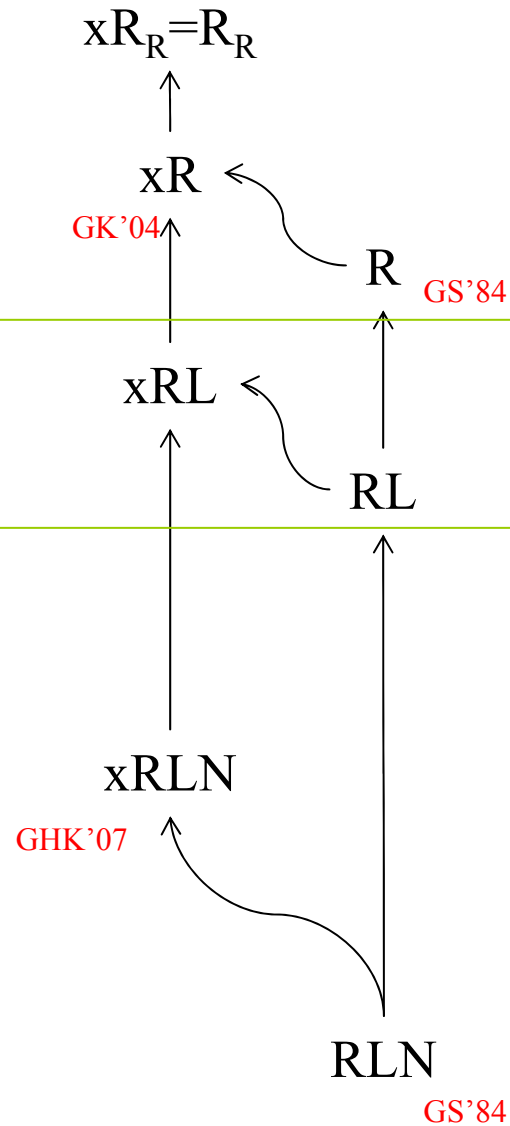# Tree Transducers

copying

non-copying

deleting

non-deleting

$$R$$

$$RL$$

$$RLN$$

Classes of Tree Transducers

$$xR_R = R_R$$

xR

GK'04

R GS'84

copying

non-copying

xRL

RL

deleting

non-deleting

xRLN

GHK'07

RLN

GS'84

Expressive power theorems in Graehl, Hopkins, Knight (submitted)

Classes of Tree Transducers

$xR_R = R_R$

R^3 F^3

xR R^2 F^2

GK'04

R   F

GS'84

bottom up transducers

copying

non-copying

xRL

$FL = RL_R$

RL

deleting

non-deleting

$xRLN$

GHK'07

RLN

GS'84

Classes of Tree Transducers

$xR_R = R_R$

R^3 F^3

R^2 F^2

xR

GK'04

R   GS'84   F

copying

non-copying

xRL

FL=RL$_R$

RL

deleting

non-deleting

GK'04

xRLN
(xRHS, input-e,
output-e)

xRLN
(xRHS, output-e)

xRLN
(xRHS, input-e)

RLN
(xRHS, input-e,
output-e)

GHK'07
GHKM'04

xRLN
(xRHS, e-free)

RLN
(xRHS, input-e)

RLN
(xRHS, output-e)

GS'84

RLN
(xRHS, e-free)

(Knight & Graehl, subm.)

Research Synergy

New applications,
prototypes,
projects…

Clean representations,
fast algorithms

Clean representations,
fast algorithms

Machine Translation Research	Tree Automata Research	Tree Automata Software

Challenges coming
from empirical data

User feedback

This Is Interdisciplinary Research

- Machine Learning
- Engineering
- Linguistics
- Efficient search algorithms
- Automata theory
- Grid computing

…

~~the end~~

still the beginning!