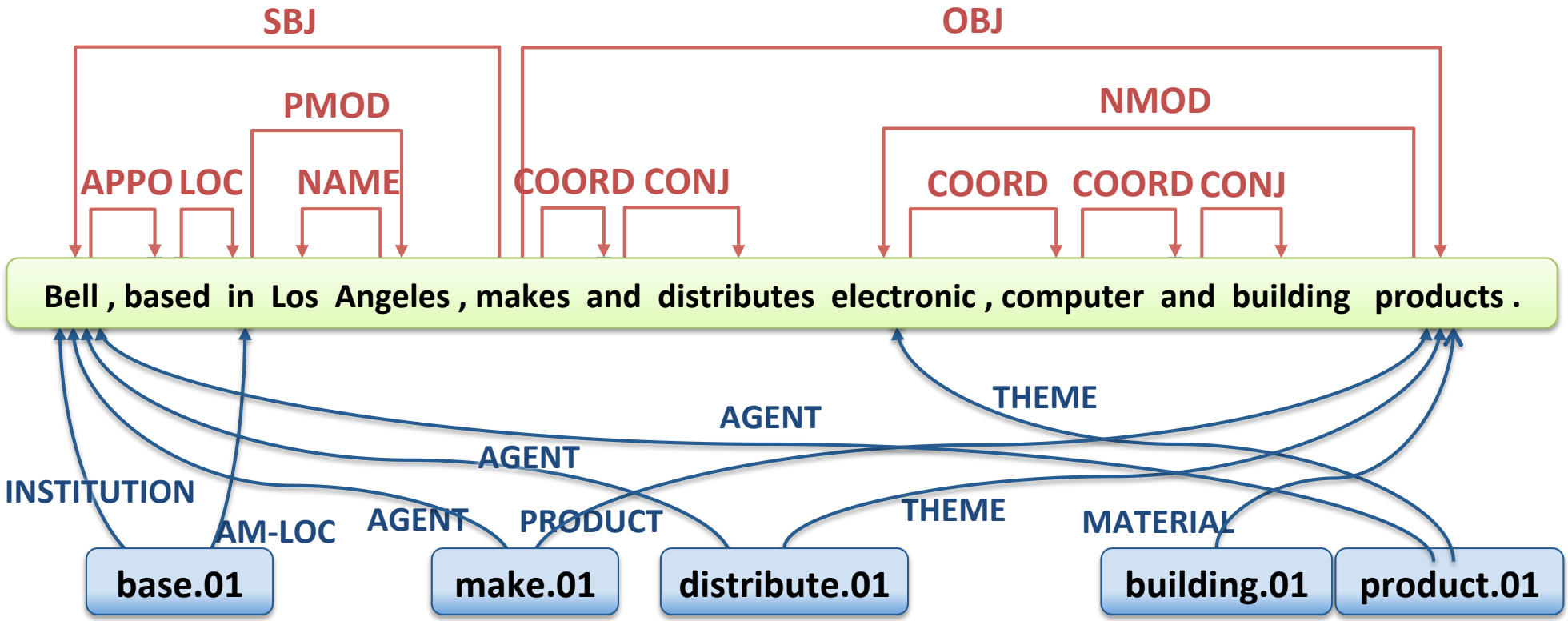
コーパスへの意味的注釈の重層的付与

乾 健太郎

東北大学大学院情報科学研究科

コーパスへの注釈付け

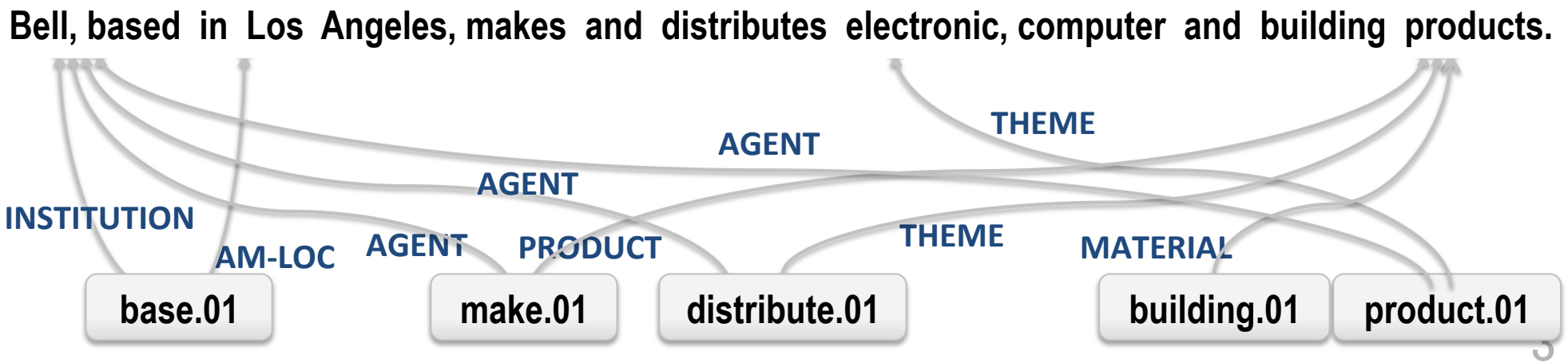
依存構造



述語項構造

意味情報の注釈付け: 何のために？

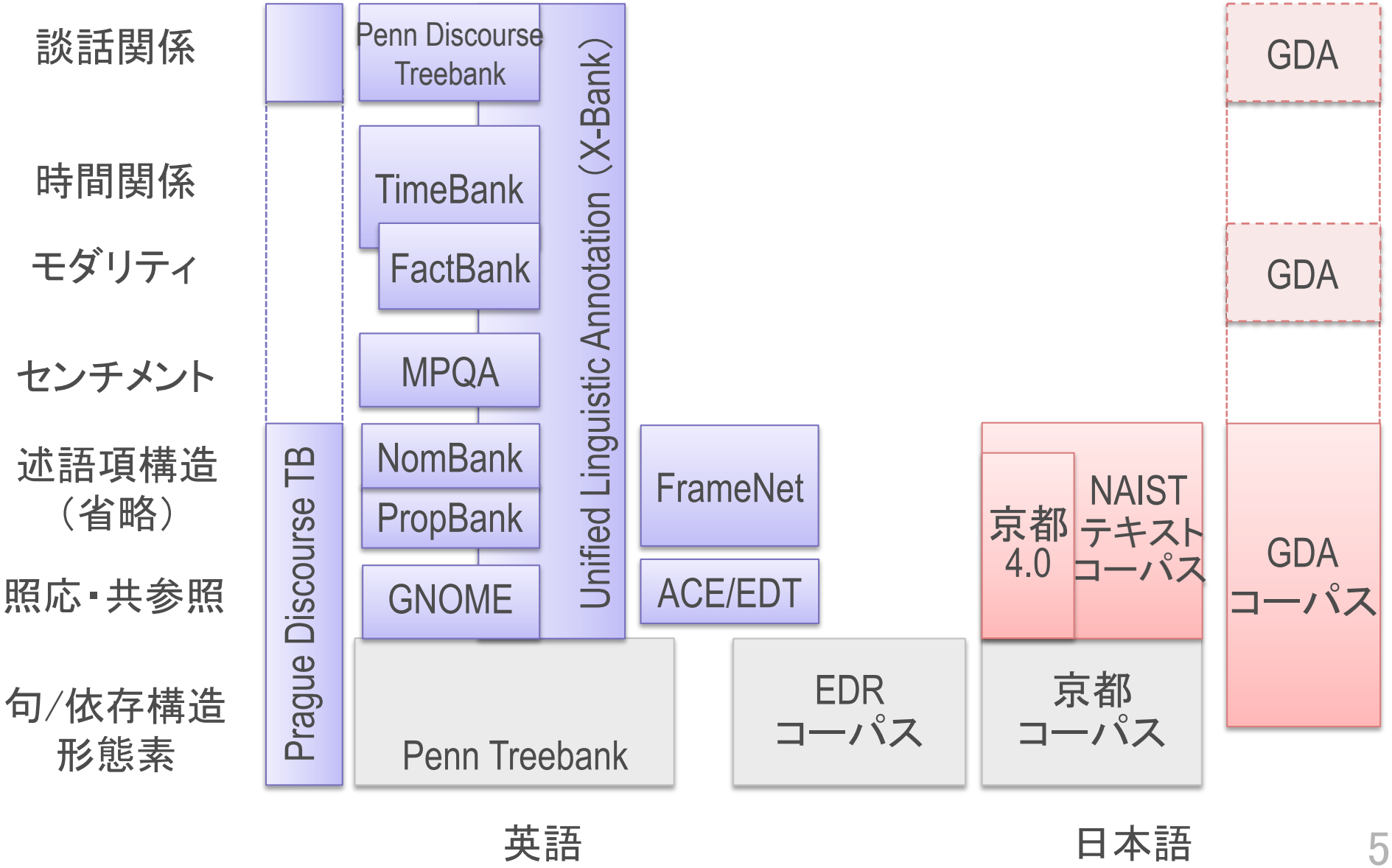
- 意味的に注釈付けされたコーパス
  ⇒ 意味解析研究のための共有資源
- 意味的注釈の仕様を設計することは
  ⇒ 意味解析の部分タスクを設計すること
  ⇒ 言語理論を実データで検証すること



講演概要

1. 意味情報の注釈付け
   - 何を何のために？
2. 海外でのコーパス構築の動向
3. 仕様設計の課題
   - 述語項構造，照応・共参照，モダリティ

注釈付きコーパスの例


談話関係
時間関係
モダリティ
センチメント
述語項構造
（省略）
照応・共参照
句/依存構造
形態素
Penn Discourse Treebank
TimeBank
FactBank
MPQA
NomBank
PropBank
GNOME
Unified Linguistic Annotation（X-Bank）
Prague Discourse TB
Penn Treebank
FrameNet
ACE/EDT
EDR コーパス
京都 4.0
NAIST テキスト コーパス
京都 コーパス
GDA
GDA
GDA コーパス
英語
日本語

計算言語学者が牽引

- C. Fillmore（認知言語学，フレーム意味論）
  - C. Baker, C. Fillmore, and J. Lowe. The Berkeley FrameNet project. COLING/ACL, 1998.
  - N. Ide, C. Baker, C. Fellbaum, C. Fillmore, and R. Passonneau. MASC: The manually annotated sub-corpus of American English. LREC, 2008.
- J. Pustejovsky（語彙意味論，生成語彙論）
  - J. Pustejovsky, A. Meyers, M. Palmer, and M. Poesio. Merging PropBank, NomBank, TimeBank, Penn Discourse Treebank and Coreference. Workshop on Frontiers in Corpus Annotation II, 2005.
  - R. Sauri, J. Littman, B. Knippen, R. Gaizauskas, A. Setzer, and J. Pustejovsky. TimeML Annotation Guidelines Version 1.2.1. 2006.
  - R. Sauri and J. Pustejovsky. FactBank: A corpus annotated with event factuality. Language Resources and Evaluation, 2009.

計算言語学者が牽引

- Prague Tectogrammatics

1. **Eva Hajicova**. (1993). Issues of Sentence Structure and Discourse Patterns. Prague: Charles University.
   Available in: BibTex item
2. **Eva Hajicová** (1998a). Movement Rules Revisited. In: Processing of Dependency-Based Grammars, *Proceedings from the Workshop COLING/ACL*, Montreal, ed. S. Kahane and A. Polguere, 49-57.
   Available in: BibTex item
3. **Eva Hajicova, Marketa Ceplova**. (2000). Deletions and Their Reconstruction in Tectogrammatical Syntactic Tagging of Very Large Corpora. In *Proceedings of COLING'2000*, pp. 228-284, Saarbruecken, Germany.
   Available in: BibTex item
4. **Eva Hajicova, Jarmila Panevova**. (1984). Valency (case) frames of verbs. In: Sgall (1984:147-188).
   Available in: BibTex item
5. **Eva Hajicova, Barbara Partee, Petr Sgall** (1998): *Topic-focus articulation, tripartite structures, and semantic content*. Amsterdam:Kluwer
   Available in: BibTex item
6. **Marcus M. P., Kim G., Marcinkiewicz M. A. et al.** (1994). The Penn Treebank: Annotating Predicate Argument Structure. Proceedings of the ARPA Human Language Technology Workshop. San Francisco: Morgan Kaufmann.
7. **Marcus M. P., Santorini B. and Marcinkiewicz M. A.** (1993). Building a Large Annotated Corpus of English: the Penn Treebank. Computational Linguistics, 19(2), 313-330.
8. **Jarmila Panevova**. 1974. "On verbal frames in Functional Generative Description". *Prague Bulletin of Mathematical Linguistics* 22:3-40; 23(1975):17-52.
   Available in: BibTex item
9. **Jarmila Panevova**. 1980. *Formy a funkce ve stavbe ceske vety*. [Forms and Functions in the Structure of the Czech Sentence]. Prague: Academia.
   Available in: BibTex item
10. **Vladimir Petkevic** (1987). A New Dependency Based Specification of Underlying Representations of Sentences. *Theoretical Linguistics* 14:143-172.
    Available in: BibTex item
11. **Vladimir Petkevic.** (1995). A New Formal Specification of Underlying Representations. *Theoretical Linguistics* 21:7-61.
    Available in: BibTex item
12. **Petr Sgall.** (1967). *Generativni popis jazyka a ceska deklinace*. [Generative Description of Czech and Czech Declension.] Prague: Academia.
    Available in: BibTex item
13. **Petr Sgall** ed. (1984). Contributions to Functional Syntax, Semantics and Language Comprehension. Amsterdam: Benjamins - Prague: Academia.
    Available in: BibTex item
14. **Petr Sgall.** 1992. Underlying Structure of Sentences and Its Relations to Semantics. *Wiener Slawistischer Almanach*. Sonderband 33. Ed. by T. Reuther. Wien: Gesellschaft zur Förderung slawistischer Studien, 273-282.
    Available in: BibTex item
15. **Petr Sgall.** (1997a). Valency and Underlying Structure. An Alternative View on Dependency. In: L. Wanner (ed.): *Recent Trends in Meaning-Text Theory*. Amsterdam/Philadelphia: Benjamins, 149-166.
    Available in: BibTex item
16. **Petr Sgall** (1997b). On the Usefulness of Movement Rules. In: Caron B. (ed.), *Actes du 16e Congres International des Linguistes* (Paris 20-25 juillet 1997), Oxford: Elsevier Sciences.
    Available in: BibTex item
17. **Petr Sgall** (in press). The Freedom of Language. To appear in *Prague Linguistic Circle Papers* 4.
18. **Petr Sgall, Eva Hajicova, Jarmila Panevova** (1986): *The Meaning of the Sentence in Its Semantic and Pragmatic Aspects*, ed. by J. L. Mey, Dordrecht:Reidel - Prague: Academia.

http://ufal.mff.cuni.cz/pdt/Corpora/PDT_1.0/Doc/tect.html

PropBank （Palmer, Gildea and Kingsbury, 2005）

- 述語項構造（意味役割）

Babko-Malaya. Propbank annotation guidelines, 2005.

[ARG0 John] broke [ARG1 the window]
[ARG1 The window] broke

Mr. Bush met him privately, in the White House,

Rel: met
Arg0: Mr. Bush
Arg1: him
ArgM-MNR: privately
ArgM-LOC: in the White House

PropBank (Palmer, Gildea and Kingsbury, 2005)

Babko-Malaya. Propbank annotation guidelines, 2005.

He was accused of conducting illegal business.

Treebank annotation:

```
(S (NP-SBJ-1 he)
  (VP was
    (VP accused
      (NP-3 *-1)
      (PP-CLR of
          (S-NOM (NP-SBJ *-3)
              (VP (VP conducting
                  (NP illegal business)))))))
```

Propbank annotation:

Arg1: [NP-3 *-1] -> he
Rel: accused
Arg2: of [*3*] conducting illegal business

TimeBank

（Sauri, Littman, Knippen, Gaizauskas, Setzer and Pustejovsky, 2006）

Newspaper reports have **said** Amir was infatuated with Har-Shefi and **may** have been **trying** to impress her by killing the prime minister.

R. Sauri. FactBank 1.0 Annotation Guidelines, 2008

Newspaper reports have
<EVENT eid="e22" class="REPORTING" tense="PRESENT" aspect="PERFECTIVE">
said </EVENT>
Amir was
<EVENT eid="e23" class="STATE" tense="PAST">
infatuated </EVENT>
with Har-Shefi and may have been
<EVENT eid="e24" class="I_ACTION" modality="may" tense="NONE" aspect="PERF_PROG">
trying </EVENT>
to
<EVENT eid="e25" class="OCCURRENCE" tense="INFINITIVE">
impress </EVENT>
her by
<EVENT eid="e26" class="OCCURRENCE" tense="PRESPART" aspect="NONE">
killing </EVENT> the prime minister.
<SLINK lid="l50" eventId="e22" subordinatedEventId="e23" relType="EVIDENTIAL"/>
<SLINK lid="l51" eventId="e22" subordinatedEventId="e24" relType="EVIDENTIAL"/>
<SLINK lid="l52" eventId="e24" subordinatedEventId="e25" relType="MODAL"/>

FactBank （Sauri and Pustejovsky, 2009）

Newspaper reports have **said** Amir was infatuated with Har-Shefi and **may** have been **trying** to impress her by killing the prime minister.

	Positive	**Negative**	**Underspecified**
Certain	Fact: <CT,+>	Counterfact: <CT,−>	Certain but unknown output: <CT, u>
Probable	Probable: <PR,+>	Not probable: <PR,−>	(NA)
Possible	Possible: <PS,+>	Not certain: <PS,−>	(NA)
Underspecif.	(NA)	(NA)	Unknown or uncommitted: <U,u>

R. Sauri. FactBank 1.0 Annotation Guidelines, 2008

FactBank （Sauri and Pustejovsky, 2009）

Newspaper reports have **said** Amir was infatuated with Har-Shefi and **may** have been **trying** to impress her by killing the prime minister.

Event (ID):	**Source (ID):**	**Fact. value:**
said (e22)	*author* (s_0)	CT+
infatuated (e23)	*reports_author* (s_2_s_0)	CT+
	author (s_0)	Uu
trying (e24)	*reports_author* (s_2_s_0)	PS+
	author (s_0)	Uu
impress (e25)	*reports_author* (s_2_s_0)	Uu
	author (s_0)	Uu
killing (e26)	*reports_author* (s_2_s_0)	Uu
	author (s_0)	Uu

R. Sauri. FactBank 1.0 Annotation Guidelines, 2008

Penn Discourse Treebank

(Miltsakaki, Prasad, Joshi and Webber, 2004)

Discourse relation

Michelle lives in a hotel room, *and* although **she drives a canary-colored Porsche**, *she hasn't time to clean or repair it*. (2402)

Most oil companies, when **they set exploration and production budgets for this year**, *forecast revenue of $15 for each barrel of crude produced*.[4] (0725)

Also unlike Mr. Ruder, Mr. Breeden appears to be in a position to get somewhere with his agenda. Implicit = BECAUSE (CAUSE) **As a former White House aide who worked closely with Congress, he is savvy in the ways of Washington**. (0955)

Prasad, et al. Attribution and its annotation in the Penn Discourse TreeBank. L'object, 2004.

Penn Discourse Treebank

(Miltsakaki, Prasad, Joshi and Webber, 1994)

Attribution

Arg2

When Mr. Green won a $240,000 verdict in a land condemnation case against the State in June 1983, he says ***Judge O'Kicki unexpectedly awarded him an additional $100,000.*** (0267)

Arg1

	REL	Arg1	Arg2
[Source]	Wr	Ot	Inh

Arg1

"*Having the dividend increases is a supportive element in the market outlook,* but I don't think **it's a main consideration,"** he says. (0090)

Arg2

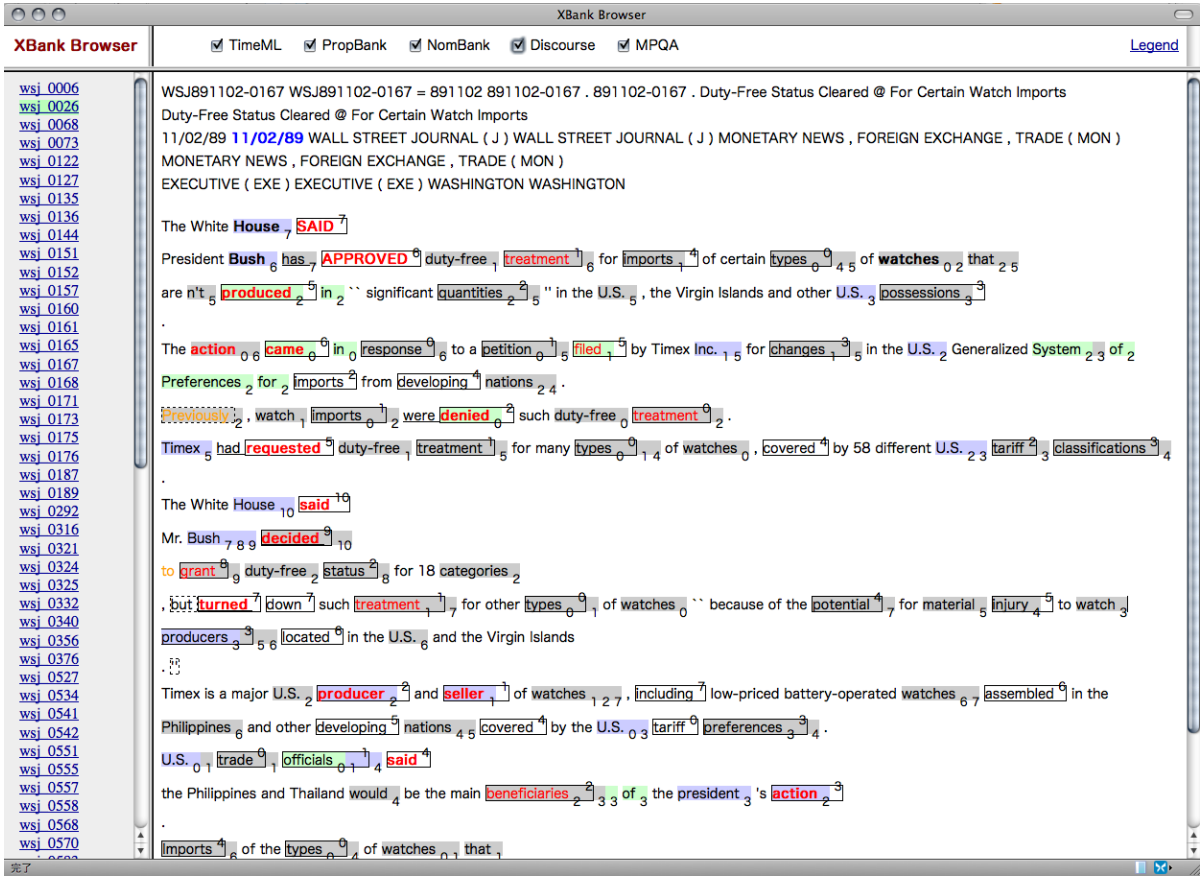
	REL	Arg1	Arg2
[Source]	Ot	Inh	Ot
[Type]	Comm	Null	PAtt
[Polarity]	Null	Null	Neg

Communication, Belief, Fact, Eventuality

Prasad, et al. Attribution and its annotation in the Penn Discourse TreeBank. L'object, 2004.

XBank

- PropBank, NomBank, TimeBank, Discourse, MPQAを統合



http://timeml.org/ula/xbank-browser/

照応・共参照 と 述語項構造 の組合せ

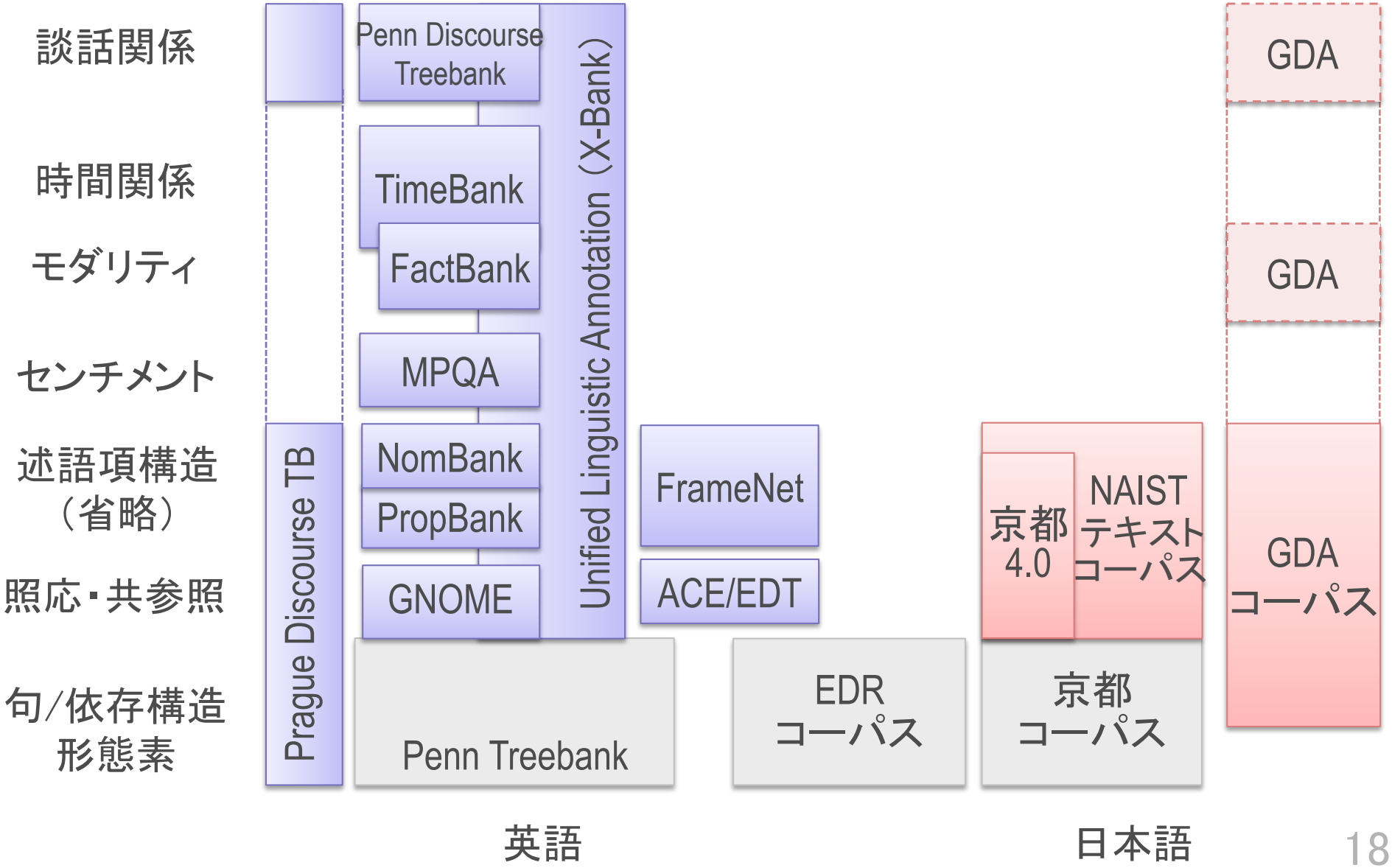
- 述語の項（ゼロ照応）の先行詞が複数ある場合

就任後初めて地元の大分県へ里帰りしていた村山富市首相$_i$ は三十一日夕，三泊四日の日程を終えて日航機で羽田空港に到着した．首相$_i$ は記者団に対し，「突然大分に帰った$_{ガ:i}$ が，温かい歓迎に接し$_{ガ:i}$『地元はいいなあ』という感謝$_{ガ:i}$ の気持ちでいっぱい$_{ガ:i}$．期待に応え$_{ガ:i}$ てしっかり頑張ら$_{ガ:i}$ないといかんという気持ちを一層強く持った$_{ガ:i}$ 」と感想を述べ$_{ガ:i}$ た．

講演概要

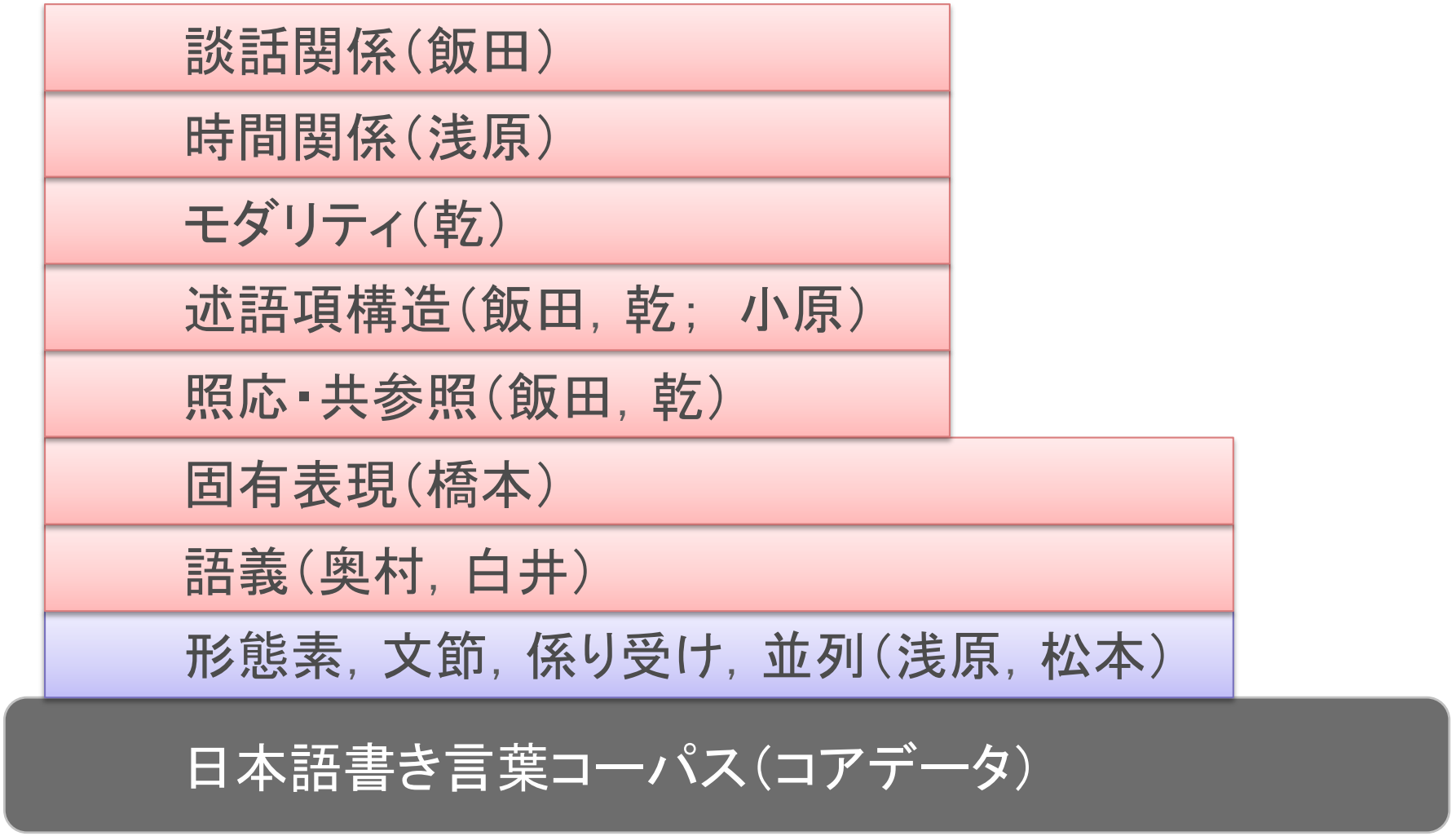
1. 意味情報の注釈付け
   - 何を何のために？
2. 海外でのコーパス構築の動向
3. 仕様設計の課題
   - 述語項構造，照応・共参照，モダリティ

注釈付きコーパスの例


談話関係
時間関係
モダリティ
センチメント
述語項構造
（省略）
照応・共参照
句/依存構造
形態素
Penn Discourse Treebank
TimeBank
FactBank
MPQA
NomBank
PropBank
GNOME
Unified Linguistic Annotation（X-Bank）
Prague Discourse TB
Penn Treebank
FrameNet
ACE/EDT
EDR コーパス
京都 4.0
NAIST テキスト コーパス
京都 コーパス
GDA
GDA
GDA コーパス
英語
日本語

科研特定領域「日本語コーパス」(2006-2011)

- 様々なレベルのアノテーションが進行/計画中

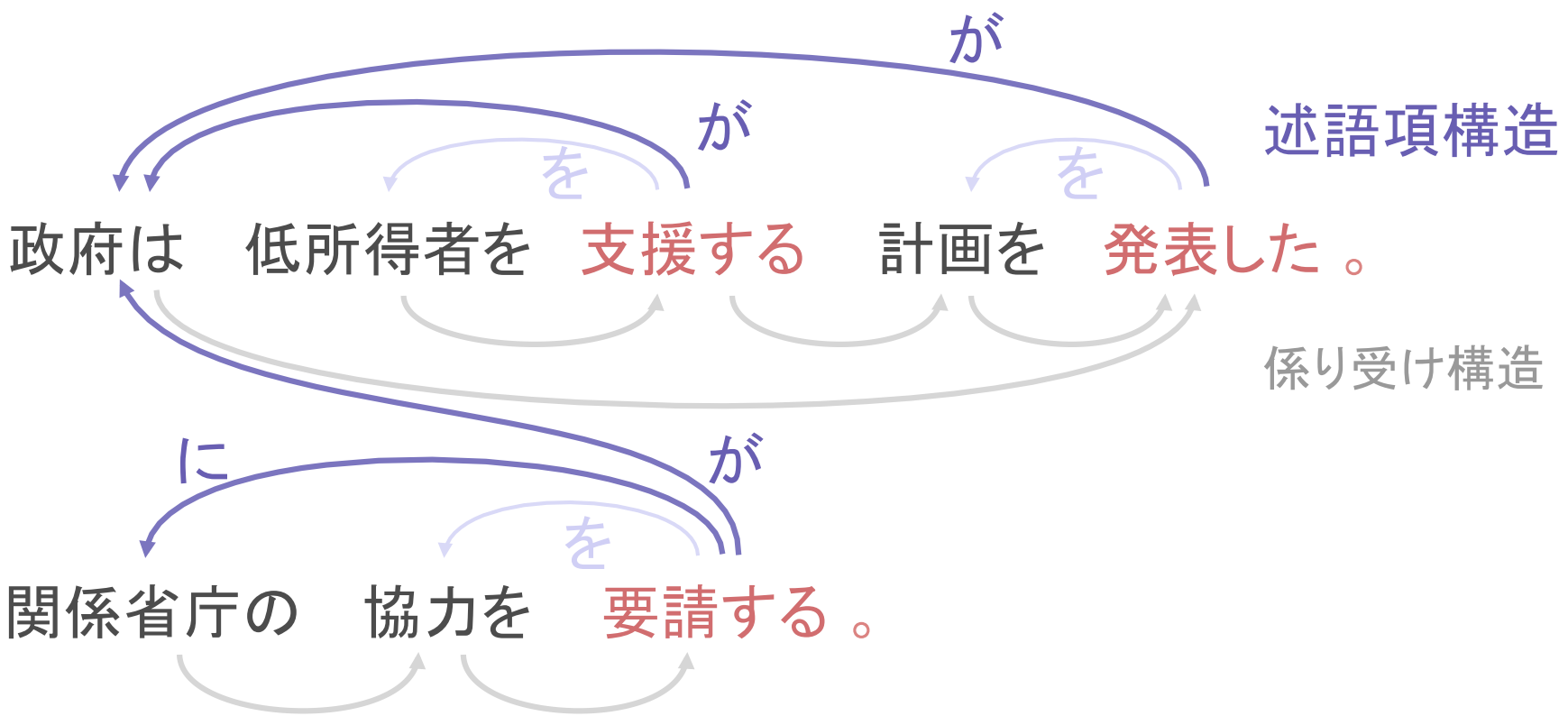


NAISTテキストコーパス

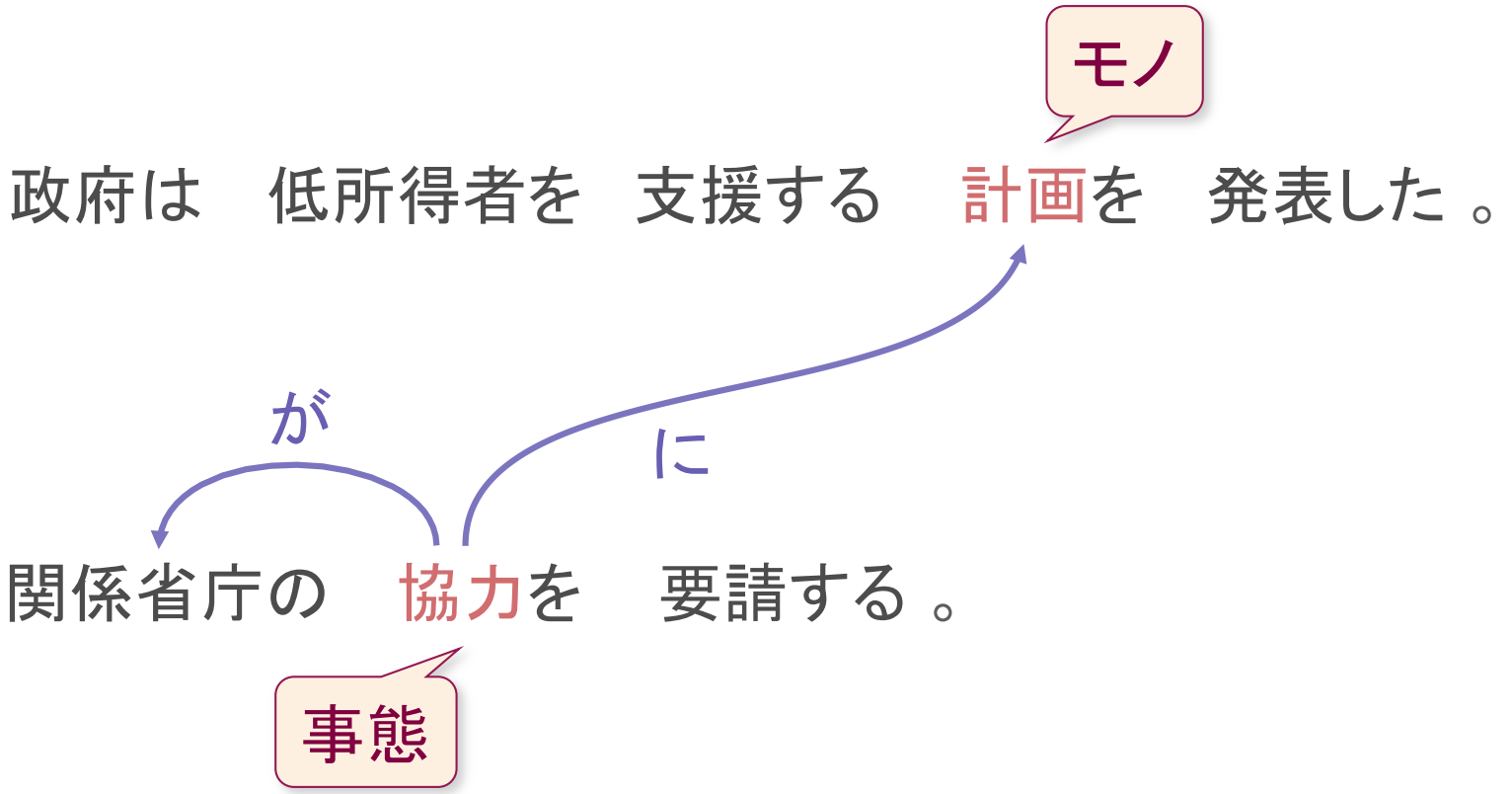
- 京都コーパス全文に述語項構造・共参照を，一部に間接照応をタグ付け

毎日新聞 2,929記事（38,384文）				
		ガ格	ヲ格	ニ格
述語 106,628	同一文節内	177 (0.002)	60 (0.001)	591 (0.027)
	係り関係	44,402 (0.419)	35,882 (0.835)	18,912 (0.879)
	ゼロ照応（文内）	32,270 (0.305)	5,625 (0.131)	1,417 (0.066)
	ゼロ照応（文間）	13,181 (0.124)	1,307 (0.030)	542 (0.025)
	ゼロ照応（文章外）	15,885 (0.150)	96 (0.002)	45 (0.002)
	全体	105,915 (1.000)	42,970 (1.000)	21,507 (1.000)
事態性名詞 28,569	同一文節内	2,195 (0.077)	5,574 (0.506)	846 (0.436)
	係り関係	4,332 (0.152)	2,890 (0.263)	298 (0.154)
	ゼロ照応（文内）	9,222 (0.324)	1,645 (0.149)	586 (0.302)
	ゼロ照応（文間）	5,190 (0.183)	854 (0.078)	201 (0.104)
	ゼロ照応（文章外）	7,525 (0.264)	42 (0.004)	10 (0.005)
	全体	28,464 (1.000)	11,005 (1.000)	1,941 (1.000)
共参照		25,764		

述語項構造（省略/ゼロ照応）



述語項構造（省略/ゼロ照応）



照応・共参照

横尾$_i$は画家でもないし、デザイナーでもない。
そんなことは彼$_i$にとってはどうでもよいことなのだ。

間接照応（bridging reference）

5 年間、水質調査を行った。このデータは機械的に処理される。

さまざまな課題

- 述語項構造
  - 真に曖昧な場合の扱い

… 自民、さきがけ、新進各党の与野党の党首会談を呼び掛けて協力を求めるべきだ。

(A) 与野党 ガ 協力する
(B) (与野党の)党首 ガ 協力する

さまざまな課題

- 事象性名詞の項構造
  - イベント か モノ（結果物，内容）か？
  - 結果物に項を認めるか？

文化庁の 2005 年の報告によると、各宗教団体の報告による信者数は合計 2 億 1100 万人 である。

文化庁 ガ 報告する（？）

さまざまな課題

- 事象性名詞の項構造
  - イベント か モノ（結果物，内容）か？
  - 結果物に項を認めるか？

文化庁の 2005 年の報告によると、各宗教団体の報告による信者数は合計 2 億 1100 万人 である。

党内には「社会党会派の離脱者は従来通り除名すべきだ」との意見が根強く…

また、経済問題については日本経済の構造変革のため規制緩和に積極的に取り組むと訴える。

さまざまな課題

- genericな名詞句間の照応・共参照関係をどのように規定するか？

フロンによる環境破壊への対策が地球的規模の課題となって久しい。特に、フロンがオゾン層を破壊することが報告されてから、…

フロンによる環境破壊への対策が地球的規模の課題となって久しい。特に、この物質がオゾン層を破壊することが報告されてから、

兵庫県内の暗やみの中で、人々が水と食べ物の不足に苦しんでいる同じ夜、隣接した大阪の繁華街ではネオンが光り、飲食店はにぎわっている。水も食料も、被災地を離れるとふんだんにある。

広義のモダリティ

- テキスト中の各事象表現にテンス，アスペクト，極性，モダリティ等の情報を付与（自然言語処理研究会，2009.9）
  - 仕様書：http://cl.naist.jp/nltools/modality/

これからは酒を飲むのを控えようと思います。

態度表明者	時制	仮想	態度	真偽判断	価値判断	焦点
書き手	未来	ー	意志	高確率から低確率	ネガティブ	ー

販売開始のめどが立たない状況に陥っている。

態度表明者	時制	仮想	態度	真偽判断	価値判断	焦点
書き手	未来	ー	叙述	低確率	ー	ー

全員がこの案に賛成しているというわけではない。

態度表明者	時制	仮想	態度	真偽判断	価値判断	焦点
書き手	非未来	ー	叙述	成立	ー	否定（全員）

さまざまな課題

- 拡張モダリティ
  - 否定のスコープ，部分否定，程度をどう扱うか？

否定の対象

薬を飲んだから元気になったわけではない。

成立　　成立

中村はあまり酒を飲まない。

還元水は体内の活性酸素を消去するのにはほとんど役に立ちません。

まとめ

- コーパスへの意味情報付与の動向
  - 述語項構造，モダリティ，照応・共参照，談話関係
  - 欧米では計算言語学者が牽引
- 意味情報付与はまだ仕様設計の模索段階
  - 言語処理にとっても言語研究にとっても興味深い研究課題の宝庫
  - 言語研究者との連携強化が不可欠
- 他のコーパス，レイヤとの相互連携性も課題