HPSG文法開発 (後半)

二宮　崇

今日の講義の予定

- 文法開発 (後半)
  - 文法開発の再解釈と展望
  - コーパス指向文法
- 教科書
  - Yusuke Miyao (2006) From Linguistic Theory to Syntactic Analysis: Corpus-Oriented Grammar Development and Feature Forest Model, Ph.D Thesis, University of Tokyo

文法開発の再解釈と展望

合理主義的文法の文法開発の難しさ

- さて、いったい何が難しくて文法開発がうまくいかなかったのだろうか？
- どこに落とし穴があったのか？

比較検討(1/2)

- 経験主義的文法開発と合理主義的文法開発の違い

	文法作成方法	コーパスの役割	評価手段
合理主義的文法	人手	生コーパス 補助的リソース	定性的評価
経験主義的文法	自動	ツリーバンク 中心的リソース	定量的評価

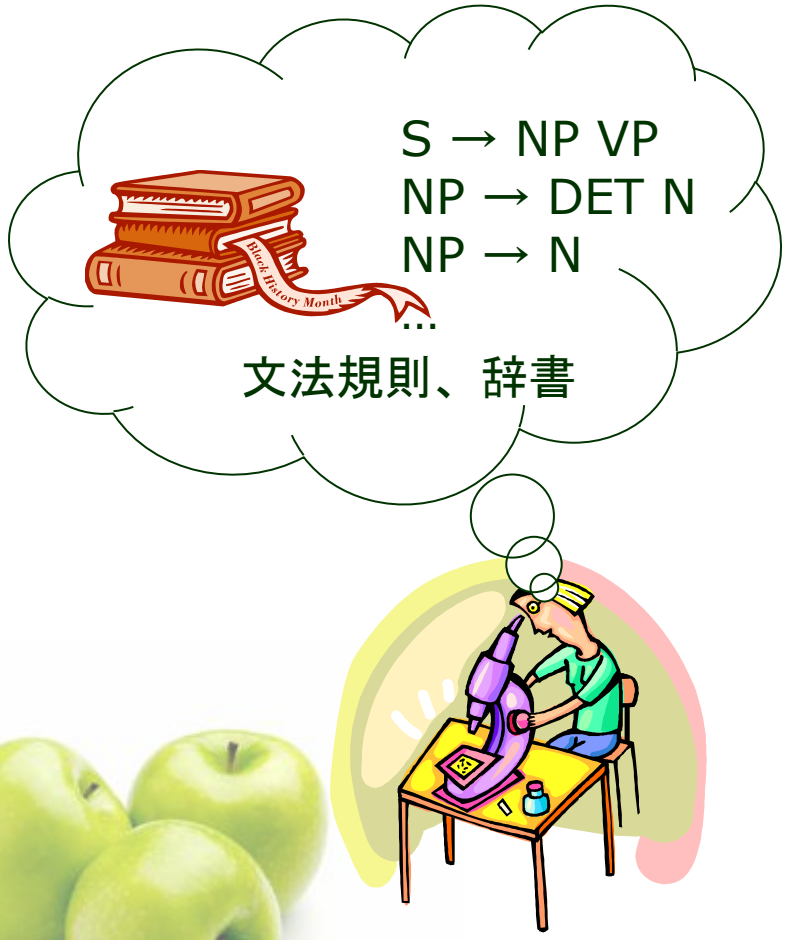
比較検討(2/2)

- 言語学者と言語処理研究者が求める文法、コーパスの役割の違い

	目的	文法	コーパス
言語学者	言語能力の法則性の発見	適格文、非文を区別するために必要な規則	人間の言語能力を調べるための資料
言語処理研究者	応用システムに有用な構文構造の自動解析	コーパスを解析するための道具	機械学習・統計学習のためのリソース。性能評価のためのリソース

合理主義的文法開発の落とし穴 (1/2)

- コーパス軽視



文法開発の対象は、文法規則と辞書。
コーパスはあくまで補助的な検証の対象にすぎない

合理主義的文法開発の落とし穴 (2/2)

- 定量的評価の不足
  - ツリーバンクの作成が困難
  - 文法を変更するとその都度正解が変化

Penn Treebankのようなツリーバンクに対して評価すれば？

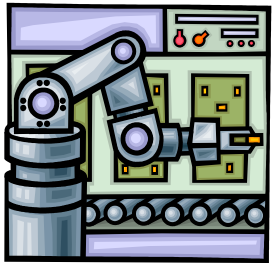
Penn Treebankにおける構文木の解釈と文法開発者の構文木の解釈が異なるため、Penn Treebankで評価するのは文法を開発するのに匹敵するほど困難

経験主義的文法と合理主義的文法の歩み寄り

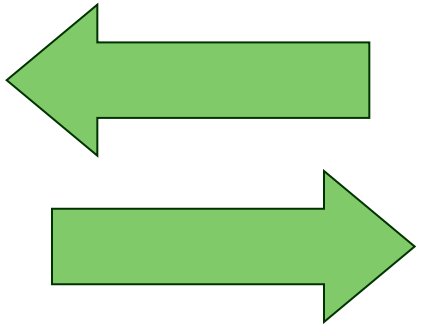
経験主義的文法

合理主義的文法

コンピュータ

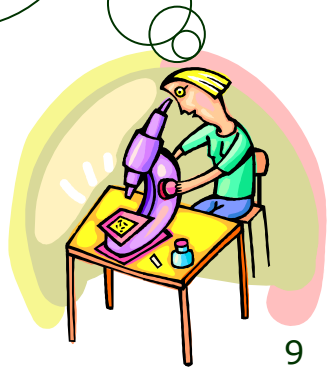


・ツリーバンク開発



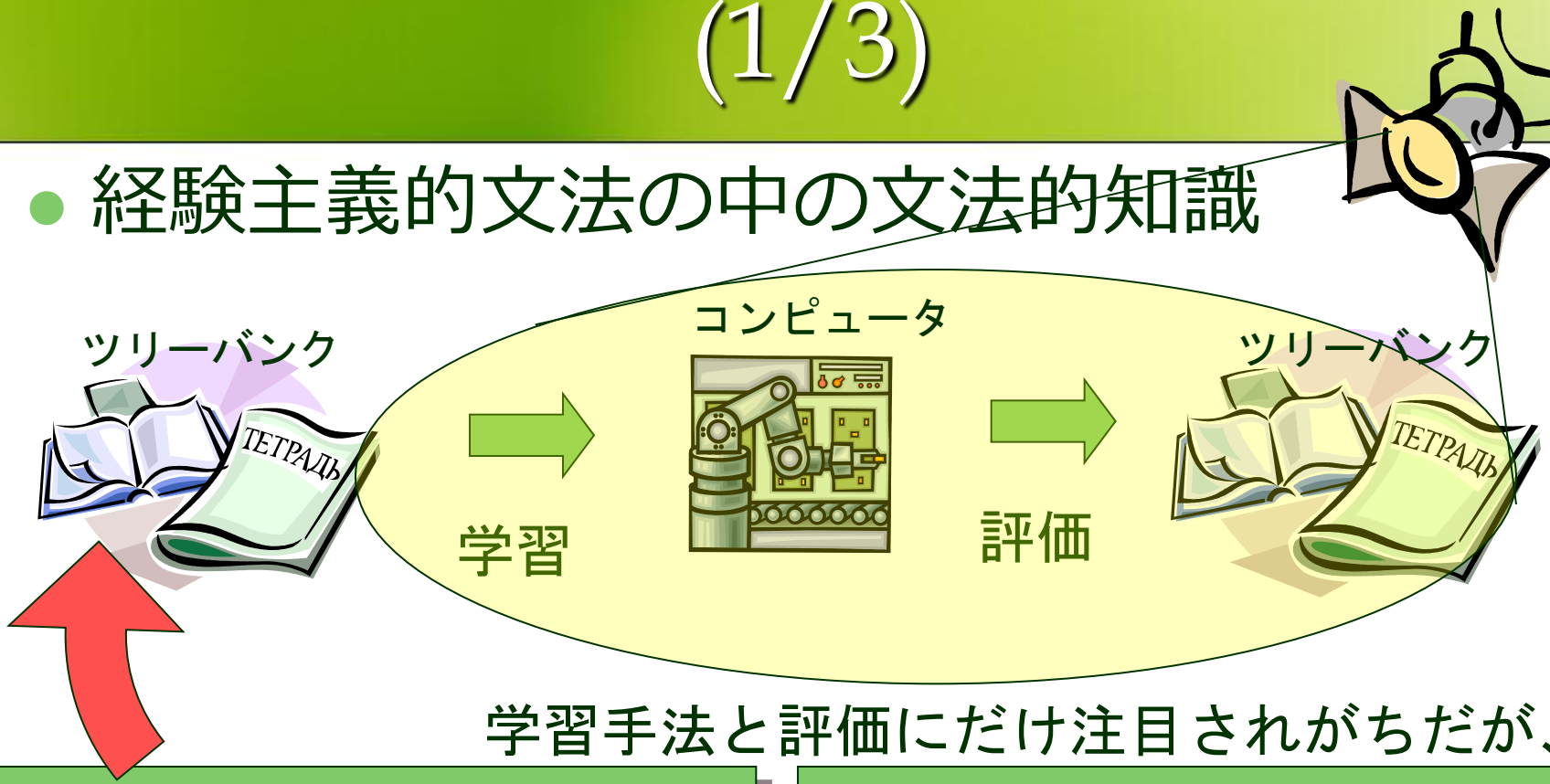
ツリーバンク

・ツリーバンクの詳細化、構造化
・ツリーバンクからの文法抽出



文法とツリーバンクの双対性 (1/3)

- 経験主義的文法の中の文法的知識



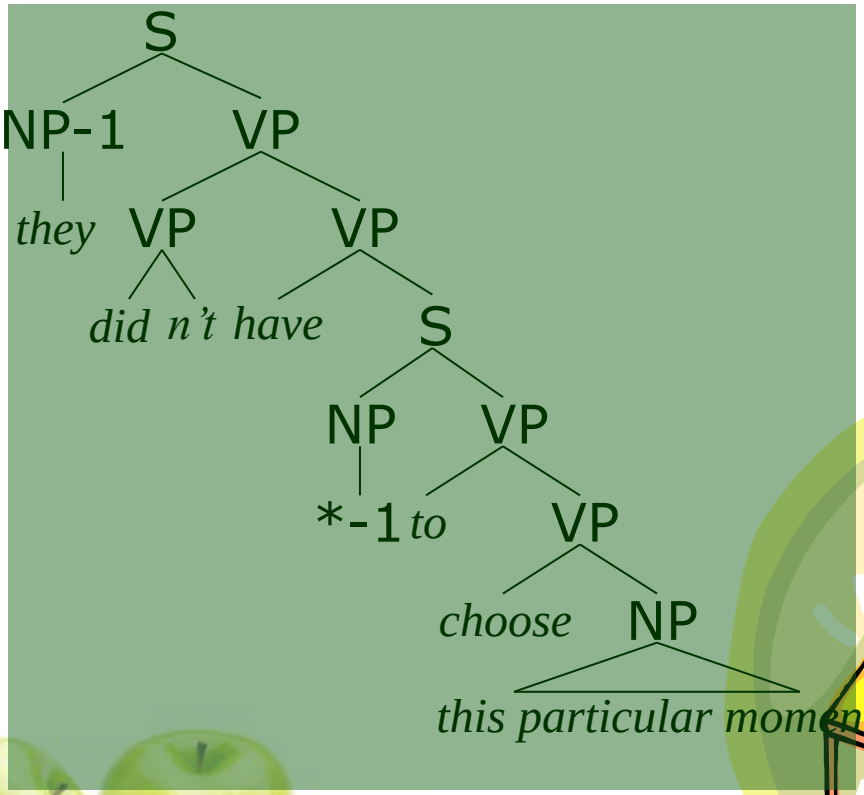
学習手法と評価にだけ注目されがちだが、、、

ツリーバンクに文法的知識
・ツリーバンク作成指針の中に暗黙的に
・構文木の構造から文法や辞書を作成するのに十分な情報

精度をあげるために文法的知識を導入
・最初から文法的制約と構造をツリーバンクに導入したほうがすっきり

文法とツリーバンクの双対性 (2/3)

- 合理主義的文法でのツリーバンク

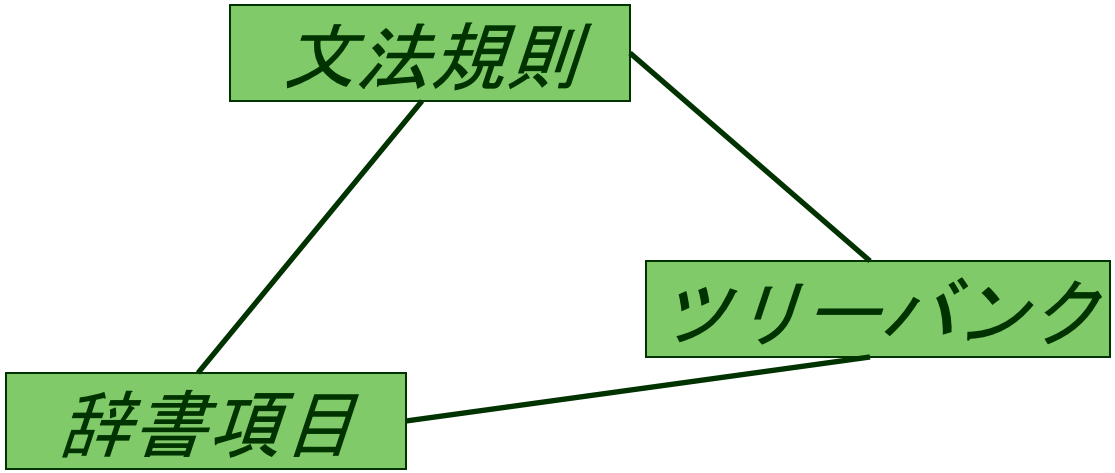


文法とツリーバンクの双対性 (3/3)

- 経験主義的文法
  - ツリーバンクに含まれる暗黙の文法
    - ツリーバンク作成の指針に含まれる文法的知識
    - 構文木の構造に含まれる文法的知識
- 合理主義的文法
  - 辞書項目と文法規則をつくる際に、構文木を想定

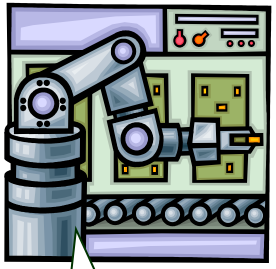
文法の３つのリソース

- 文法開発では３つのリソースを想定している



経験主義的文法と合理主義的文法を超えて

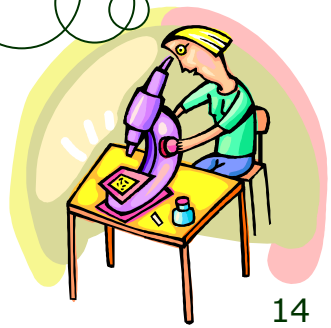
- 三つのリソースを同時につくれば万事解決？



ツリーバンクだけいただきます

S → NP VP
NP → DET N
NP → N
…

ツリーバンク、
文法規則、辞書



合理主義的文法開発のジレンマ

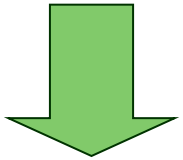
- ツリーバンクと文法の不一致

データと理論の不一致？

辞書

文法規則

S → NP VP
NP → DET N
NP → N
…



≠

作成したツリーバンク　　導出されたツリーバンク

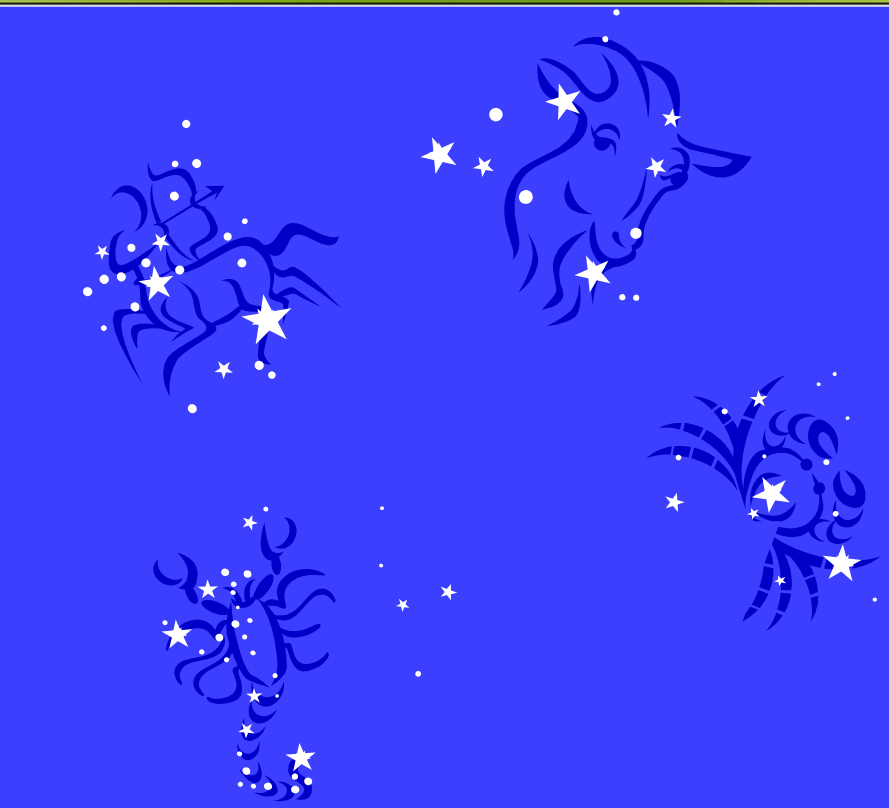
文法理論の恣意性

- 同じような機能・構造によって分類
- 観点・基準によって分類が異なる
  - HPSG
  - LFG
  - TAG
  - …

c.f. 分類学 (進化分類学, 分岐分類学, 表形分類学)

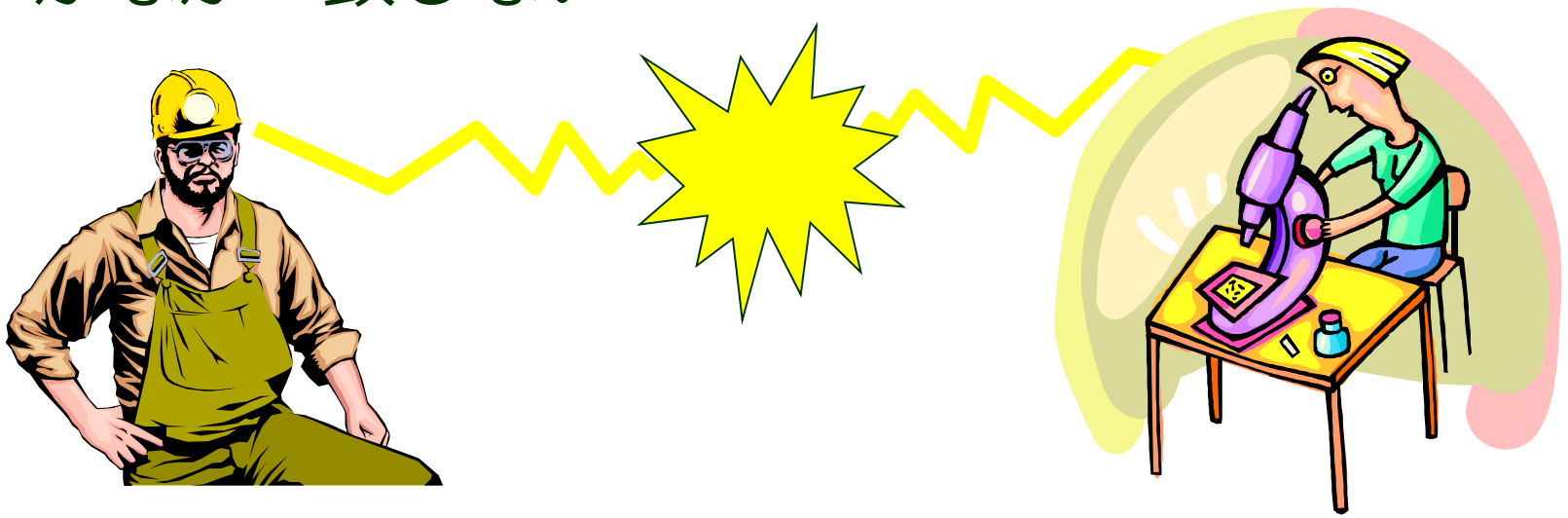
極端な話、百人いれば百の文法理論がありうる!

星座と文法理論



まぁ、星座の話はおいといて、、、

- 違う基準・違う方法論でつくるツリーバンクはなかなか一致しない



直感＋アノテーションガイドライン

辞書と文法規則による文法理論

文法が先かツリーバンクが先か？

ツリーバンク

文法

S → NP VP
NP → DET N
NP → N
…

文法なんかいらねーよ

不一致が生じたときにどちらを修正すべきか？

どちらにあわせればいいのだろうか？

どちらを先に開発すべきか？

文法を先につくる

辞書と文法規則による文法理論

- 文法がツリーバンクを説明

・ツリーバンクは文法に導出される副産物

・文を解釈するときの観点・基準を与えるのが文法なのだから、ツリーバンクは文法に従うべき

ツリーバンクを先につくる

- ツリーバンクが文法を説明

直感＋アノテーションガイドライン

S → NP VP
NP → DET N
NP → N
…

自分の頭の中にある文法解析結果をまず外在化

外在化されたツリーバンクを説明できるように文法を開発、導出

合理主義的文法と経験主義的文法を超えて

ツリーバンクと文法の協調関係

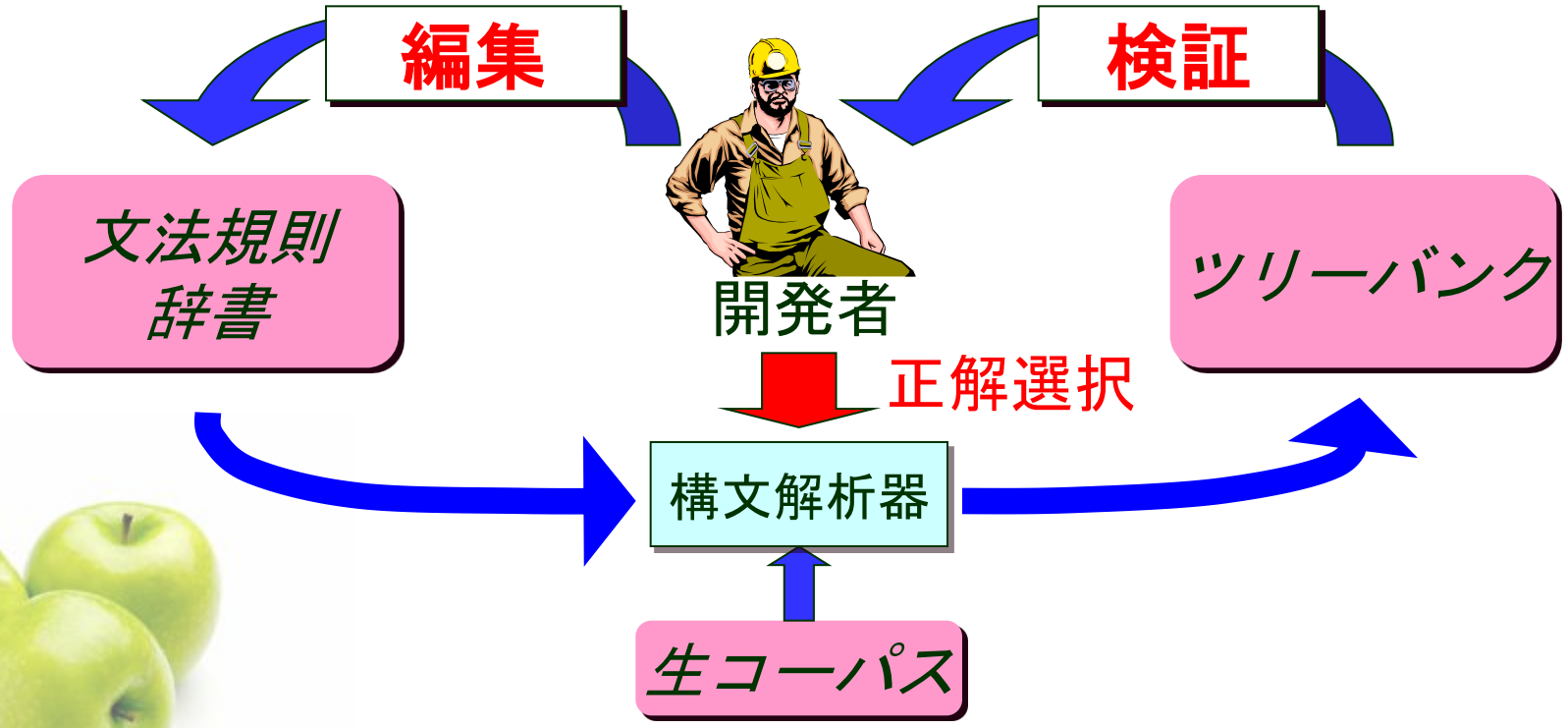
- 文法開発ではツリーバンクの役割が重要
  - 曖昧性解消モデルのための統計情報を提供する
  - 文法の不備・矛盾・間違いを検出する
  - 構文解析・文生成の性能を客観的に評価する
- ツリーバンク開発では合理的な構造化が必要
  - 文法理論による構文構造の明示化
    - より複雑な構造のアノテーション・文法開発を容易にする
    - 統語構造の一般化 （例、能動態と受動態)
  - 性能向上のために文法的知識を断片的に導入
    - 最初から文法的制約と構造化を導入したほうが良い
  - ツリーバンクの一貫性の向上

ツリーバンクと文法の開発

- 文法評価のためのツリーバンク
  - PARC 700 Dependency Bank [King et al. 2003]
    - Penn Treebank Section 23 から無作為に700文を抽出
    - English XLE パーザで構文解析し、人手で正解の f-structure を選択
  - XLE パーザと Collins パーザを客観的に比較 [Kaplan et al. 2004]
    - 構文解析時間は Collins パーザが速い
    - 構文解析精度は XLE パーザの方が高い
- 文法開発のためのツリーバンク
  - ツリーバンキング (文法が先の文法開発)
  - コーパス指向文法開発 (ツリーバンクが先の文法開発)

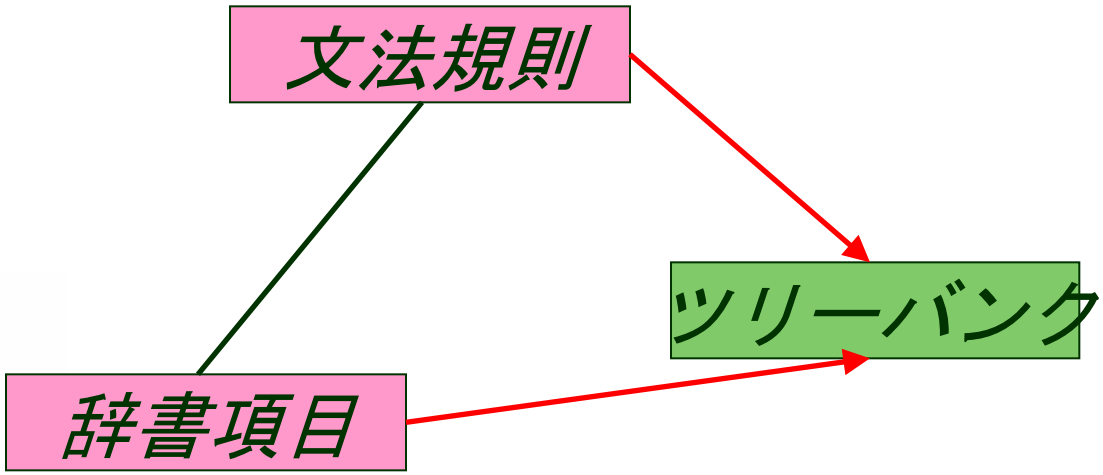
ツリーバンキング (文法が先)

- 文法開発過程にツリーバンク開発を組み込む
- 生コーパスを構文解析し、人手で正解を選択
  - Redwoods [Oepen et al. 2002], Hinoki [Bond et al. 2004]

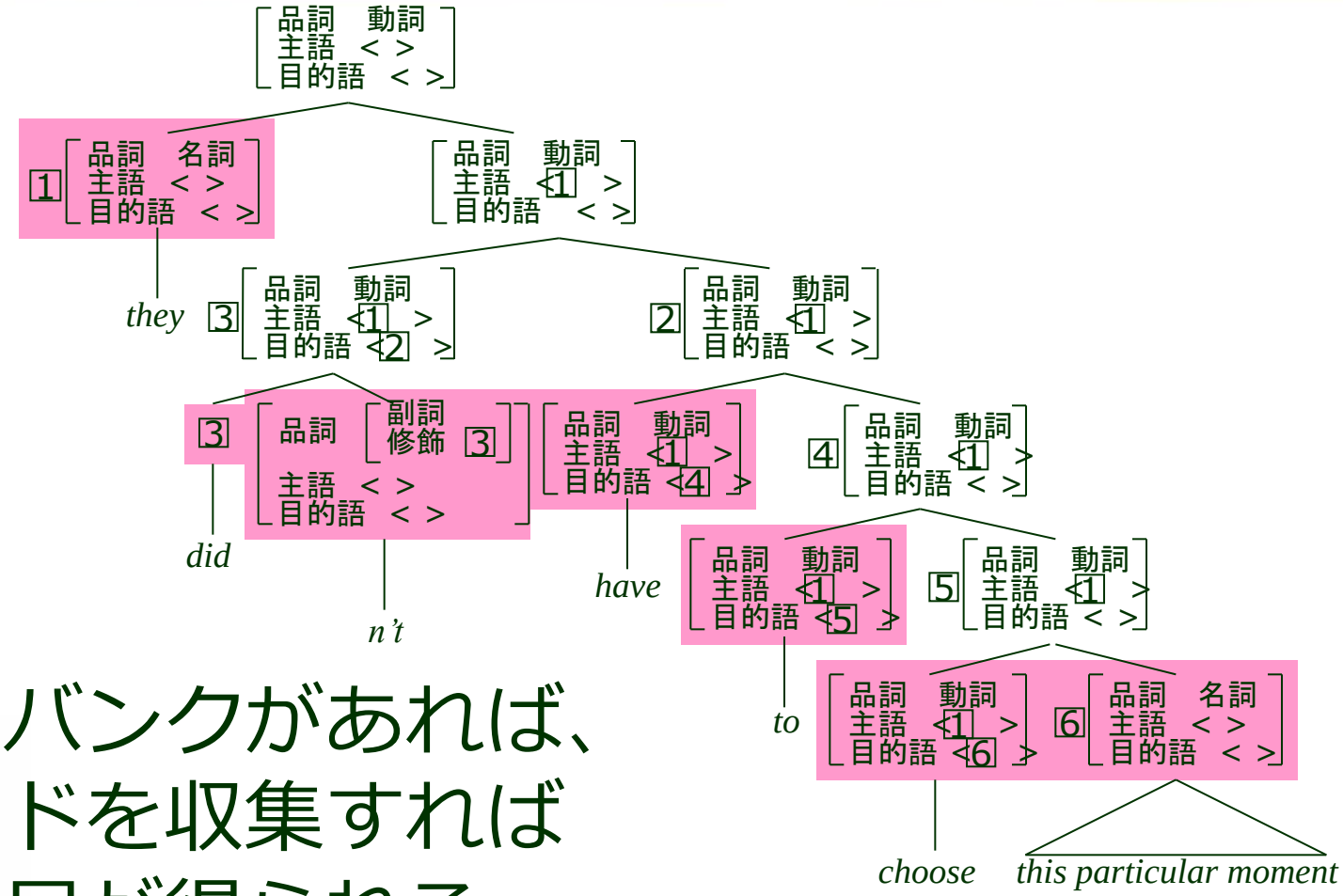


ツリーバンキングの利点

- 効率的・系統的にツリーバンクが開発できる
- ツリーバンクを曖昧性解消モデルの学習データとして利用する [Toutanova et al. 2002]
- ツリーバンク開発を通して、文法の不備・矛盾・間違いを発見できる



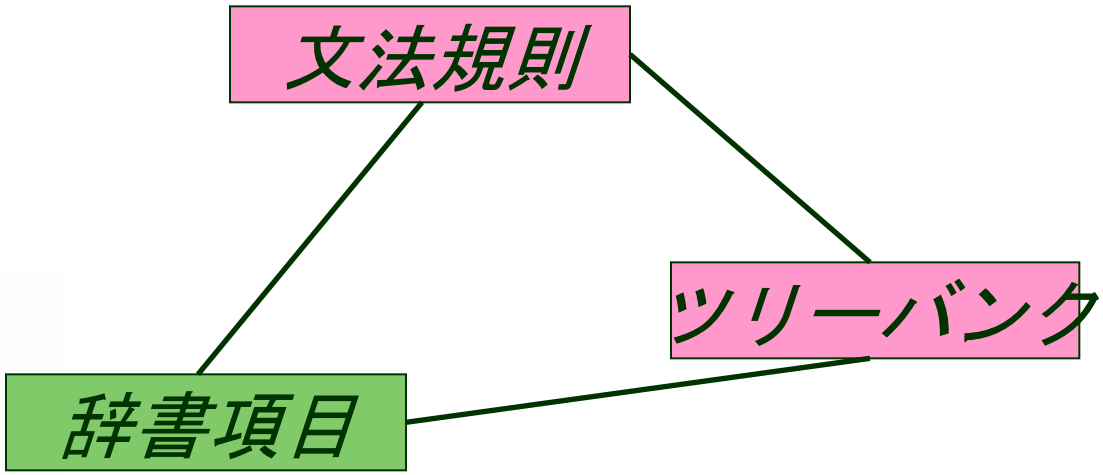
再考：辞書とツリーバンクの関係



- ツリーバンクがあれば、葉ノードを収集すれば辞書項目が得られる

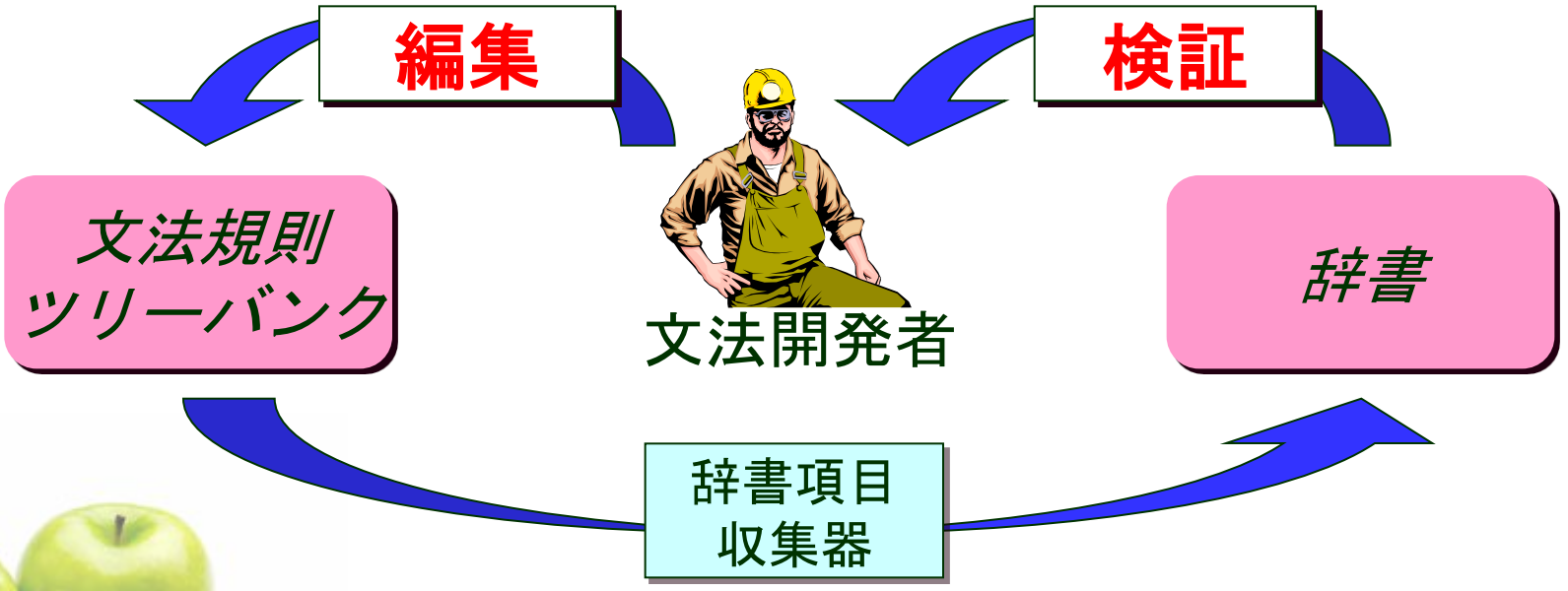
ツリーバンク＞ 辞書

- ツリーバンクがあれば辞書は得られる
- ツリーバンクの方が辞書より情報が多い
  - 文法の不備・矛盾・間違いが検出できる
  - 統計情報が得られる



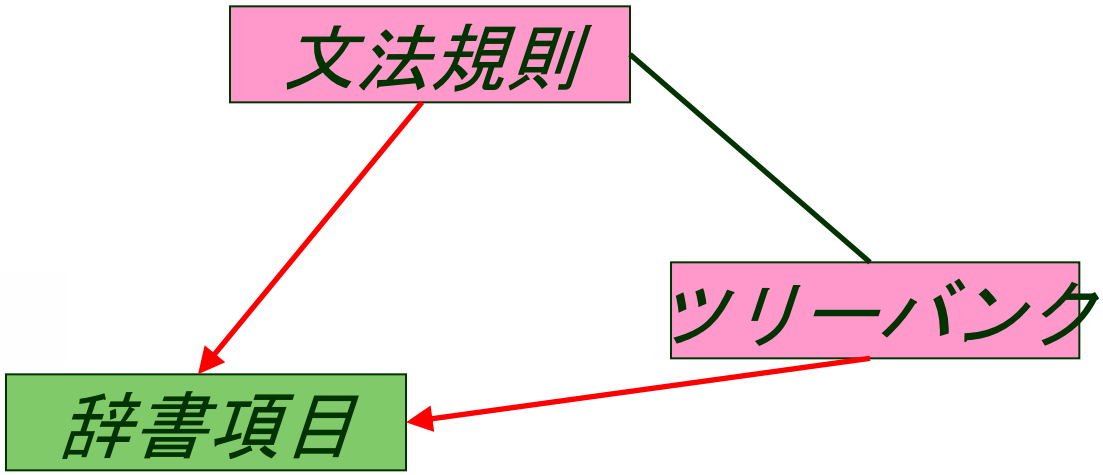
コーパス指向文法開発 (ツリーバンクが先)

- 辞書の代わりにツリーバンクを作る
  - CCG [Hockenmaier et al. 2002], HPSG [Miyao et al. 2004]
- 辞書項目はツリーバンクから収集する



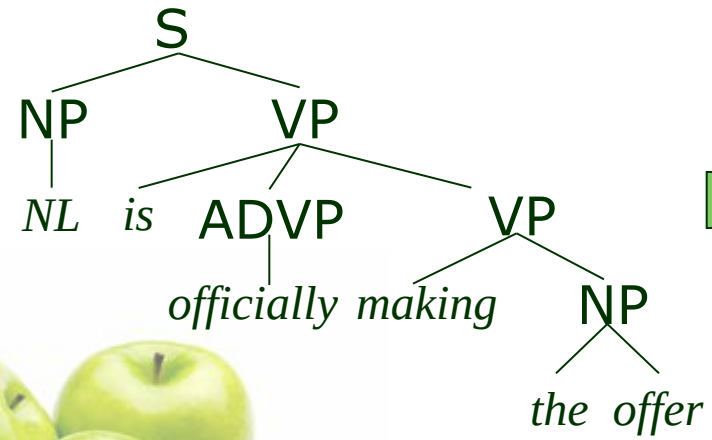
コーパス指向文法開発の利点

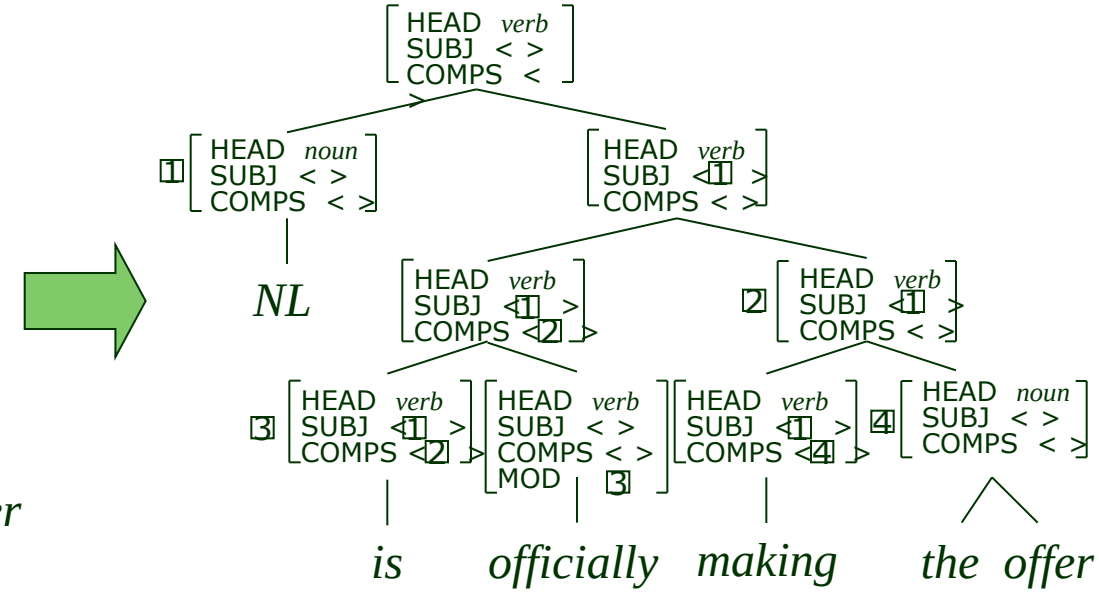
- ツリーバンクと辞書が同時に得られる
- ツリーバンク開発を通して、ツリーバンクや文法規則の不備・矛盾・間違いを発見できる



どうやってツリーバンクを作るのか？

- Penn Treebank を再利用し、文法規則に合致するように変換
- 文法開発＝文法規則に合致するようにツリーバンクを編集する過程



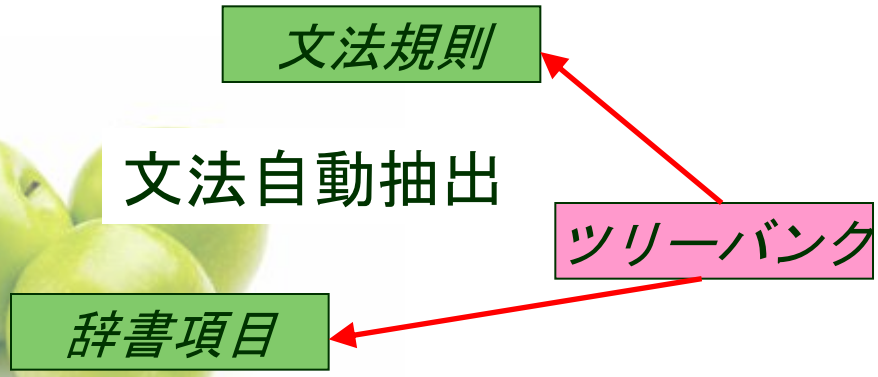


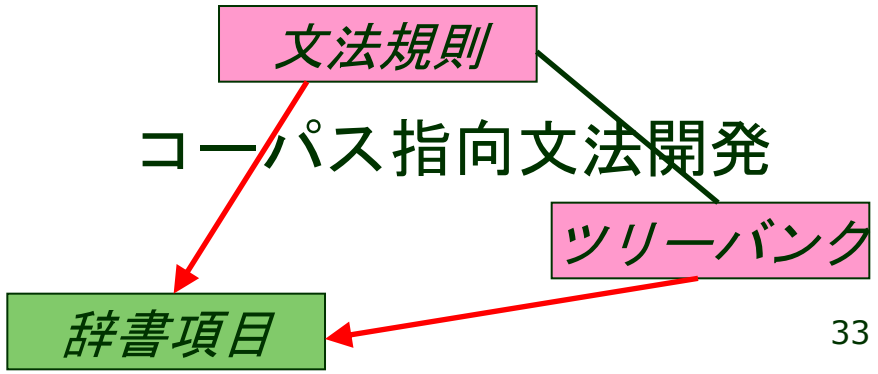
文法自動抽出との違い (1/3)

- 目標
  - 文法自動抽出: なるべく人手を介在させず、すでにあるリソースからいかに楽をして文法を獲得できるか
  - コーパス指向: なるべく人手を介在させて、いかに良いコーパスをつくれるか（＝良い文法をつくれるか）
- 開発過程
  - 文法自動抽出：全自動なので、アルゴリズムができれば数時間から数日
  - コーパス指向: 手作業で半年から数年

文法自動抽出との違い (2/3)

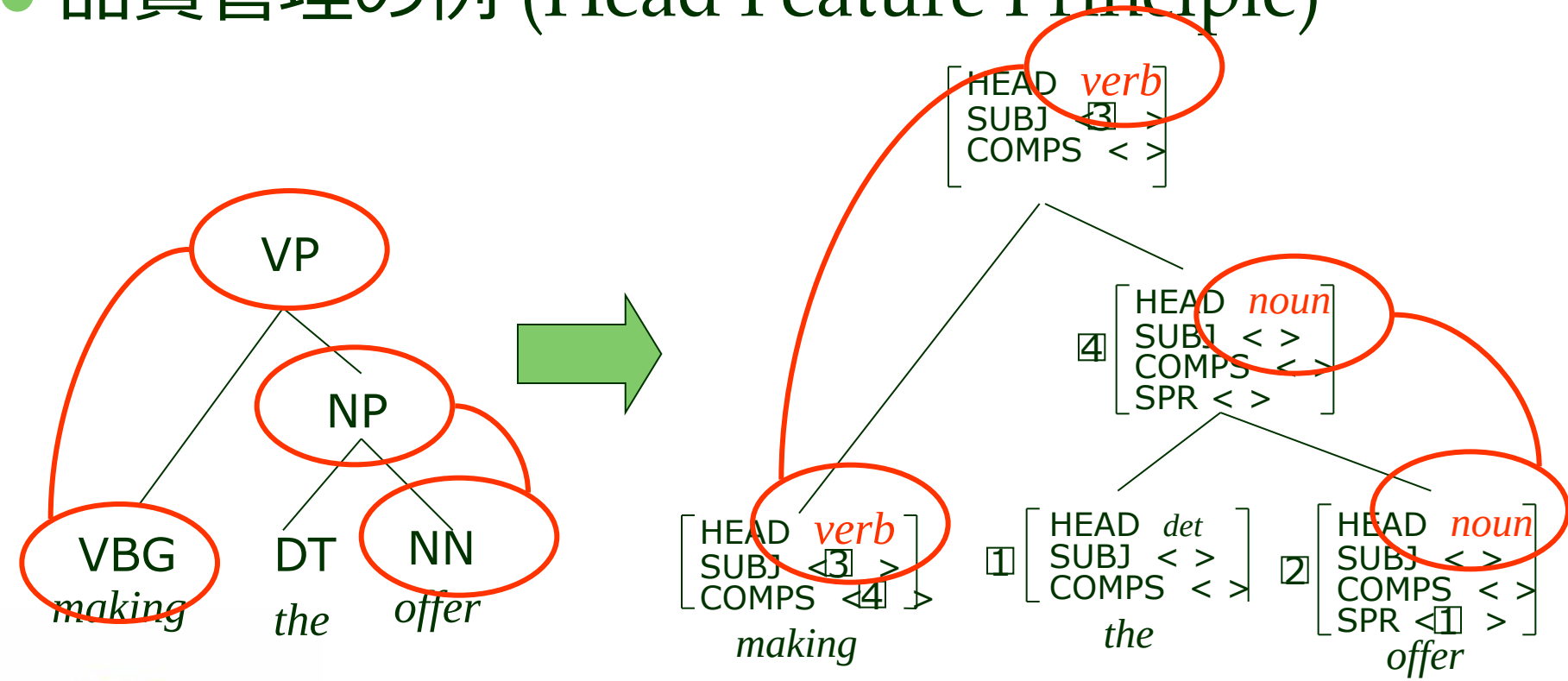
- 品質管理
  - 文法自動抽出: 抽出された文法を主に評価
  - コーパス指向:
    - ツリーバンク、文法規則は人間が管理する
    - 文法規則によるツリーバンクの構造化
      - ツリーバンクの品質が必然的に検証される
      - 得られる辞書は文法規則に従うことが保証される





文法自動抽出との違い (3/3)

- 品質管理の例 (Head Feature Principle)

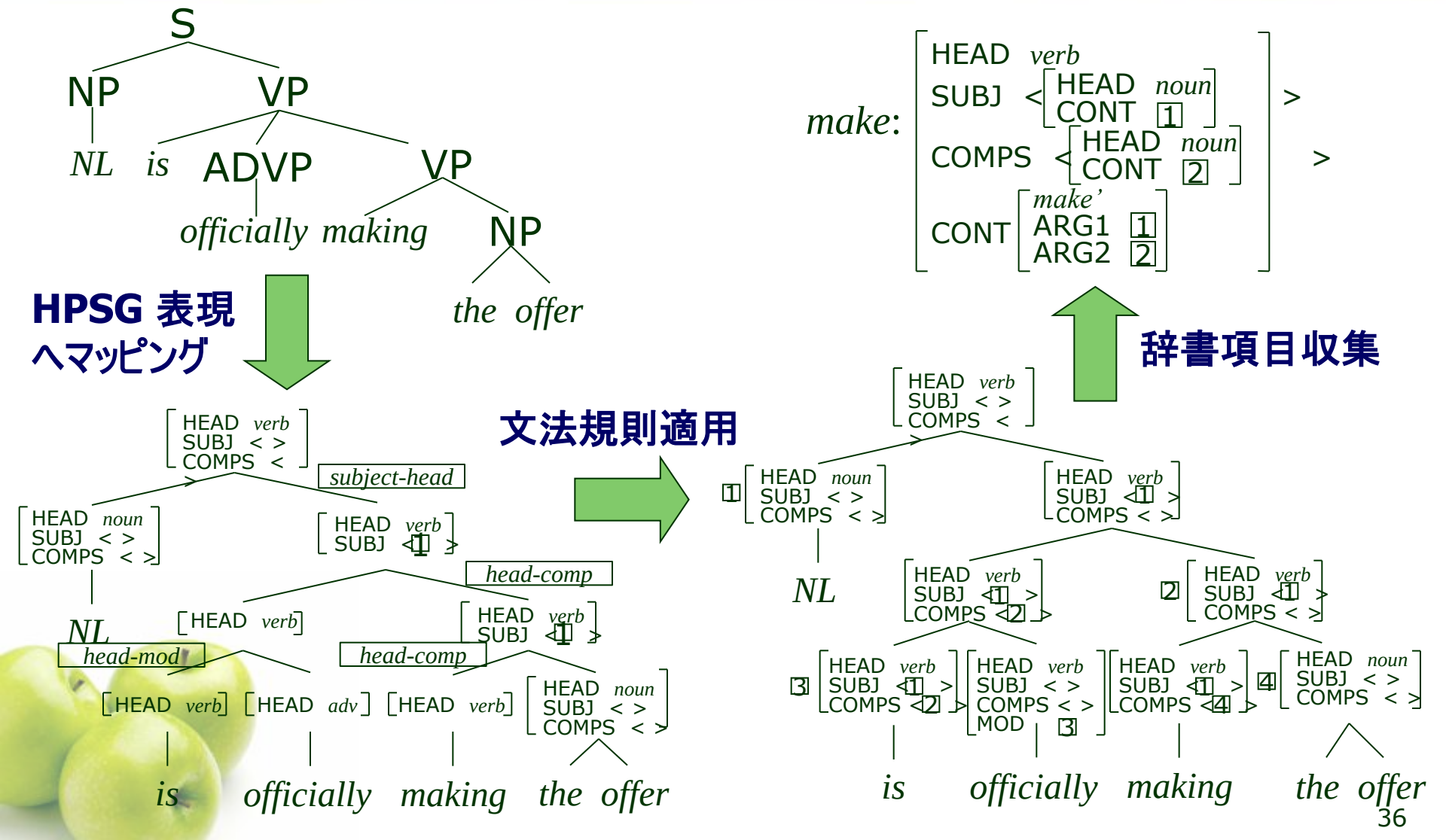


ツリーバンクの句構造が文法的制約を満たしているかチェックされる

HPSG ツリーバンクの開発

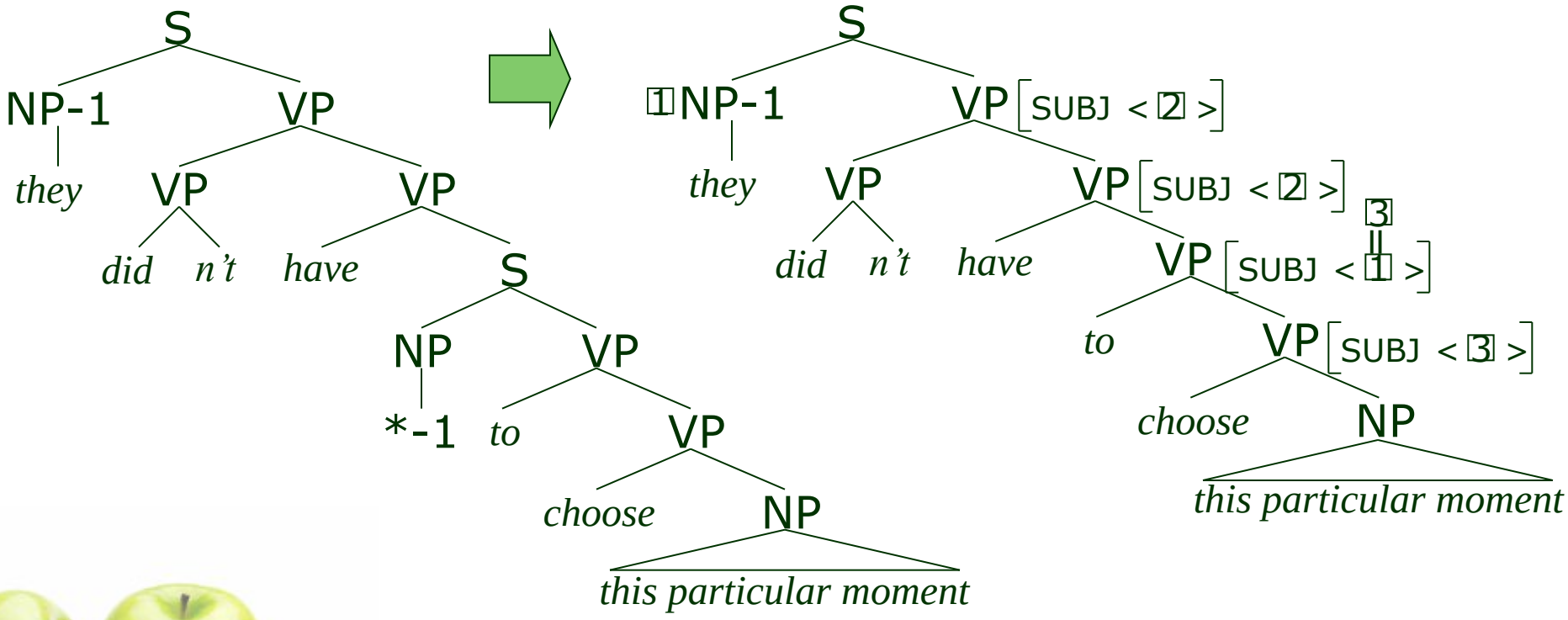
- Penn Treebank の構造をHPSG理論に基づく構造に変換する
  - 木構造変換・素性の追加
    - 下位範疇化、受身、命令形・疑問形、時制、格、量化、control/raising、 small clause、長距離依存、関係節、tough 構文、自由関係詞、並列構造、外置変形、倒置、挿入、同格、引用、etc.
- HPSG の文法規則を適用
  - 文法規則やツリーバンクの不備・矛盾・間違いは、制約違反として検出される

辞書・ツリーバンク開発の概要



助動詞・control/raising

- 不飽和構成素を補語としてとるようにする



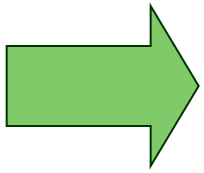
長距離依存・関係節

- SLASH: 移動した項を表す
- REL: 関係詞と先行詞の関係を表す

HPSGのカテゴリへマッピング

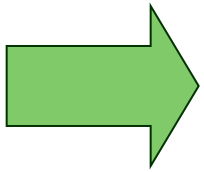
- （非）終端記号を素性構造へマッピング

NN
（普通名詞）



$$\begin{bmatrix} \text{HEAD: noun} \\ \text{AGR: 3sg} \end{bmatrix}$$

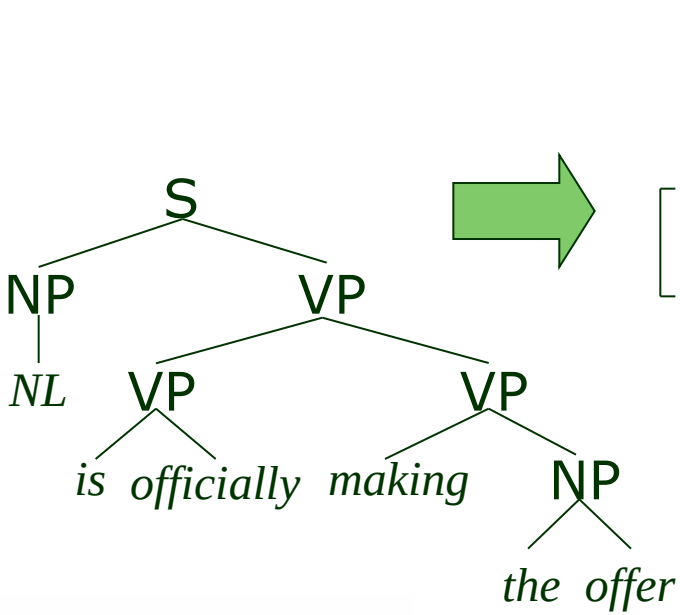
VBZ
（三単現動詞）

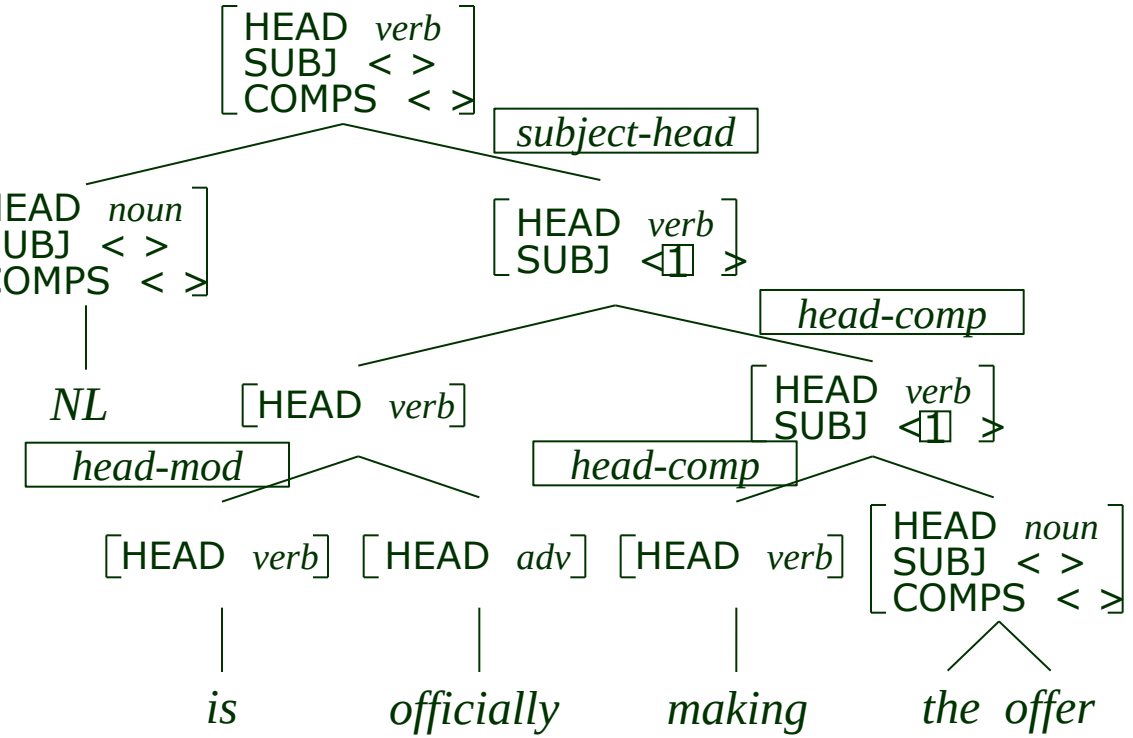


$$\begin{bmatrix} \text{HEAD: verb} \\ \text{AGR: 3sg} \\ \text{VFORM: finite} \\ \text{TENSE: present} \end{bmatrix}$$

具体例

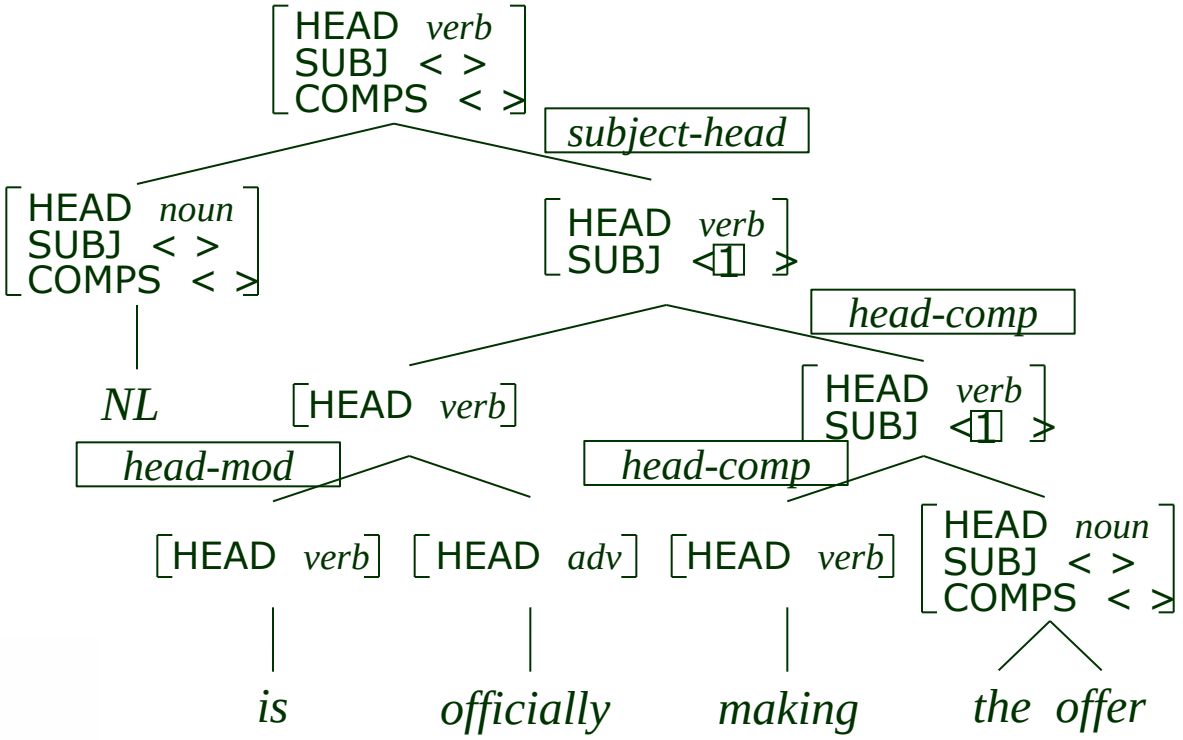
- *"NL is officially making the offer"*





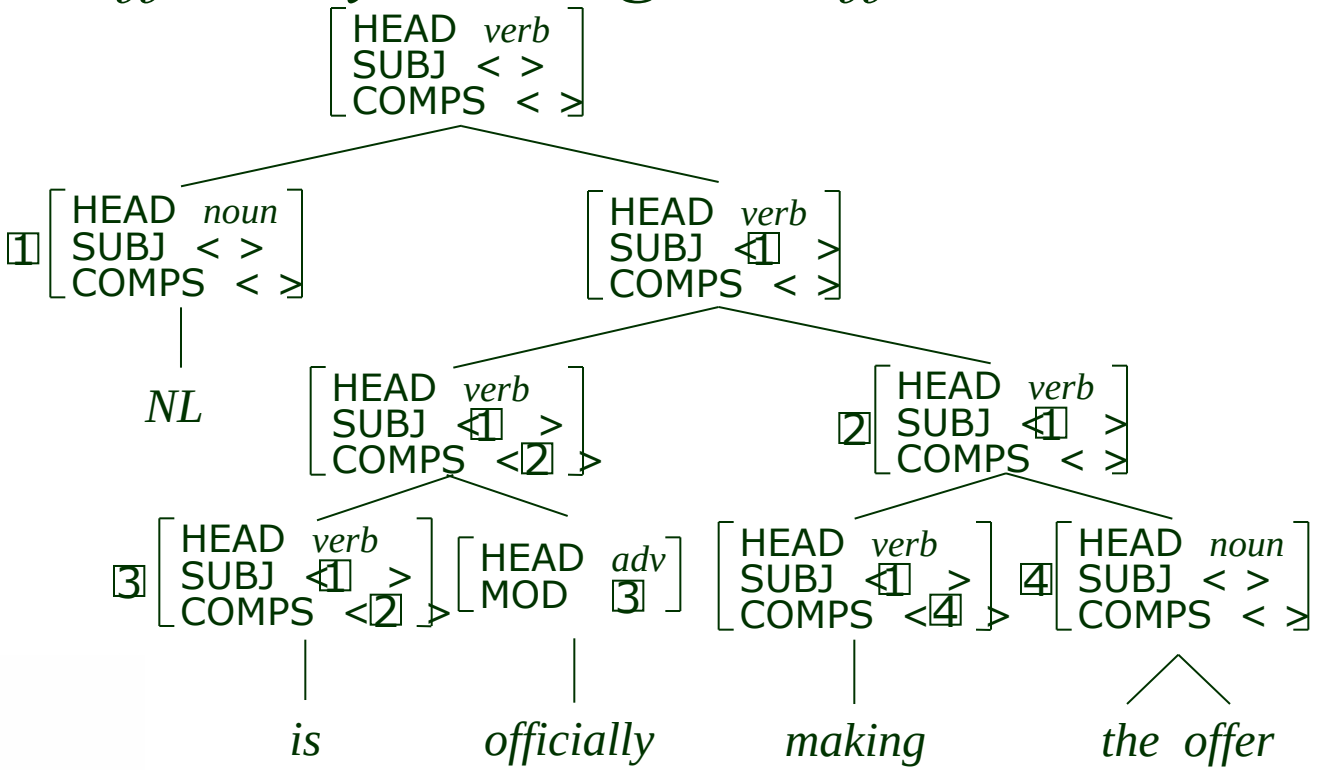
文法規則の適用

- *"NL is officially making the offer"*

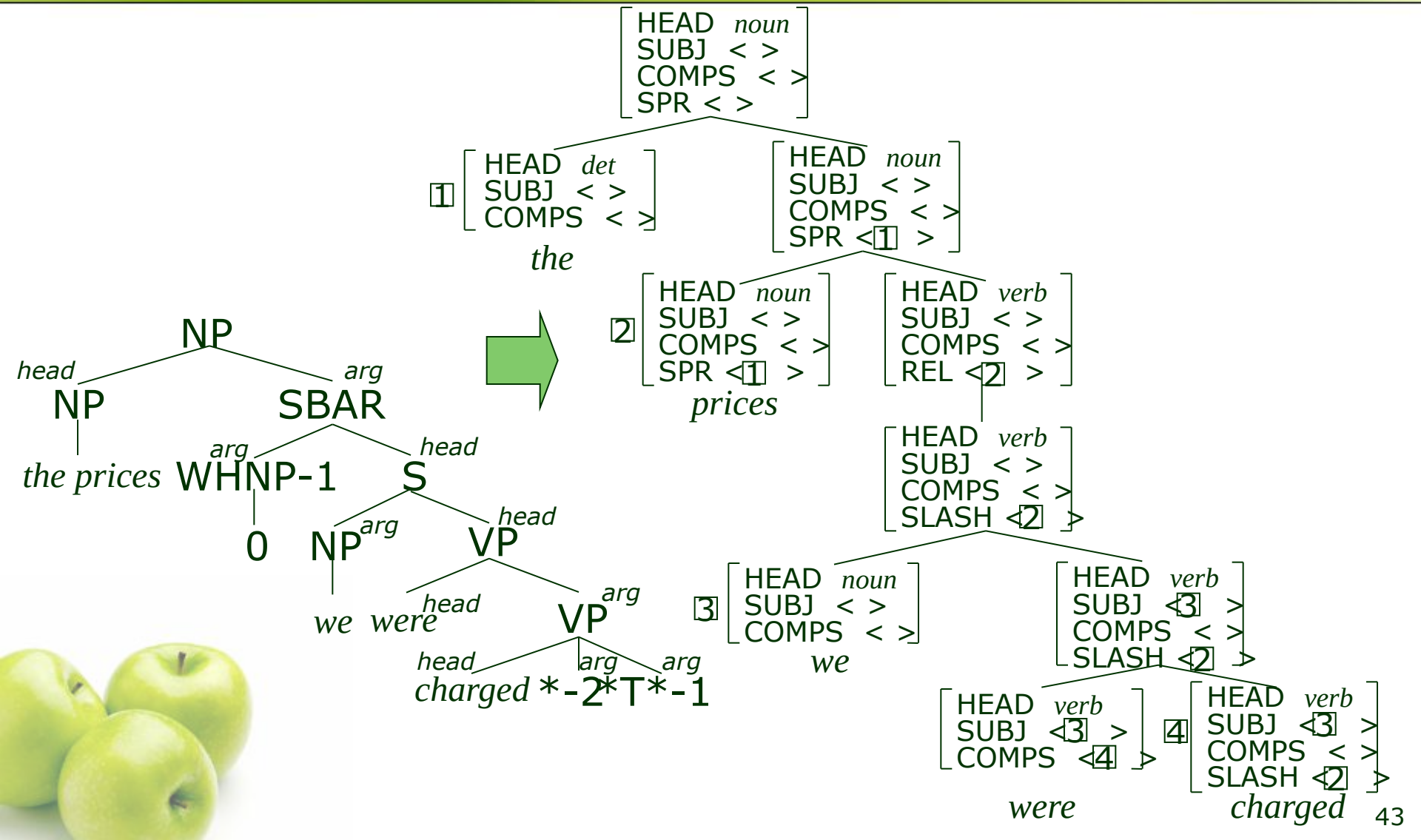


文法規則の適用

- *"NL is officially making the offer"*

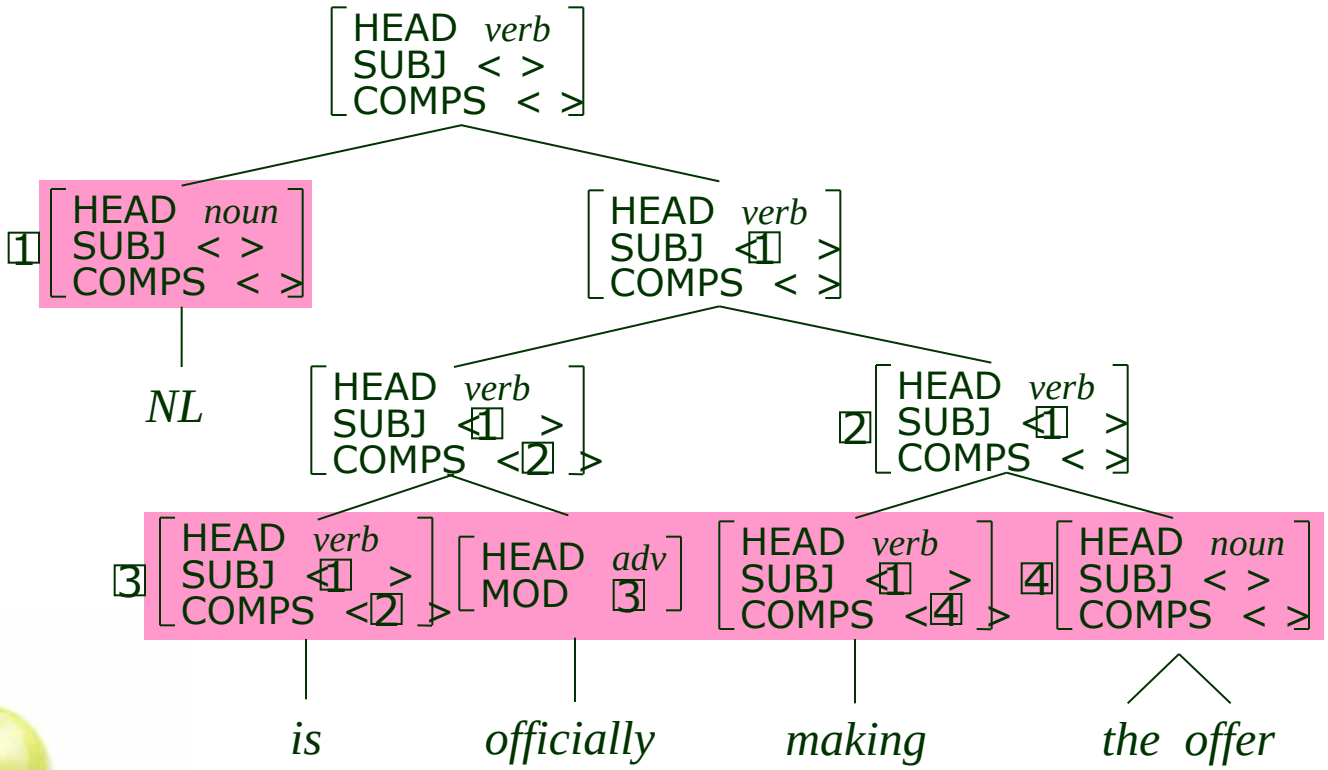


複雑な例


NP
head
NP
arg
SBAR
the prices
arg
WHNP-1
head
S
0
NP
arg
head
VP
we
were
head
VP
arg
head
arg
arg
charged *-2*T*-1
HEAD noun
SUBJ < >
COMPS < >
SPR < >
1
HEAD det
SUBJ < >
COMPS < >
the
HEAD noun
SUBJ < >
COMPS < >
SPR <1 >
2
HEAD noun
SUBJ < >
COMPS < >
SPR <1 >
prices
HEAD verb
SUBJ < >
COMPS < >
REL <2 >
HEAD verb
SUBJ < >
COMPS < >
SLASH <2 >
3
HEAD noun
SUBJ < >
COMPS < >
we
HEAD verb
SUBJ <3 >
COMPS < >
SLASH <2 >
HEAD verb
SUBJ <3 >
COMPS <4 >
4
HEAD verb
SUBJ <3 >
COMPS < >
SLASH <2 >
were
charged

辞書項目の収集

- HPSG構文木の葉ノードは、辞書項目の実例

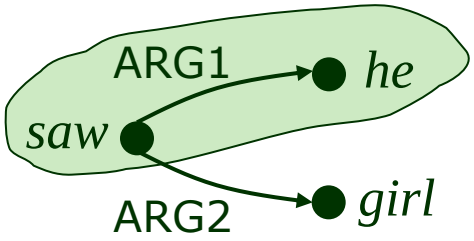


HPSG文法の評価実験

- HPSG ツリーバンクから収集した辞書項目を評価
  - 未知の文に対する被覆率
  - ツリーバンクのサイズと被覆率の関係
- Penn Treebank Section 02-21 (39,832文) をHPSG ツリーバンクに変換し、辞書項目を収集
- テストデータ：Section 23 を HPSG ツリーバンクに変換したもの (2,299 文)

被覆率と構文解析精度

- 被覆率： 99.8%
  - 構文解析に成功した文の割合
- 構文解析精度： 適合率 90.44%，再現率 90.19%
  - 述語-項関係の精度



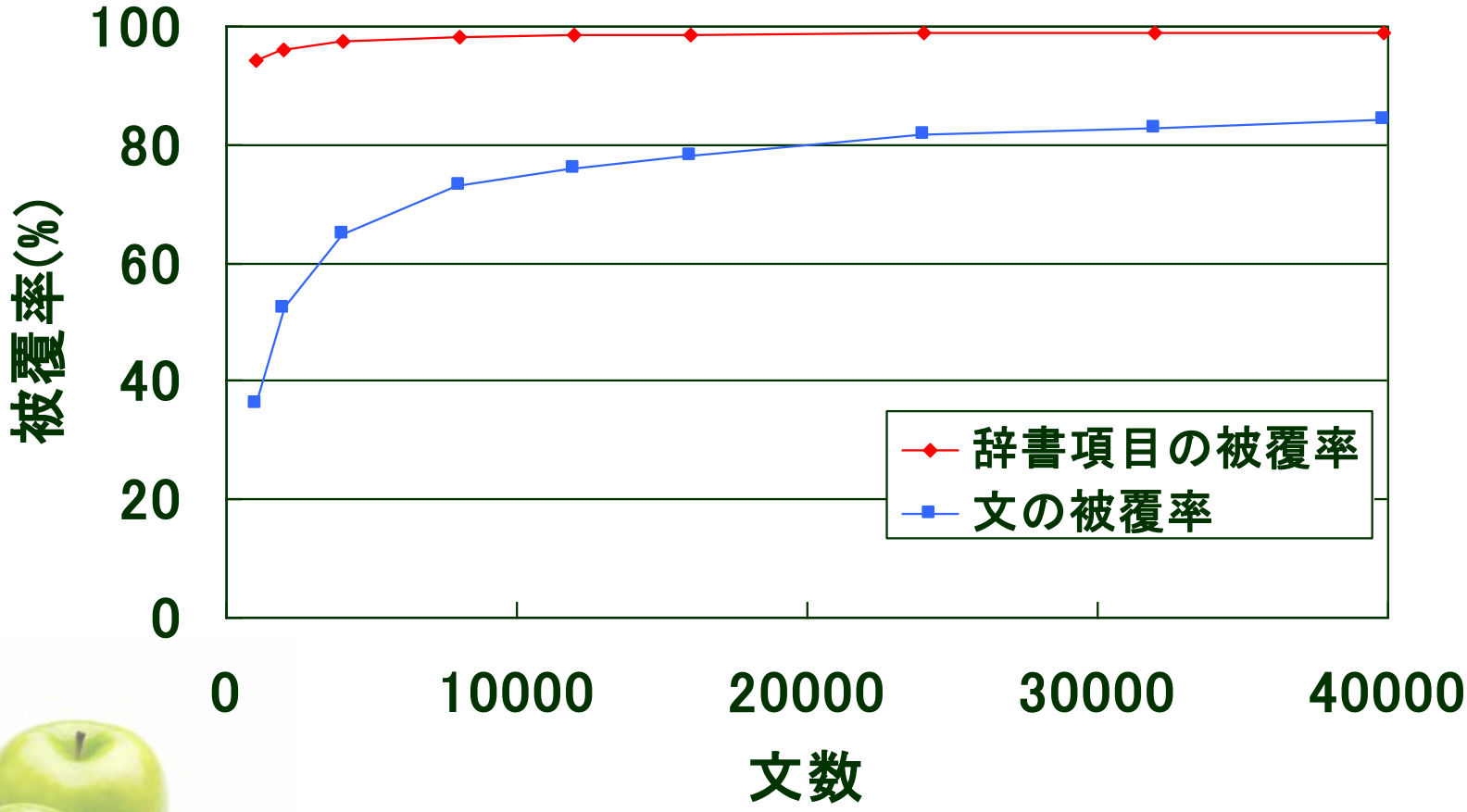
- 強意の被覆率： 84.4%
  - 構文森が完全一致の構文木を含む文の割合

被覆率

- テストデータ中の辞書項目を文法が含んでいる割合を測定
- 文中の全ての辞書項目が被覆されていれば文が被覆されていると判定

	辞書項目単位	文単位
未知語処理なし	96.52%	54.7%
未知語処理あり	99.15%	84.8%

ツリーバンクのサイズ vs. 被覆率



まとめ (1/3)

- 合理主義的文法
  - 人手による文法規則と辞書の開発と中心とした文法開発
  - 合理的な利点
    - 言語学的な妥当性
    - 複雑な構造、深い構造の記述が容易
  - 問題点
    - 網羅性と一貫性のトレードオフ
    - 曖昧性解消の先送り
    - 性能評価の問題
- 経験主義的文法
  - ツリーバンクを中心とした文法開発
  - 経験的な利点
    - 網羅性
    - 一貫性
    - 機械学習・統計学習が容易
    - 評価も容易
  - 問題点
    - 正解の客観的基準が存在しない
    - 深い構造・複雑な構造の品質管理は困難
    - 自動的な文法抽出の妥当性

文法とツリーバンクの両方を開発することが重要!

まとめ (2/3)

- 違う基準・違う方法論でつくるツリーバンクはなかなか一致しない

経験主義的文法開発

直感＋アノテーションガイドライン

合理主義的文法開発

辞書と文法規則による文法理論

まとめ (3/3)

- コーパスと文法の両方を開発
  - 経験主義的文法と合理主義的文法の双方の利点
  - 理論（＝文法）とデータ（＝ツリーバンク）をいかに一致させるか？
    - ツリーバンキング(文法が先の文法開発)
      - 文法規則や辞書を優先し、ツリーバンクを開発
      - 例: Redwoods, Hinoki, PARC 700 Dependency Bank
    - コーパス指向文法開発(ツリーバンクが先の文法開発)
      - ツリーバンクを優先し、文法規則や辞書を開発
      - 文法的知識をツリーバンクとして外在化
      - 例： CCGツリーバンクからCCG文法、HPSGツリーバンクからHPSG文法

参考文献

- H. Alshawi (Ed.) (1992) The Core Language Engine. MIT Press.
- A. K. Joshi and Y. Schabes (1997) Tree Adjoining Grammars. in G. Rosenberg and A. Salomaa, (eds.), Handbook of Formal Languages, vol. 3, pp. 69-124.
- XTAG Research Group (2001) A lexicalized tree adjoining grammar for English. Technical Report IRCS-01-03, University of Pennsylvania.
- A. Abeillé and M.-H. Candito and A. Kinyon (2000) FTAG: developping and maintaining a wide-coverage grammar for French. ESSLLI-2000.
- J. Bresnan (1982) The Mental Representation of Grammatical Relations. MIT Press.

参考文献

- S. Riezler, T. H. King, R. S. Crouch, J. T. Maxwell, R. M. Kaplan (2002) Parsing the Wall Street Journal using a lexical-functional grammar and discriminative estimation techniques. In Proc. of ACL 2002.
- R. M. Kaplan, S. Riezler, T. H. King, J. T. Maxwell, A. Vasserman (2004) Speed and accuracy in shallow and deep stochastic parsing. In Proc. of HLT/NAACL-2004.
- M. Forst and C. Rohrer (2006). Improving coverage and parsing quality of a large-scale LFG for German. In Proc. of LREC 2006.
- C. Pollard and I. A. Sag (1994) Head-Driven Phrase Structure Grammar. University of Chicago Press.

参考文献

- S. Müller (1996) The Babel-System – An HPSG Prolog Implementation. In Proc. of 4th International Conference on the Practical Application of Prolog, pp. 263—277.
- M. Siegel and E. M. Bender (2002) Efficient Deep Processing of Japanese. In Proc. of the 3rd Workshop on Asian Language Resources and International Standardization. COLING 2002 Post-Conference Workshop.
- G. Bouma, G. van Noord, R. Malouf (2000) Alpino: Wide-coverage Computational Analysis of Dutch. Computational Linguistics in the Netherlands. Selected Papers from the 11th CLIN Meeting.
- J. Carroll and T. Briscoe (2002) High Precision Extraction of Grammatical Relations. In Proc. of COLING 2002.

参考文献

- M. Butt, H. Dyvik, T. H. King, H. Masuichi, and C. Rohrer (2002) The Parallel Grammar Project. In Proceedings of COLING-2002 Workshop on Grammar Engineering and Evaluation. pp. 1-7.
- D. Flickinger (2002) On building a more efficient grammar by exploiting types. In Stephan Oepen, Dan Flickinger, Jun'ichi Tsujii and Hans Uszkoreit (eds.) Collaborative Language Engineering. Stanford: CSLI Publications, pp. 1-17.
- E. M. Bender, D. Flickinger, and S. Oepen (2002) The grammar Matrix. An open-source starter-kit for the rapid development of cross-linguistically consistent broad-coverage precision grammar. In Proc. of the Workshop on Grammar Engineering and Evaluation at COLING 2002.

参考文献

- T. Götz and D. Meurers (1997) The ConTroll System as Large Grammar Development Platform. ``ENVGRAM'' ACL-Workshop.
- A. Copestake and D. Flickinger (2000) An open-source grammar development environment and broadcoverage English grammar using HPSG. In Proc. LREC-2000.
- S. Oepen and J. Carroll (2000) Performance profiling for parser engineering. Natural Language Engineering, 6 (1) (Special Issue on Efficient Processing with HPSG):81–97.
- T. Baldwin, E. M. Bender, D. Flickinger, A. Kim, and S. Oepen (2004) Road-testing the English Resource Grammar over the British National Corpus. In Proc. LREC 2004, pages 2047–2050.

参考文献

- A. Frank, T. H. King, J. Kuhn, J. Maxwell (1998) Optimality Theory Style Constraint Ranking in Large-scale LFG Grammars. In Proc. of the 3rd LFG Conference.
- M. Marcus, B. Santorini, Marcinkiewicz (1993) Building a large annotated corpus of English: the Penn Treebank. Computational Linguistics 19.
- A. Bies, M. Ferguson, K. Katz, R. MacIntyre, V. Tredinnick, Grace Kim, M. A. Marcinkiewicz, B. Schasberger (1995) Bracketing Guidelines for Treebank II Style Penn Treebank Project
- G. Sampson (1995) English for the computer. Oxford University Press.
- S. Brants, S. Dipper, S. Hansen, W. Lezius, and G. Smith (2002) The TIGER Treebank. In Proc. Workshop on Treebanks and Linguistic Theories.

参考文献

- J. Hajic (1998) Building a syntactically annotated corpus: The Prague Dependency Treebank. In Issues of Valency and Meaning.
- E. Hinrichs, J. Bartels, Y. Kawata, V. Kordoni, and H. Telljohann (2000) The Tubingen treebanks for spoken German, English, and Japanese. In W. Wahlster (ed.), Verbmobil: Foundations of Speech-to-Speech Translation. Springer.
- EDR (1995) EDR 電子化辞書使用説明書第2版. Technical Report TR-045.
- 黒橋、長尾 (1997) 京都大学テキストコーパス・プロジェクト.言語処理学会第３回年次大会発表論文集.
- 前川、籠宮、小磯、小椋、菊池 (2000) 日本語話し言葉コーパスの設計. 音声研究 4-2.

参考文献

- E. Charniak (1996) Tree-bank Grammars. Technical Report CS-96-02, Department of Computer Science, Brown University.
- E. Charniak (1997) Statistical parsing with a context-free grammar and word statistics. In Proc. 14th National Conference on Artificial Intelligence.
- S. Sekine and R. Grishman (1995) A Corpus-based Probabilistic Grammar with Only Two Non-terminals. In IWPT ’95.
- B. Carpenter and C. Manning (1997) Probabilistic parsing using left corner language models. In 5th IWPT.
- D. Magerman (1995) Statistical decision-tree models for parsing. In Proc. 33rd ACL.

参考文献

- A. Krotov, M. Hepple, R. Gaizauskas, Y. Wilks (1998) Compacting the Penn Treebank grammar. In Proc. 17th COLING.
- A. Krotov, M. Hepple, R. Gaizauskas, Y. Wilks (1999) Evaluating two methods for Treebank grammar compaction. Natural Language Engineering 5(4).
- M. Collins (1996) A new statistical parser based on bigram lexical dependencies. In Proc. 34th ACL.
- M. Collins (1997) Three generative lexicalised models for statistical parsing. In Proc. 35th ACL.

参考文献

- E. Charniak (2000) A maximum-entropy-inspired parser. In Proc. NAACL-2000.
- M. Johnson (1998) PCFG models of linguistic tree representations. Computational Linguistics 24(4).
- D. Gildea (2001) Corpus variation and parser performance. In Proc. 2001.
- D. Bikel (2004) Intricacies of Collins' parsing model. Computational Linguistics 30(4).
- D. Klein and C. Manning (2003) Accurate unlexicalized parsing. In Proc. ACL 2003.

参考文献

- F. Xia (1999) Extracting tree adjoining grammars from bracketed corpora. In Proc. 5th NLPRS.
- J. Chen and K. Vijay-Shanker (2000) Automated extraction of LTAGs from the Penn Treebank. In Proc. 6th IWPT.
- D. Chiang (2000) Statistical parsing with an automatically-extracted tree adjoining grammar. In Proc. 38th ACL.
- A. Cahill, M. McCarthy, J. van Genabith, and A. Way (2002) Parsing with PCFGs and automatic f-structure annotation. In Proc. 7th International Lexical-Functional Grammar Conference.

参考文献

- A. Frank, L. Sadler, J. van Genabith, and A. Way (2003) From treebank resources to LFG f-structures: Automatic f-structure annotation of treebank trees and CFGs extracted from treebanks. In A. Abeille (ed), Building and Using Syntactically Annotated Corpora. Kluwer Academic Publishers.
- T. H. King, R. Crouch, S. Riezler, M. Dalrymple, and R. Kaplan (2003) The PARC 700 Dependency Bank. In Proc. LINC 2003.
- S. Oepen, K. Toutanova, S. Shieber, C. Manning, D. Flickinger, and T. Brants (2002) The LinGO Redwoods Treebank: Motivation and preliminary applications. In Proc. COLING 2002.

参考文献

- F. Bond, S. Fujita, C. Hashimoto, K. Kasahara, S. Nariyama, E. Nichols, A. Ohtani, T. Tanaka, S. Amano (2004) The Hinoki Treebank: A treebank for text understanding. In IJCNLP-04.
- K. Toutanova, C. Manning, and S. Oepen (2002) Parse ranking for a rich HPSG grammar. In Proc. TLT2002.
- J. Hockenmaier and M. Steedman (2002) Acquiring compact lexicalized grammars from a cleaner treebank. In Proc. 3rd LREC.
- Y. Miyao, T. Ninomiya, and J. Tsujii (2004) Corpus-oriented grammar development for acquiring a Head-Driven Phrase Structure Grammar from the Penn Treebank. In Proc. IJCNLP-04.