日本語複合辞の言語処理のための辞書・用例データベースの設計と作成

宇津呂武仁
筑波大学大学院　システム情報工学研究科
知能機能システム専攻

松吉 俊
京都大学大学院 情報学研究科 知能情報学専攻
黒橋研究室 博士後期課程3回生 （実質的には名大佐藤研）

土屋　雅稔
豊橋技術科学大学　情報メディア基盤センター

本日の内訳

1. プロジェクト全体の紹介（宇津呂）
2. 日本語複合辞の階層的辞書の編纂（松吉）
   （おまけ）複合辞の言い換え
3. 日本語複合辞用例データベースの作成（土屋）
4. 日本語文中の複合辞の言語解析
   4-1: 複合辞の検出（土屋）
   4-2: 統計的係り受け解析（宇津呂）
5. その他の周辺の話題・進行中の研究（宇津呂）

研究グループメンバーと研究経過

2003～2004年度

- 京大情報・知能情報・言語メディア研
  佐藤理史（現名大）、宇津呂、
  土屋（2003年度D単位認定⇒豊橋助手）、松吉（M取得）、他

2005年度

- 京大：松吉（D）、注連隆夫（M、検出・係り受け）
- 佐藤：京大⇒名大

2006年度

- 宇津呂：京大⇒筑波大、土屋：D取得、注連：M取得⇒NEC関西研

2007年度

- 松吉：D取得見込⇒某所ポスドク予定
- 筑波大宇津呂研：長坂（B4、全機能表現の検出）
  坂本（来年度M1、日英翻訳）

2008年度：宇津呂研上記二名M1

本日の内訳

1. プロジェクト全体の紹介（宇津呂）
2. 日本語複合辞の階層的辞書の編纂（松吉）
   （おまけ）複合辞の言い換え
3. 日本語複合辞用例データベースの作成（土屋）
4. 日本語文中の複合辞の言語解析
   4-1: 複合辞の検出（土屋）
   4-2: 統計的係り受け解析（宇津呂）
5. その他の周辺の話題・進行中の研究（宇津呂）

研究背景

・機能表現

幾つかの語から構成される、機能的な意味をもった
ひとまとまりの表現

- 助詞、助動詞とともに、文の構造を形作る要素

「について」

（内容的用法）　私は、彼の車について走った。

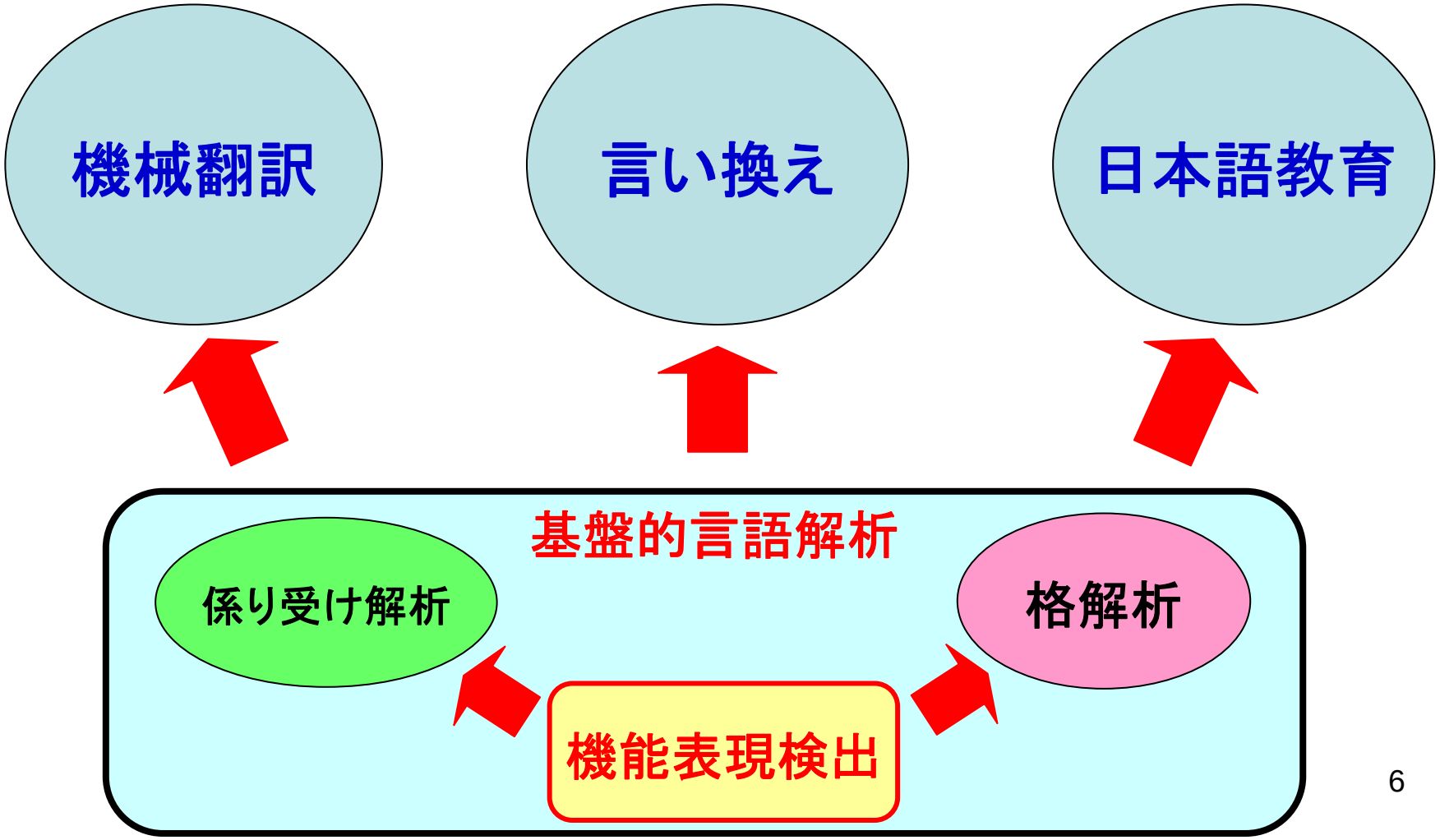
⇒　I drove following his car.

（機能的用法）　私は、自分の夢について話した。

⇒　I talked about my dream.

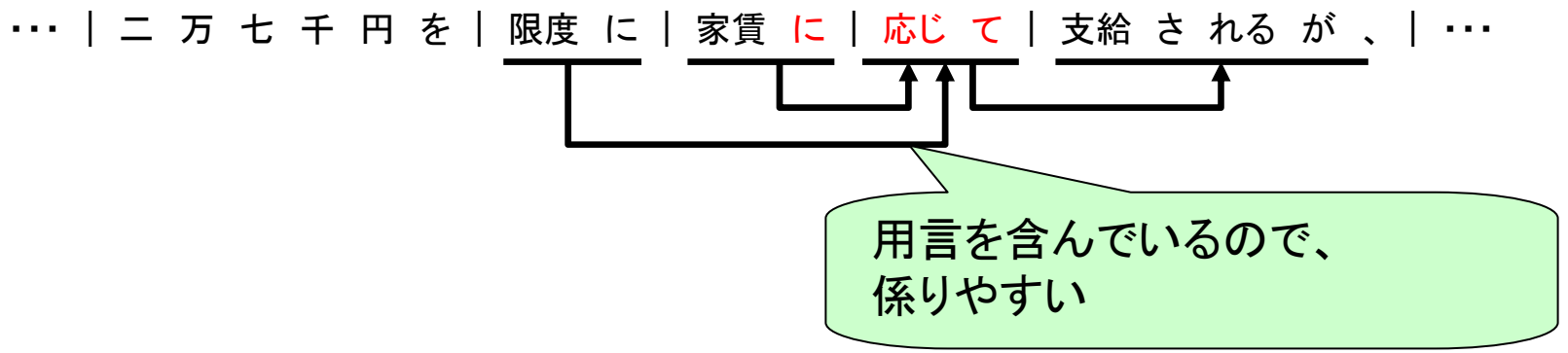
機能的用法と内容的用法とを、正しく判別する必要がある。

機能表現の基盤的言語解析 およびその応用

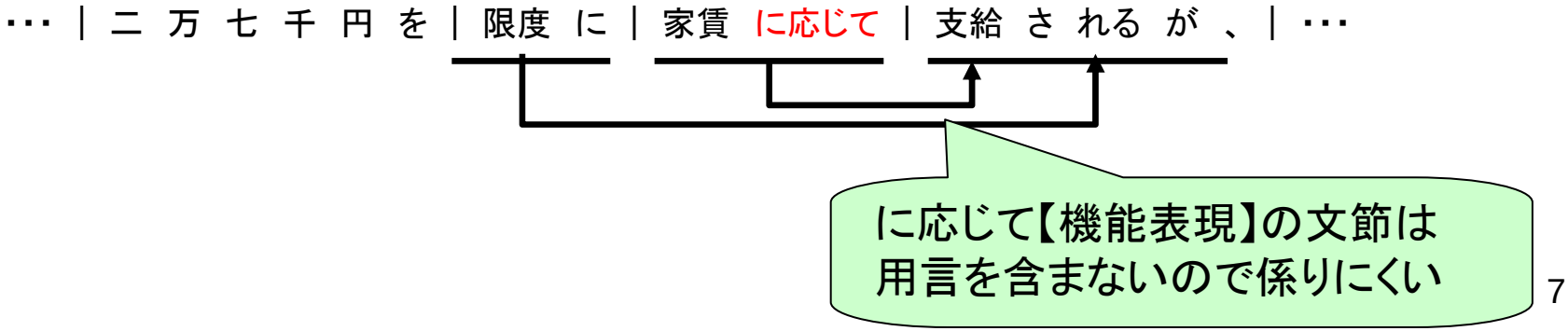


機能表現を考慮した係り受け解析

- 機能表現を考慮すると，係り先の曖昧さが減少
  - 機能表現「～に応じて」を考慮しない場合

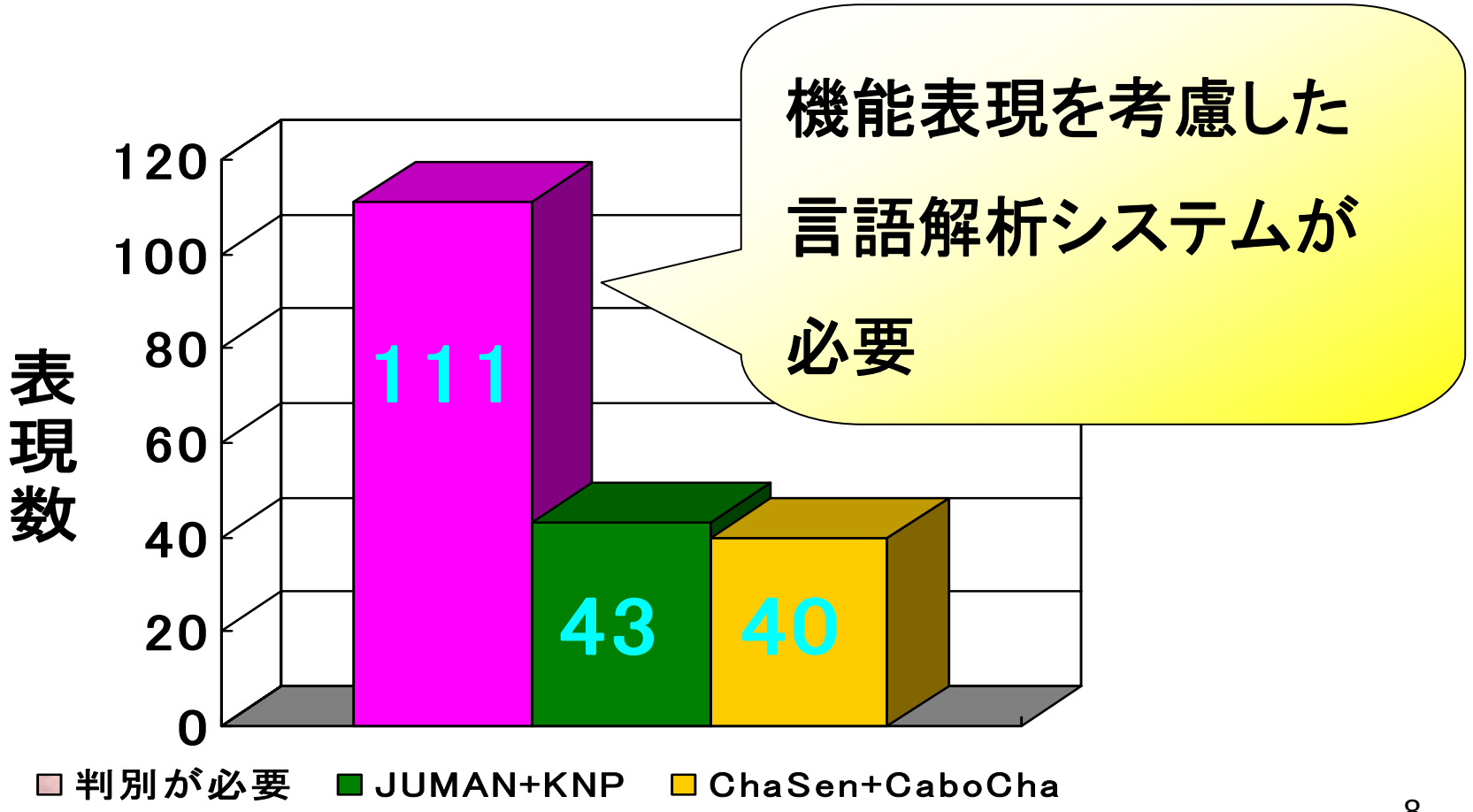


  - 機能表現「～に応じて」を考慮した場合



既存の解析系における扱い

- 日本語複合辞用例データベース 337表現中
  〔土屋2006〕



本日の内訳

1. プロジェクト全体の紹介（宇津呂）
2. 日本語複合辞の階層的辞書の編纂（松吉）
   （おまけ）複合辞の言い換え
3. 日本語複合辞用例データベースの作成（土屋）
4. 日本語文中の複合辞の言語解析
   4-1: 複合辞の検出（土屋）
   4-2: 統計的係り受け解析（宇津呂）
5. その他の周辺の話題・進行中の研究（宇津呂）

<table>
<tr><td colspan="2"></td><td>100表現
（機能的用法・
自立的用法が
バランス）</td><td>300表現
（「現代語複合辞
用例集」の
表現リストを展開）</td><td>17,000表現
（松吉リストにより収集済み）</td></tr>
<tr><td colspan="2">機能表現
リストの作成</td><td colspan="2" rowspan="2">[土屋，情処論文誌2006]
（毎日新聞1995年分から
50～100用例/表現を
収集・用法判定）</td><td>[松吉，NLPジャーナル2007]</td></tr>
<tr><td rowspan="3">用例
データベー
スの
作成</td><td>機能的・
自立的識別</td><td>19年度～の計画</td></tr>
<tr><td>係り受け関係付与</td><td colspan="2">[土屋，NLP大会2007]
（京都テキストコーパス中の
機能表現の用法判定・
係り受け関係誤り修正済み）</td><td rowspan="2"></td></tr>
<tr><td>意味分類付与</td><td colspan="2">20年度～の計画</td></tr>
<tr><td rowspan="4">日本語文中
の
機能表現の
言語解析</td><td>検出</td><td rowspan="2">[土屋・注連，
NLPジャーナル2007]</td><td colspan="2">19年度の計画
（派生表現を代表的表現で代用する方式の確立）</td></tr>
<tr><td>係り受け</td><td colspan="2" rowspan="3">20年度～の計画</td></tr>
<tr><td>格解析との統合</td></tr>
<tr><td>意味分類付与</td></tr>
<tr><td colspan="2">応用　（言い換え・
日本語教育・機械翻訳）</td><td colspan="3">言い換え：[松吉，NLP大会2007・NLPジャーナル2008]
その他：　19年度～の計画</td></tr>
</table>

各種情報・データ・スナップショット集

言語資源の公開状況

公開済み

- 日本語複合辞用例データベースv1.0

  337表現×50用例(国語研「現代語複合辞用例集」の125表現の派生表現を収録)

  – 毎日新聞1995年分を別途購入

  – http://nlp.iit.tsukuba.ac.jp/must/ から複合辞用法判定ラベル情報をダウンロード

今後の予定:

- 松吉「日本語機能表現一覧」
- 京都テキストコーパスへの用法判定ラベル付与情報
- 松吉「日本語機能表現一覧」中の表現の用例DB

参考文献

機能表現階層辞書

- 日本語機能表現辞書の編纂

  松吉俊, 佐藤理史, 宇津呂武仁.

  自然言語処理(言語処理学会論文誌), Vol.14, No.5, pp. 123-146, 2007

データベース本体に関する発表

- 日本語複合辞用例データベースの作成と分析

  土屋雅稔, 宇津呂武仁, 松吉俊, 佐藤理史, 中川聖一.

  情報処理学会論文誌, Vol. 47, No. 6, pp. 1728-1741, 2006.

データベースを利用した複合辞の検出及び係り受け解析に関する発表

- 日本語機能表現の自動検出と統計的係り受け解析への応用

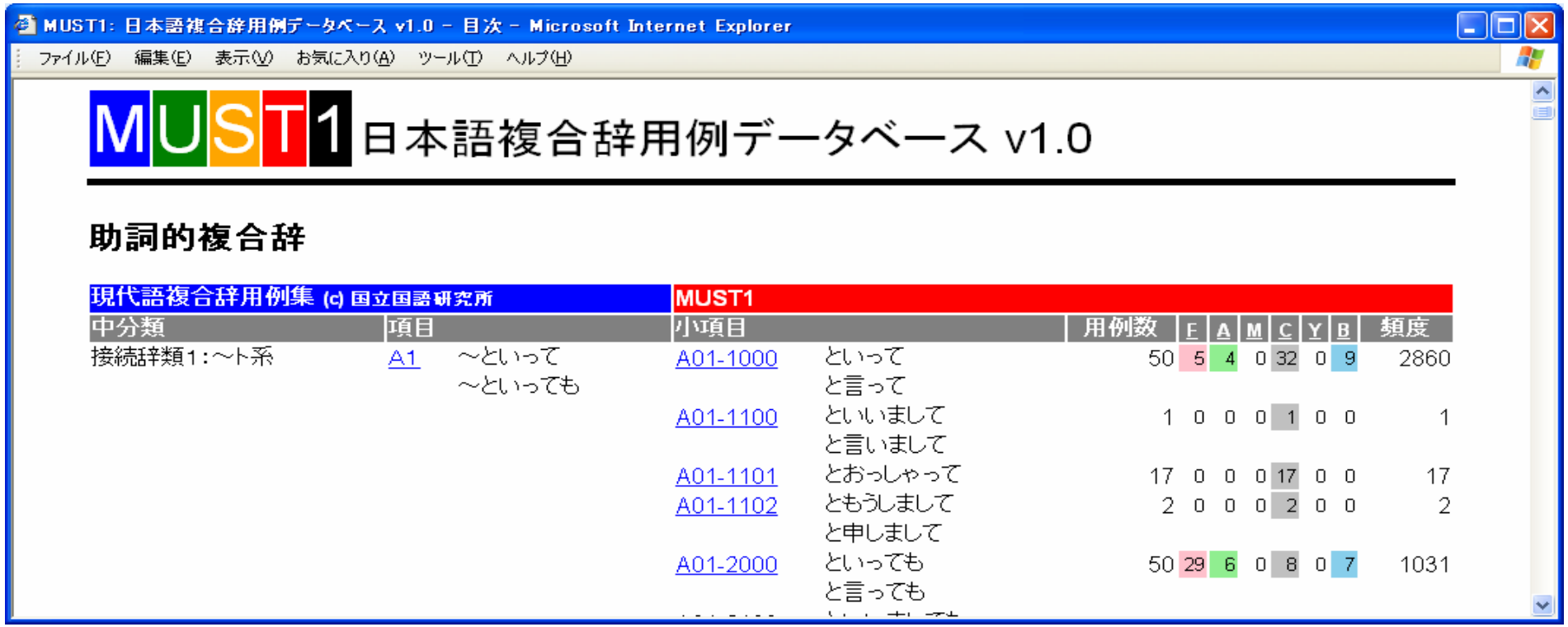
  注連隆夫, 土屋雅稔, 松吉俊, 宇津呂武仁, 佐藤理史.

  自然言語処理(言語処理学会論文誌), Vol.14, No.5, pp. 167-197, 2007

機能表現の言い換え

- 文体と難易度を制御可能な日本語機能表現の言い換え

  松吉俊, 佐藤理史. 自然言語処理(言語処理学会論文誌), 2008

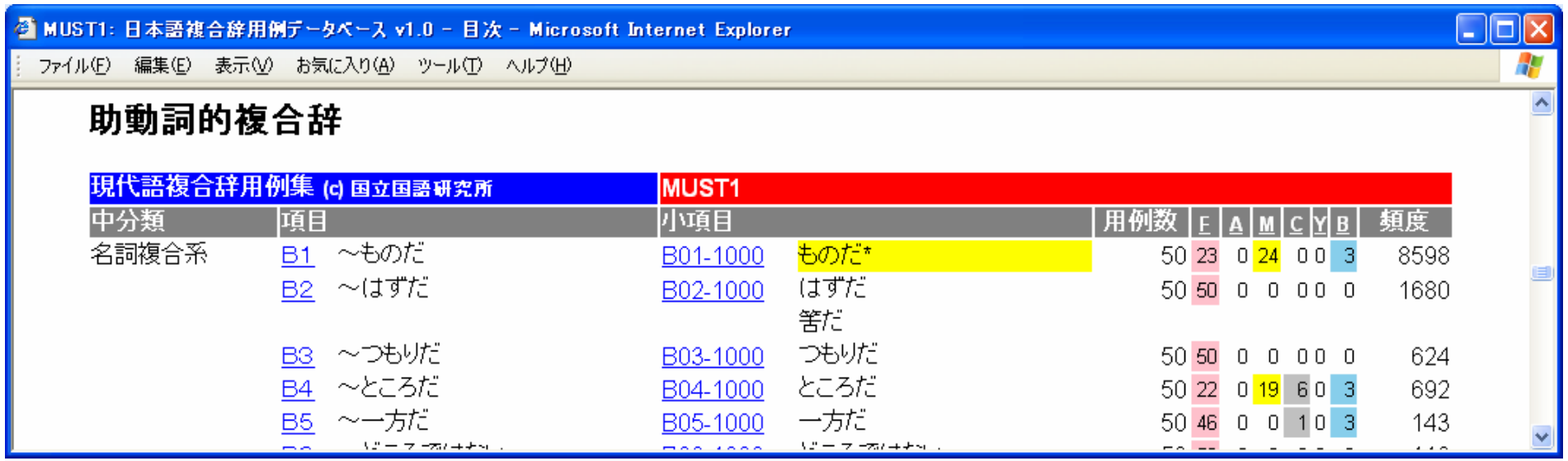

MUST1: 日本語複合辞用例データベース v1.0 - 目次 - Microsoft Internet Explorer

ファイル(F) 編集(E) 表示(V) お気に入り(A) ツール(T) ヘルプ(H)

MUST1 日本語複合辞用例データベース v1.0

助詞的複合辞

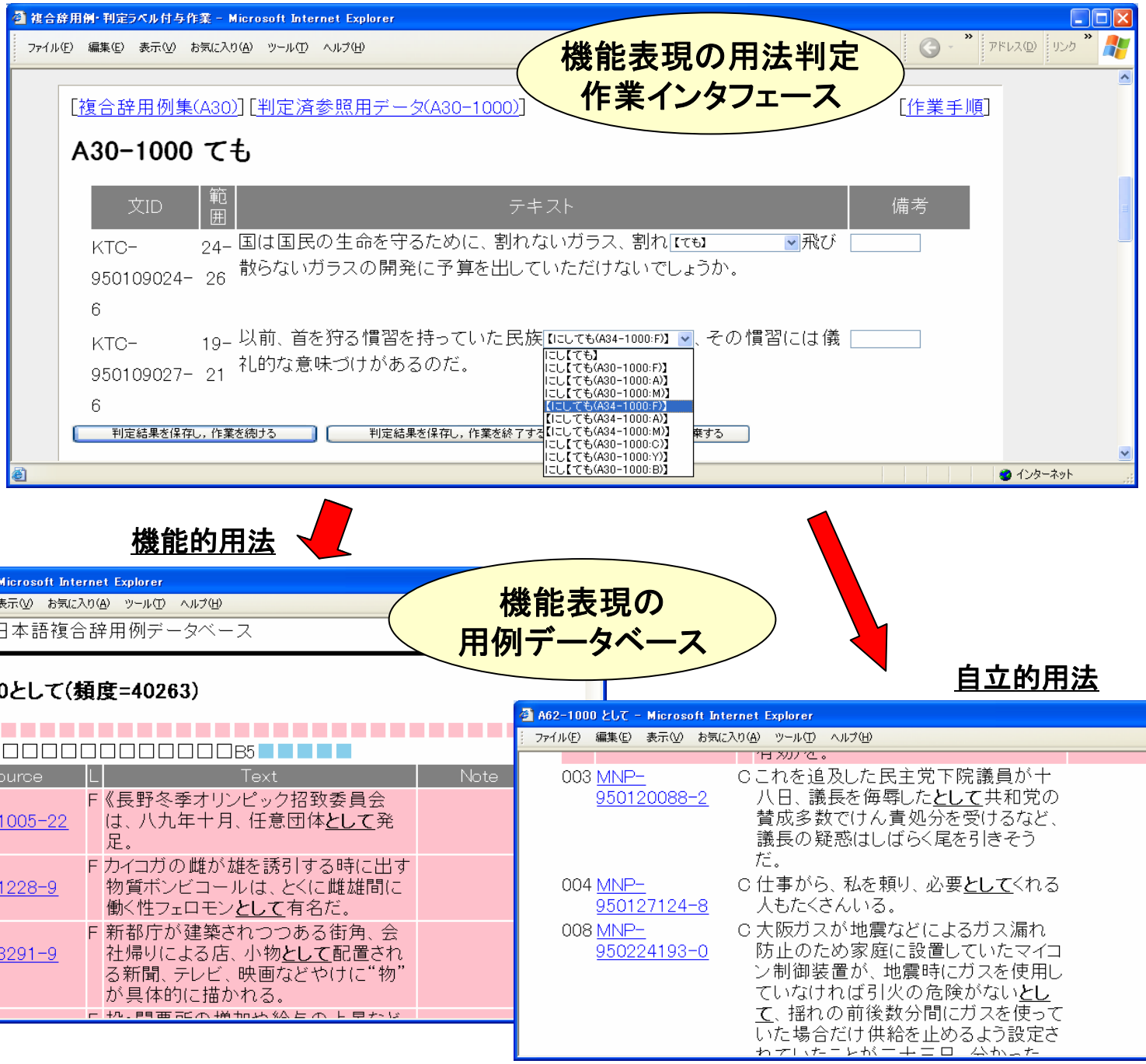
現代語複合辞用例集 (c) 国立国語研究所			MUST1										
中分類	項目		小項目		用例数	F	A	M	C	Y	B	頻度
接続辞類1：～ト系	A1	～といって ～といっても	A01-1000	といって と言って	50	5	4	0	32	0	9	2860
			A01-1100	といいまして と言いまして	1	0	0	0	1	0	0	1
			A01-1101	とおっしゃって	17	0	0	0	17	0	0	17
			A01-1102	ともうしまして と申しまして	2	0	0	0	2	0	0	2
			A01-2000	といっても と言っても	50	29	6	0	8	0	7	1031


MUST1: 日本語複合辞用例データベース v1.0 - 目次 - Microsoft Internet Explorer

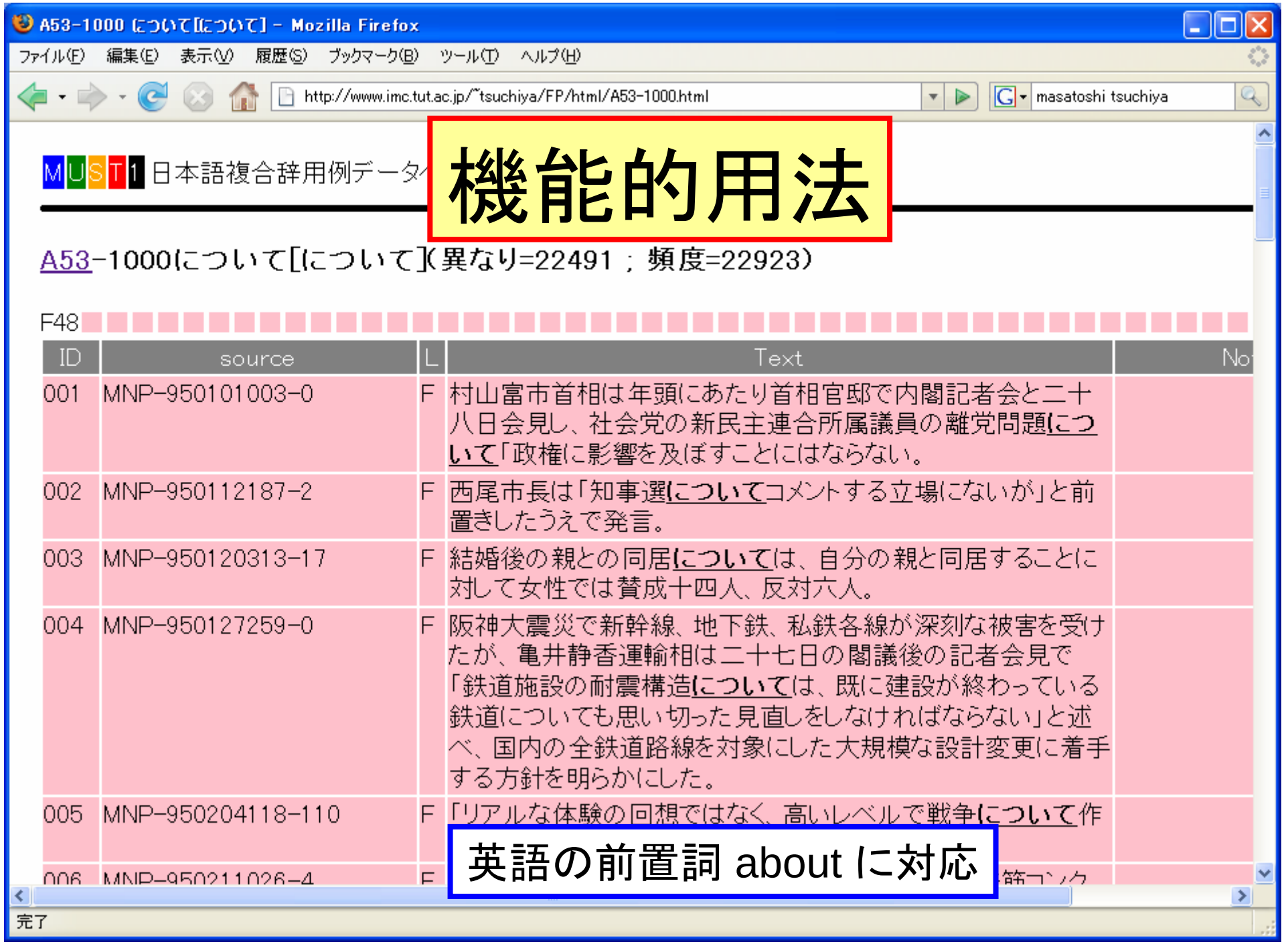
ファイル(F) 編集(E) 表示(V) お気に入り(A) ツール(T) ヘルプ(H)

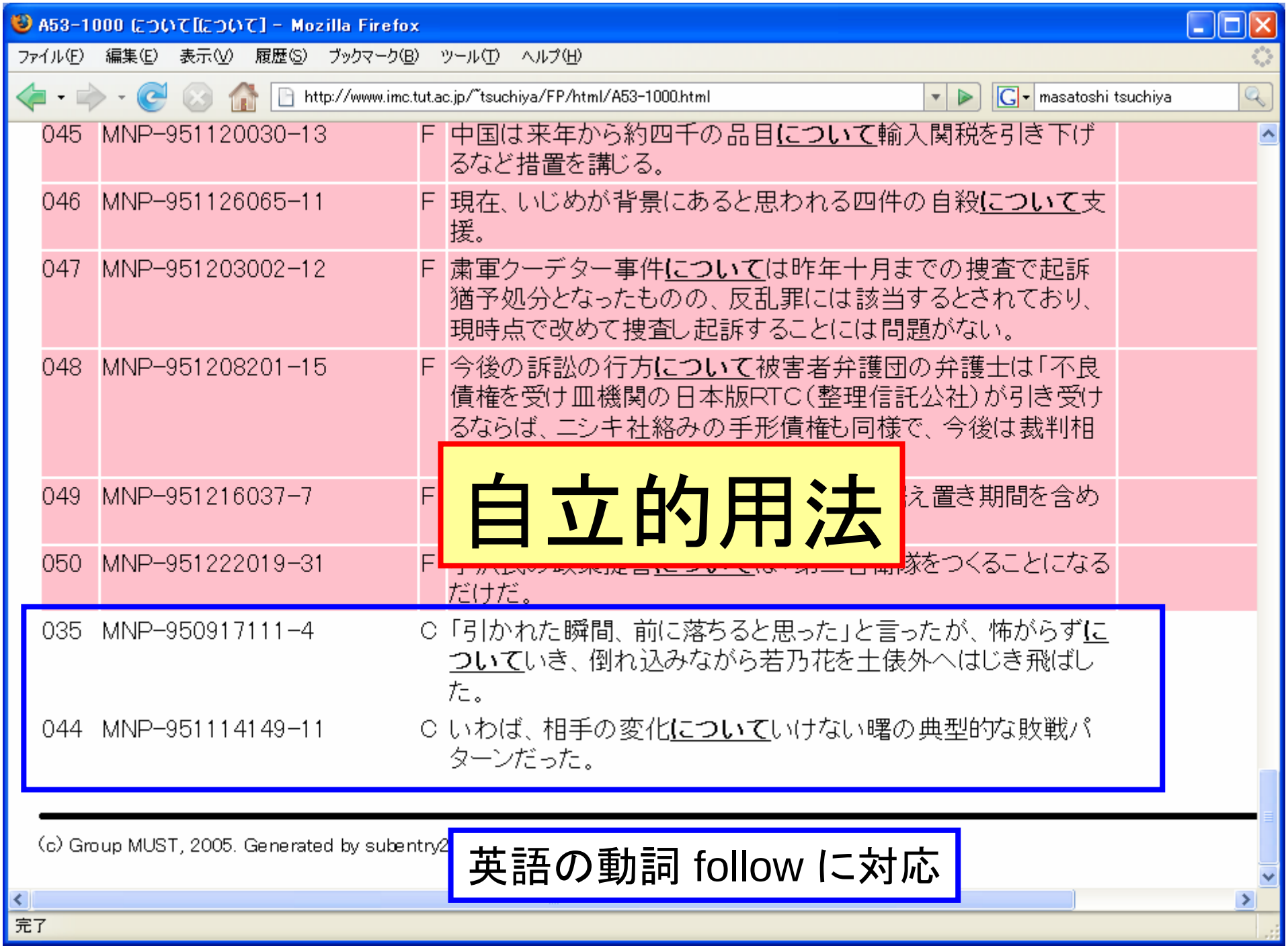
助動詞的複合辞

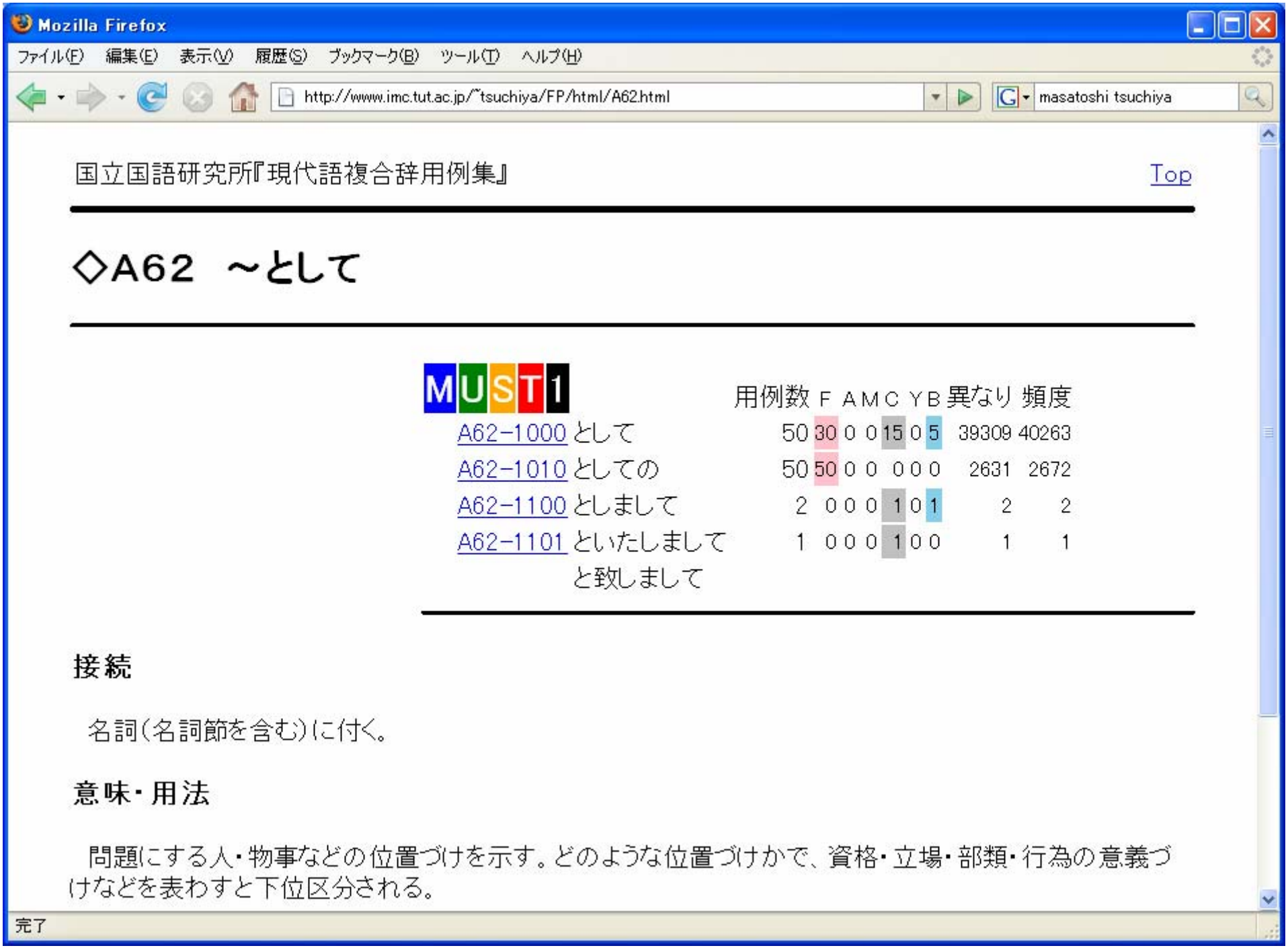
現代語複合辞用例集 (c) 国立国語研究所			MUST1										
中分類	項目		小項目		用例数	F	A	M	C	Y	B	頻度
名詞複合系	B1	～ものだ	B01-1000	ものだ*	50	23	0	24	0	0	3	8598
	B2	～はずだ	B02-1000	はずだ 筈だ	50	50	0	0	0	0	0	1680
	B3	～つもりだ	B03-1000	つもりだ	50	50	0	0	0	0	0	624
	B4	～ところだ	B04-1000	ところだ	50	22	0	19	6	0	3	692
	B5	～一方だ	B05-1000	一方だ	50	46	0	0	1	0	3	143

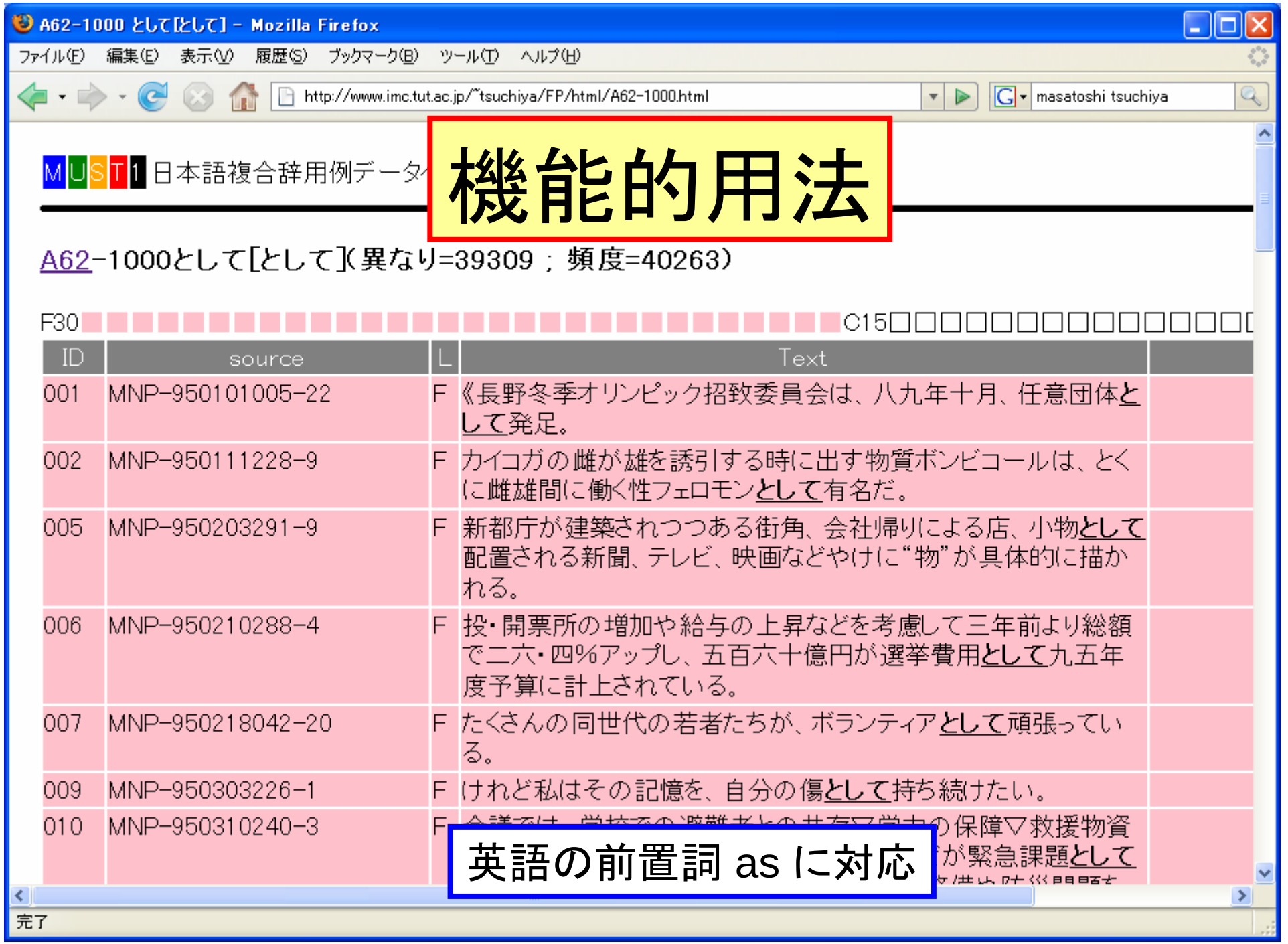

機能表現の用法判定
作業インタフェース
機能的用法
機能表現の
用例データベース
自立的用法
複合辞用例-判定ラベル付与作業 - Microsoft Internet Explorer
[複合辞用例集(A30)] [判定済参照用データ(A30-1000)]
[作業手順]
A30-1000 ても
文ID
範囲
テキスト
備考
KTC-950109024-6
24-26
国は国民の生命を守るために、割れないガラス、割れ【ても】飛び散らないガラスの開発に予算を出していただけないでしょうか。
KTC-950109027-6
19-21
以前、首を狩る慣習を持っていた民族【にしても(A34-1000:F)】、その慣習には儀礼的な意味づけがあるのだ。
にし【ても】
にし【ても(A30-1000:F)】
にし【ても(A30-1000:A)】
にし【ても(A30-1000:M)】
【にしても(A34-1000:F)】
【にしても(A34-1000:A)】
【にしても(A34-1000:M)】
にし【ても(A30-1000:C)】
にし【ても(A30-1000:Y)】
にし【ても(A30-1000:B)】
判定結果を保存し，作業を続ける
判定結果を保存し，作業を終了する
A62-1000 として - Microsoft Internet Explorer
MUST1 日本語複合辞用例データベース
A62-1000として(頻度=40263)
F30
C15□□□□□□□□□□□□□□□B5
ID
source
L
Text
Note
001 MNP-950101005-22
F 《長野冬季オリンピック招致委員会は、八九年十月、任意団体として発足。
002 MNP-950111228-9
F カイコガの雌が雄を誘引する時に出す物質ボンビコールは、とくに雌雄間に働く性フェロモンとして有名だ。
005 MNP-950203291-9
F 新都庁が建築されつつある街角、会社帰りによる店、小物として配置される新聞、テレビ、映画などやけに“物”が具体的に描かれる。
A62-1000 として - Microsoft Internet Explorer
003 MNP-950120088-2
C これを追及した民主党下院議員が十八日、議長を侮辱したとして共和党の賛成多数でけん責処分を受けるなど、議長の疑惑はしばらく尾を引きそうだ。
004 MNP-950127124-8
C 仕事がら、私を頼り、必要としてくれる人もたくさんいる。
008 MNP-950224193-0
C 大阪ガスが地震などによるガス漏れ防止のため家庭に設置していたマイコン制御装置が、地震時にガスを使用していなければ引火の危険がないとして、揺れの前後数分間にガスを使っていた場合だけ供給を止めるよう設定さ

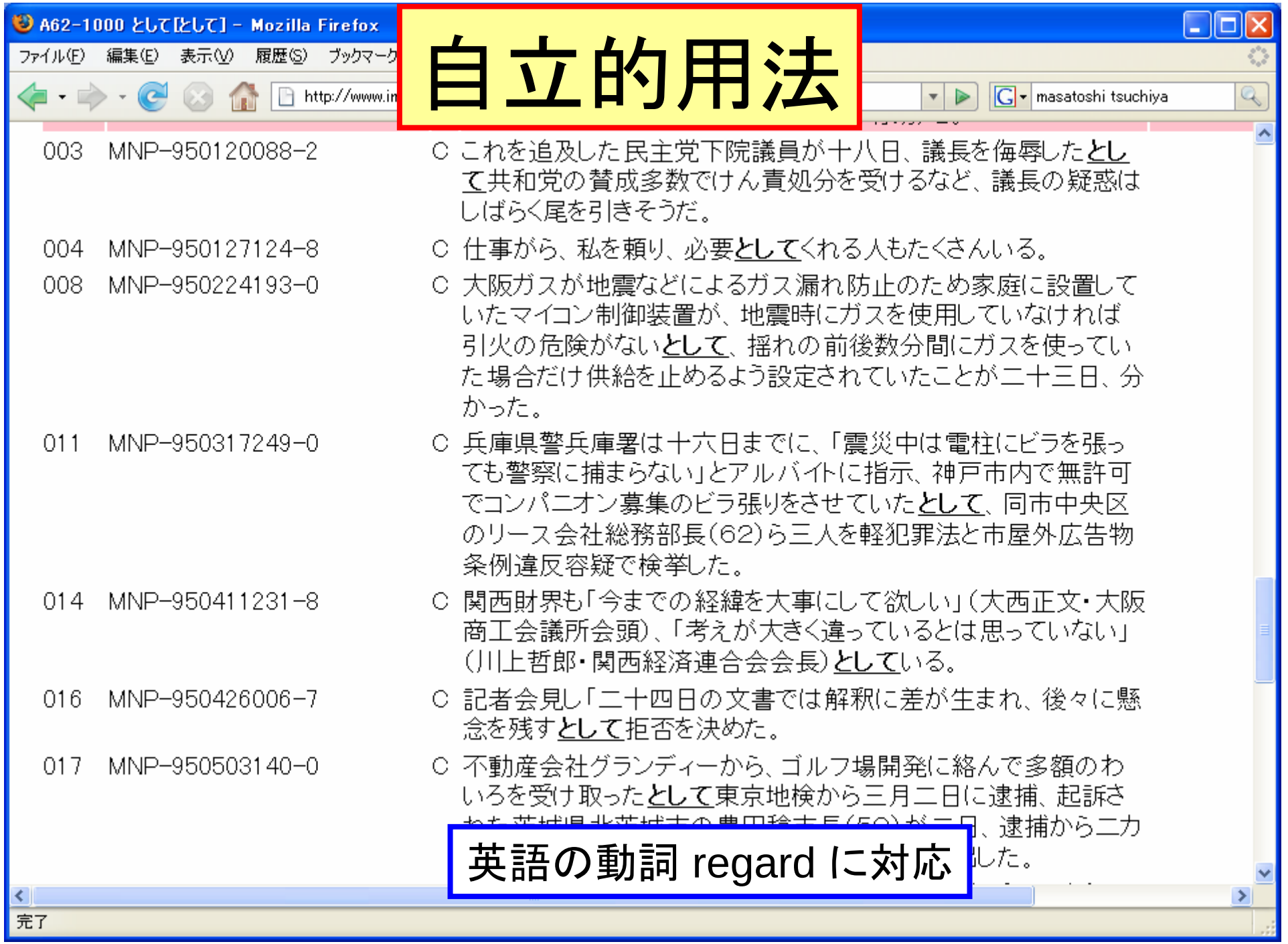

Mozilla Firefox
ファイル(F) 編集(E) 表示(V) 履歴(S) ブックマーク(B) ツール(T) ヘルプ(H)
http://www.imc.tut.ac.jp/~tsuchiya/FP/html/A53.html
masatoshi tsuchiya
国立国語研究所『現代語複合辞用例集』
Top
◇A53 ～について
MUST1
用例数 F A M C Y B 異なり 頻度
A53-1000 について 50 48 0 0 2 0 0 22491 22923
A53-1010 についての 50 50 0 0 0 0 0 1284 1312
A53-1100 につきまして 8 8 0 0 0 0 0 8 8
A53-2000 につき 50 6 0 29 8 0 7 443 454
接続
名詞(名詞節を含む)に付く。
意味・用法
言語・思考行動の対象・内容や、検討・判定・評価がなされる観点・指標を示す。
関連項目
A46「～に関して」
完了

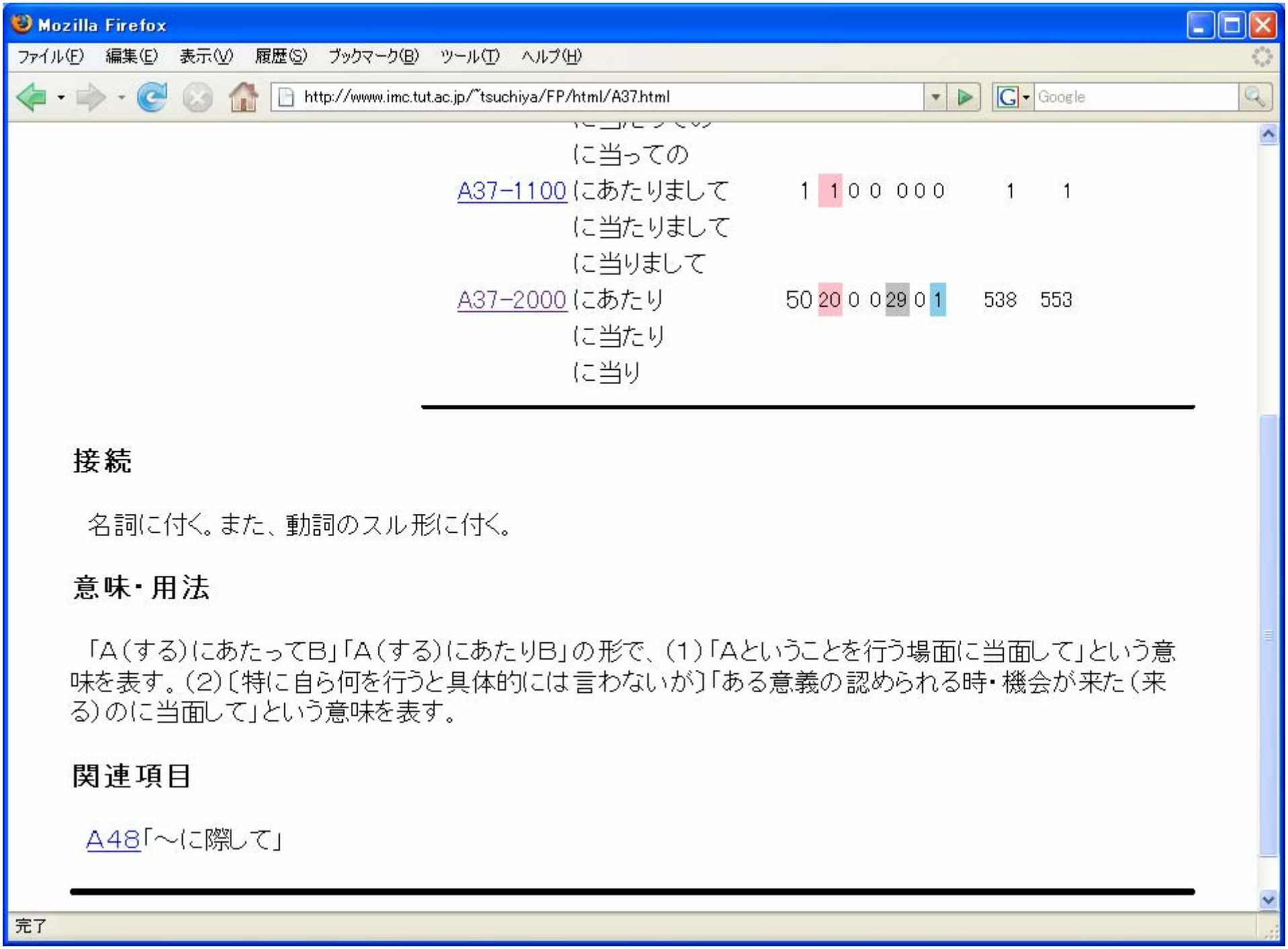

A53-1000 について[について] - Mozilla Firefox
ファイル(F) 編集(E) 表示(V) 履歴(S) ブックマーク(B) ツール(T) ヘルプ(H)
http://www.imc.tut.ac.jp/~tsuchiya/FP/html/A53-1000.html
masatoshi tsuchiya
MUST1 日本語複合辞用例データ
機能的用法
A53-1000について[について](異なり=22491 ; 頻度=22923)
F48
ID source L Text No
001 MNP-950101003-0 F 村山富市首相は年頭にあたり首相官邸で内閣記者会と二十八日会見し、社会党の新民主連合所属議員の離党問題について「政権に影響を及ぼすことにはならない。
002 MNP-950112187-2 F 西尾市長は「知事選についてコメントする立場にないが」と前置きしたうえで発言。
003 MNP-950120313-17 F 結婚後の親との同居については、自分の親と同居することに対して女性では賛成十四人、反対六人。
004 MNP-950127259-0 F 阪神大震災で新幹線、地下鉄、私鉄各線が深刻な被害を受けたが、亀井静香運輸相は二十七日の閣議後の記者会見で「鉄道施設の耐震構造については、既に建設が終わっている鉄道についても思い切った見直しをしなければならない」と述べ、国内の全鉄道路線を対象にした大規模な設計変更に着手する方針を明らかにした。
005 MNP-950204118-110 F 「リアルな体験の回想ではなく、高いレベルで戦争について作
英語の前置詞 about に対応
完了

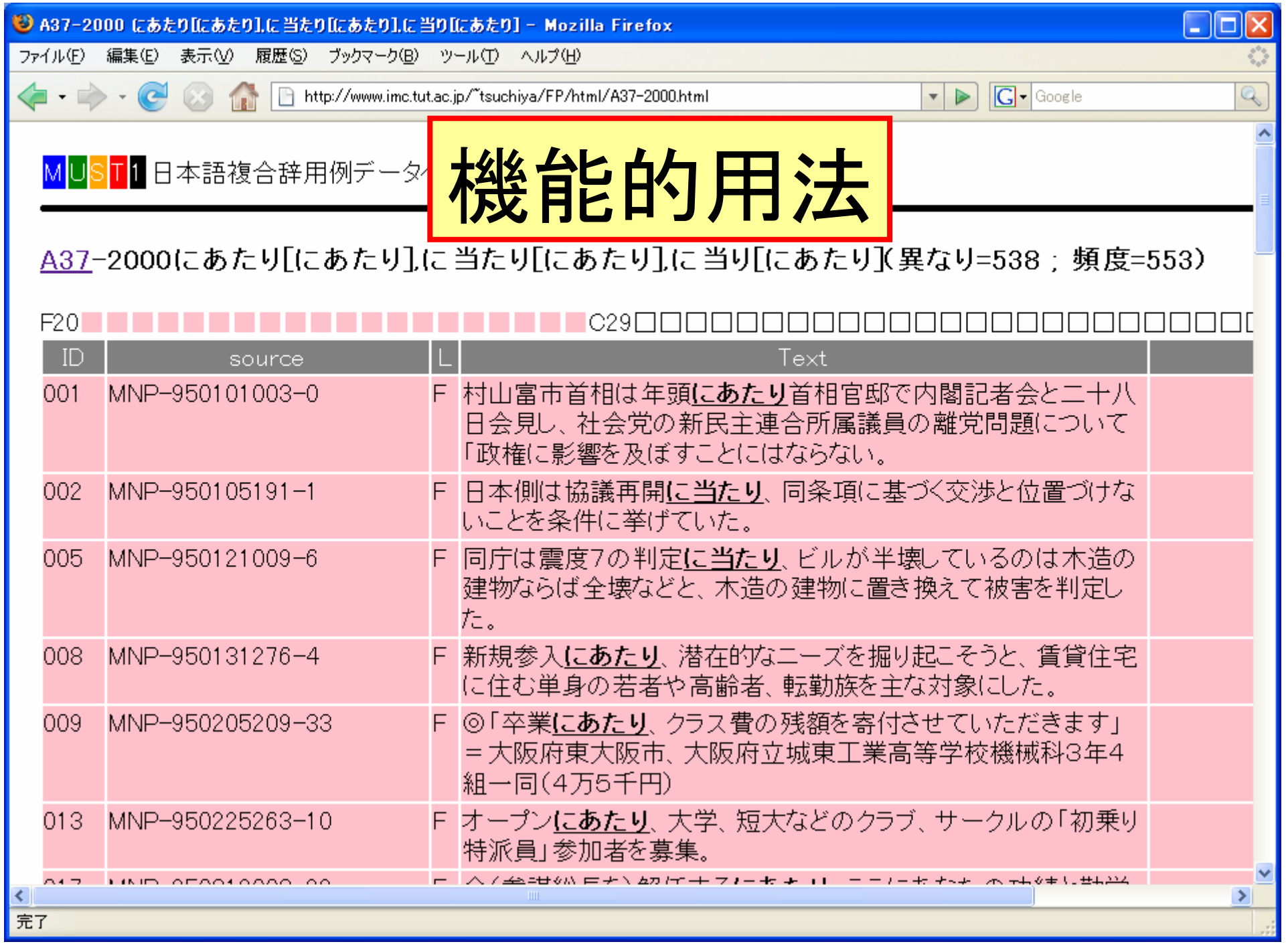

A53-1000 について[について] - Mozilla Firefox
ファイル(F) 編集(E) 表示(V) 履歴(S) ブックマーク(B) ツール(T) ヘルプ(H)
http://www.imc.tut.ac.jp/~tsuchiya/FP/html/A53-1000.html
masatoshi tsuchiya
045 MNP-951120030-13 F 中国は来年から約四千の品目について輸入関税を引き下げるなど措置を講じる。
046 MNP-951126065-11 F 現在、いじめが背景にあると思われる四件の自殺について支援。
047 MNP-951203002-12 F 粛軍クーデター事件については昨年十月までの捜査で起訴猶予処分となったものの、反乱罪には該当するとされており、現時点で改めて捜査し起訴することには問題がない。
048 MNP-951208201-15 F 今後の訴訟の行方について被害者弁護団の弁護士は「不良債権を受け皿機関の日本版RTC(整理信託公社)が引き受けるならば、ニシキ社絡みの手形債権も同様で、今後は裁判相
049 MNP-951216037-7 F え置き期間を含め
050 MNP-951222019-31 F 様をつくることになるだけだ。
自立的用法
035 MNP-950917111-4 C 「引かれた瞬間、前に落ちると思った」と言ったが、怖がらずについていき、倒れ込みながら若乃花を土俵外へはじき飛ばした。
044 MNP-951114149-11 C いわば、相手の変化についていけない曙の典型的な敗戦パターンだった。
(c) Group MUST, 2005. Generated by subentry2
英語の動詞 follow に対応
完了

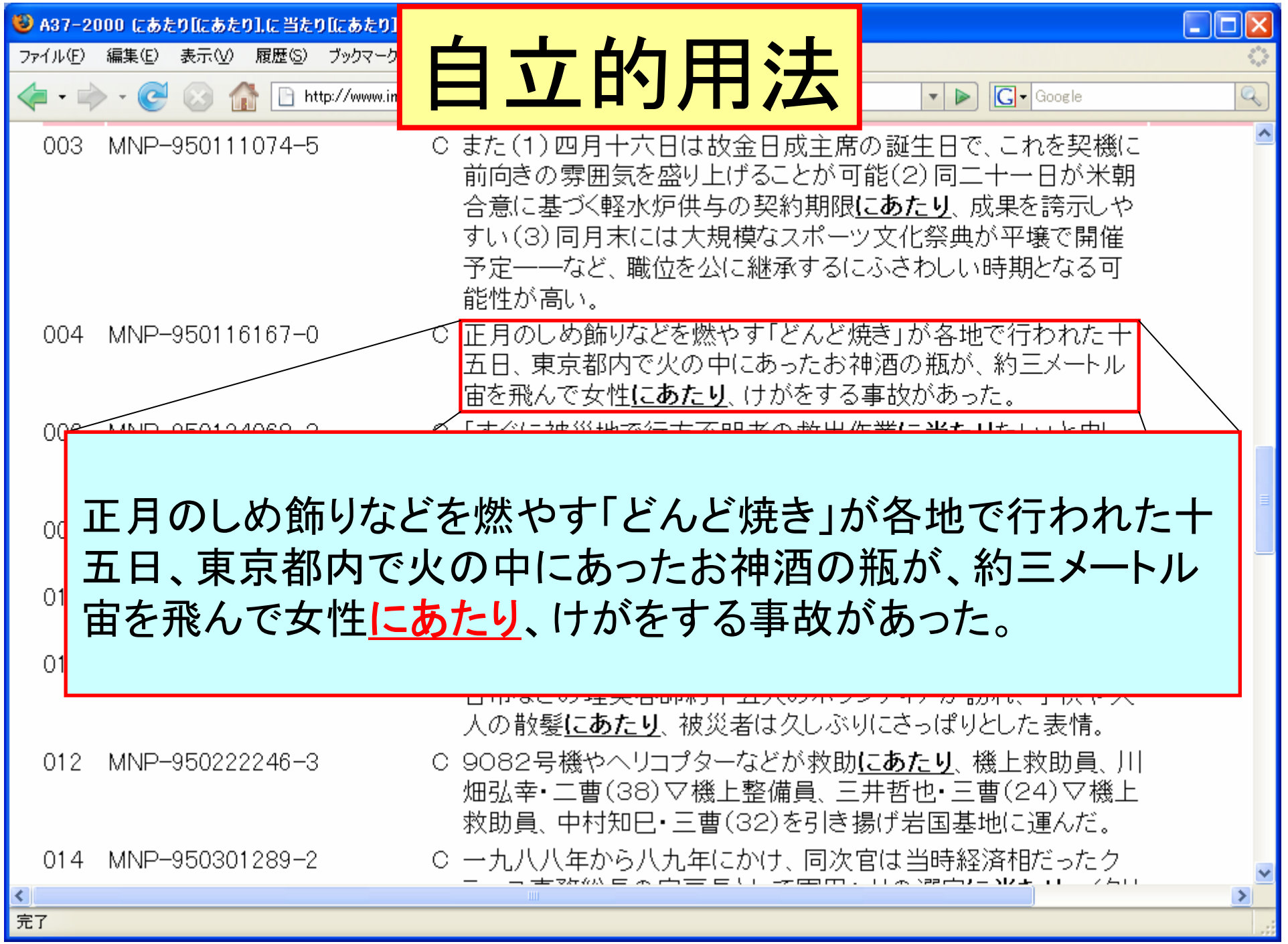

Mozilla Firefox
ファイル(F) 編集(E) 表示(V) 履歴(S) ブックマーク(B) ツール(T) ヘルプ(H)
http://www.imc.tut.ac.jp/~tsuchiya/FP/html/A62.html
masatoshi tsuchiya
国立国語研究所『現代語複合辞用例集』
Top
◇A62 ～として
MUST1
A62-1000 として
A62-1010 としての
A62-1100 としまして
A62-1101 といたしまして
と致しまして
用例数 F A M C Y B 異なり 頻度
50 30 0 0 15 0 5 39309 40263
50 50 0 0 0 0 0 2631 2672
2 0 0 0 1 0 1 2 2
1 0 0 0 1 0 0 1 1
接続
名詞(名詞節を含む)に付く。
意味・用法
問題にする人・物事などの位置づけを示す。どのような位置づけかで、資格・立場・部類・行為の意義づけなどを表わすと下位区分される。
完了


A62-1000 として[として] - Mozilla Firefox
ファイル(F) 編集(E) 表示(V) 履歴(S) ブックマーク(B) ツール(T) ヘルプ(H)
http://www.imc.tut.ac.jp/~tsuchiya/FP/html/A62-1000.html
masatoshi tsuchiya
MUST1 日本語複合辞用例データ
機能的用法
A62-1000として[として]（異なり=39309 ； 頻度=40263）
F30
C15
ID
source
L
Text
001 MNP-950101005-22 F 《長野冬季オリンピック招致委員会は、八九年十月、任意団体として発足。
002 MNP-950111228-9 F カイコガの雌が雄を誘引する時に出す物質ボンビコールは、とくに雌雄間に働く性フェロモンとして有名だ。
005 MNP-950203291-9 F 新都庁が建築されつつある街角、会社帰りによる店、小物として配置される新聞、テレビ、映画などやけに“物”が具体的に描かれる。
006 MNP-950210288-4 F 投・開票所の増加や給与の上昇などを考慮して三年前より総額で二六・四%アップし、五百六十億円が選挙費用として九五年度予算に計上されている。
007 MNP-950218042-20 F たくさんの同世代の若者たちが、ボランティアとして頑張っている。
009 MNP-950303226-1 F けれど私はその記憶を、自分の傷として持ち続けたい。
010 MNP-950310240-3 F
が緊急課題として
英語の前置詞 as に対応
完了


A62-1000 として[として] - Mozilla Firefox
ファイル(F) 編集(E) 表示(V) 履歴(S) ブックマーク
http://www.in
masatoshi tsuchiya
自立的用法
003 MNP-950120088-2 C これを追及した民主党下院議員が十八日、議長を侮辱したとして共和党の賛成多数でけん責処分を受けるなど、議長の疑惑はしばらく尾を引きそうだ。
004 MNP-950127124-8 C 仕事がら、私を頼り、必要としてくれる人もたくさんいる。
008 MNP-950224193-0 C 大阪ガスが地震などによるガス漏れ防止のため家庭に設置していたマイコン制御装置が、地震時にガスを使用していなければ引火の危険がないとして、揺れの前後数分間にガスを使っていた場合だけ供給を止めるよう設定されていたことが二十三日、分かった。
011 MNP-950317249-0 C 兵庫県警兵庫署は十六日までに、「震災中は電柱にビラを張っても警察に捕まらない」とアルバイトに指示、神戸市内で無許可でコンパニオン募集のビラ張りをさせていたとして、同市中央区のリース会社総務部長(62)ら三人を軽犯罪法と市屋外広告物条例違反容疑で検挙した。
014 MNP-950411231-8 C 関西財界も「今までの経緯を大事にして欲しい」(大西正文・大阪商工会議所会頭)、「考えが大きく違っているとは思っていない」(川上哲郎・関西経済連合会会長)としている。
016 MNP-950426006-7 C 記者会見し「二十四日の文書では解釈に差が生まれ、後々に懸念を残すとして拒否を決めた。
017 MNP-950503140-0 C 不動産会社グランディーから、ゴルフ場開発に絡んで多額のわいろを受け取ったとして東京地検から三月二日に逮捕、起訴さ
英語の動詞 regard に対応
完了


Mozilla Firefox
ファイル(F) 編集(E) 表示(V) 履歴(S) ブックマーク(B) ツール(T) ヘルプ(H)
http://www.imc.tut.ac.jp/~tsuchiya/FP/html/A37.html
Google
に当っての
A37-1100 にあたりまして 1 1 0 0 0 0 0 1 1
に当たりまして
に当りまして
A37-2000 にあたり 50 20 0 0 29 0 1 538 553
に当たり
に当り
接続
名詞に付く。また、動詞のスル形に付く。
意味・用法
「A(する)にあたってB」「A(する)にあたりB」の形で、(1)「Aということを行う場面に当面して」という意味を表す。(2)〔特に自ら何を行うと具体的には言わないが〕「ある意義の認められる時・機会が来た(来る)のに当面して」という意味を表す。
関連項目
A48「～に際して」
完了


A37-2000 にあたり[にあたり],に 当たり[にあたり],に 当り[にあたり] - Mozilla Firefox
ファイル(F) 編集(E) 表示(V) 履歴(S) ブックマーク(B) ツール(T) ヘルプ(H)
http://www.imc.tut.ac.jp/~tsuchiya/FP/html/A37-2000.html
Google
MUST1 日本語複合辞用例データ
機能的用法
A37-2000にあたり[にあたり],に 当たり[にあたり],に 当り[にあたり](異なり=538 ; 頻度=553)
F20
C29
ID source L Text
001 MNP-950101003-0 F 村山富市首相は年頭にあたり首相官邸で内閣記者会と二十八日会見し、社会党の新民主連合所属議員の離党問題について「政権に影響を及ぼすことにはならない。
002 MNP-950105191-1 F 日本側は協議再開に当たり、同条項に基づく交渉と位置づけないことを条件に挙げていた。
005 MNP-950121009-6 F 同庁は震度7の判定に当たり、ビルが半壊しているのは木造の建物ならば全壊などと、木造の建物に置き換えて被害を判定した。
008 MNP-950131276-4 F 新規参入にあたり、潜在的なニーズを掘り起こそうと、賃貸住宅に住む単身の若者や高齢者、転勤族を主な対象にした。
009 MNP-950205209-33 F ◎「卒業にあたり、クラス費の残額を寄付させていただきます」＝大阪府東大阪市、大阪府立城東工業高等学校機械科3年4組一同(4万5千円)
013 MNP-950225263-10 F オープンにあたり、大学、短大などのクラブ、サークルの「初乗り特派員」参加者を募集。
完了

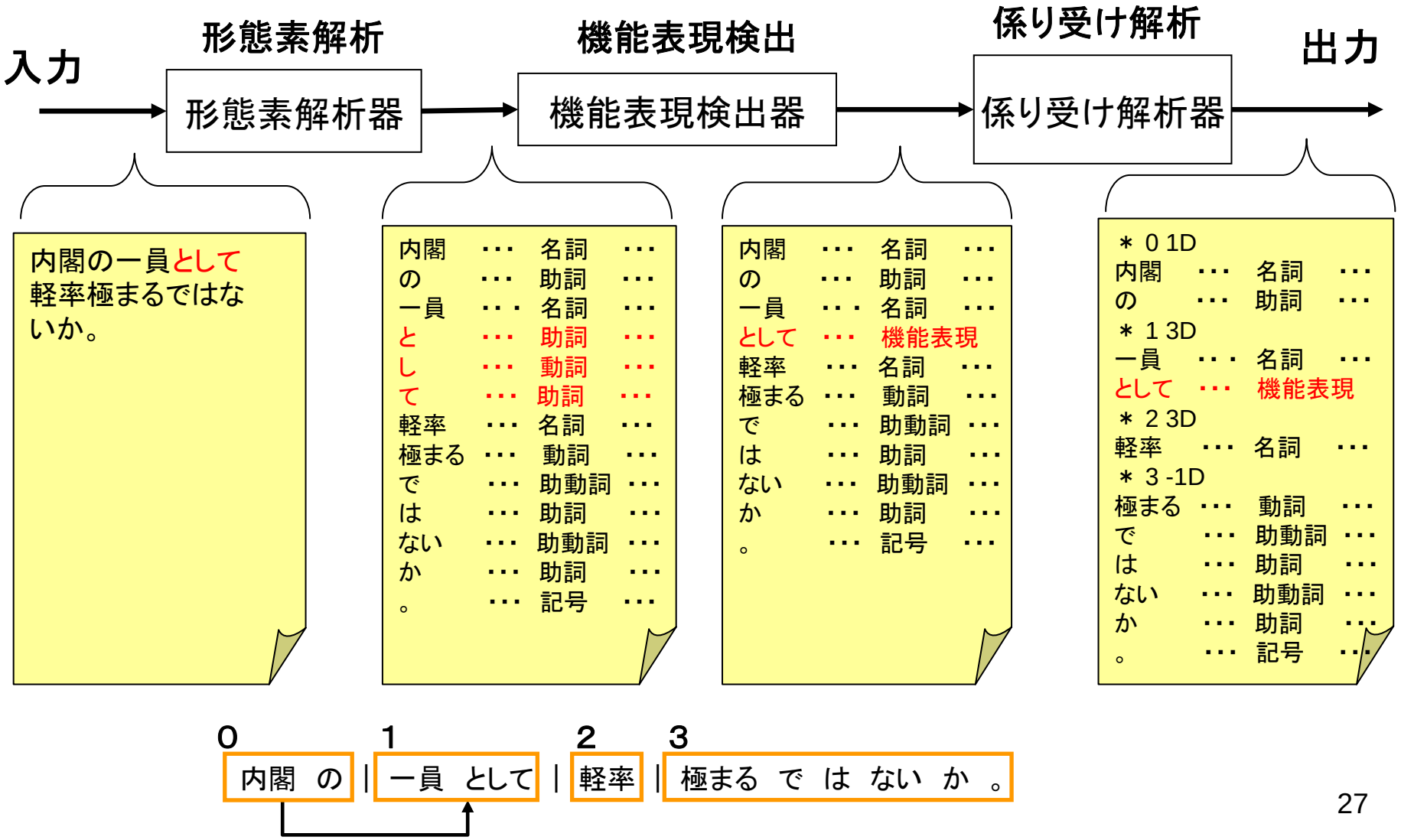

A37-2000 にあたり[にあたり],に 当たり[にあたり]
ファイル(F) 編集(E) 表示(V) 履歴(S) ブックマーク
http://www.in
Google
自立的用法
003 MNP-950111074-5 C また(1)四月十六日は故金日成主席の誕生日で、これを契機に前向きの雰囲気を盛り上げることが可能(2)同二十一日が米朝合意に基づく軽水炉供与の契約期限にあたり、成果を誇示しやすい(3)同月末には大規模なスポーツ文化祭典が平壌で開催予定――など、職位を公に継承するにふさわしい時期となる可能性が高い。
004 MNP-950116167-0 C 正月のしめ飾りなどを燃やす「どんど焼き」が各地で行われた十五日、東京都内で火の中にあったお神酒の瓶が、約三メートル宙を飛んで女性にあたり、けがをする事故があった。
012 MNP-950222246-3 C 9082号機やヘリコプターなどが救助にあたり、機上救助員、川畑弘幸・二曹(38)▽機上整備員、三井哲也・三曹(24)▽機上救助員、中村知巳・三曹(32)を引き揚げ岩国基地に運んだ。
014 MNP-950301289-2 C 一九八八年から八九年にかけ、同次官は当時経済相だったク
正月のしめ飾りなどを燃やす「どんど焼き」が各地で行われた十五日、東京都内で火の中にあったお神酒の瓶が、約三メートル宙を飛んで女性にあたり、けがをする事故があった。
人の散髪にあたり、被災者は久しぶりにさっぱりとした表情。
完了

機能表現を考慮した
係り受け解析

本日の内訳

1. プロジェクト全体の紹介（宇津呂）
2. 日本語複合辞の階層的辞書の編纂（松吉）
   （おまけ）複合辞の言い換え
3. 日本語複合辞用例データベースの作成（土屋）
4. 日本語文中の複合辞の言語解析
   4-1: 複合辞の検出（土屋）
   4-2: 統計的係り受け解析（宇津呂）
5. その他の周辺の話題・進行中の研究（宇津呂）

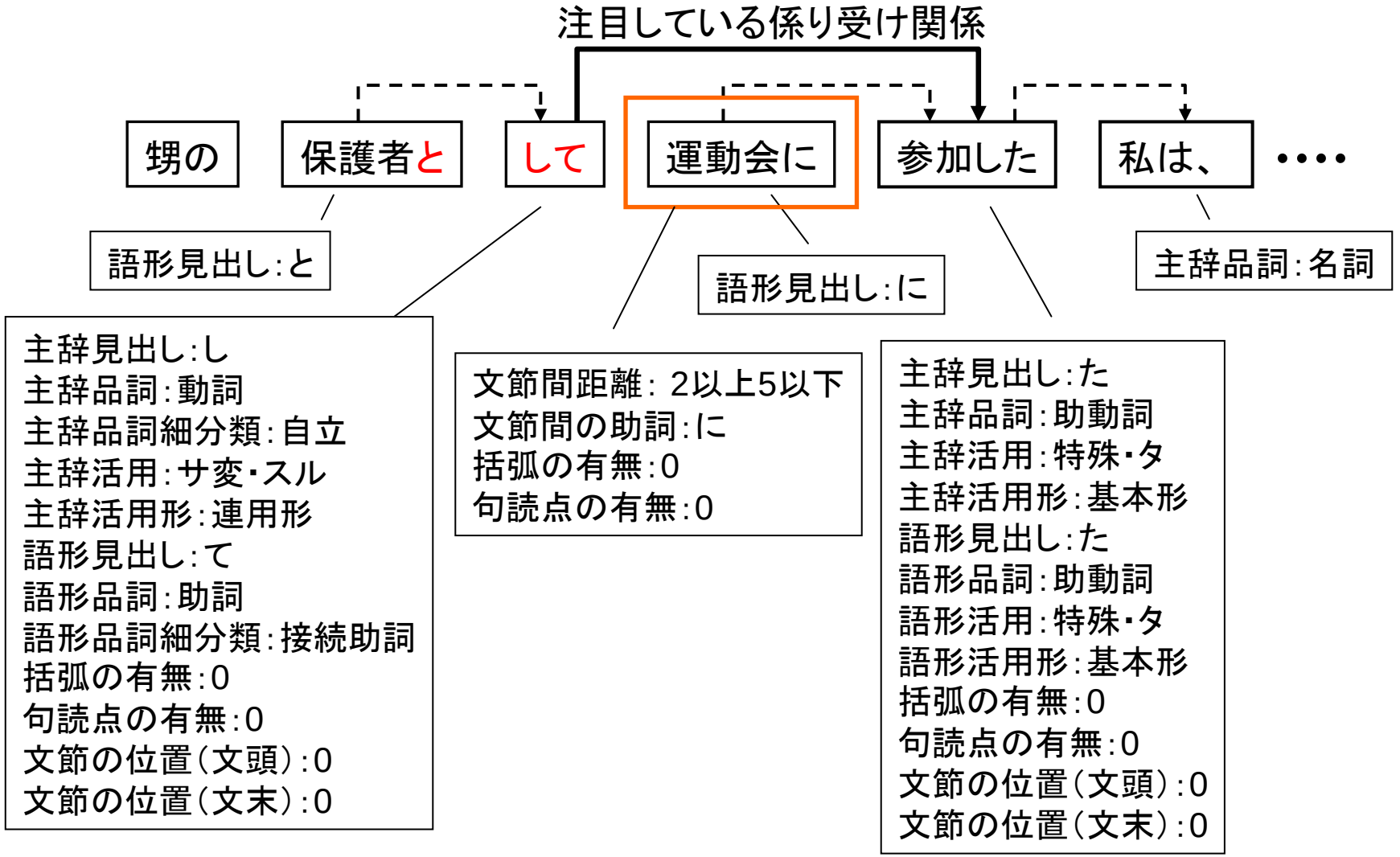
機能表現を考慮した係り受け解析システムの一構成案



機能表現を考慮した係り受け解析器の実現

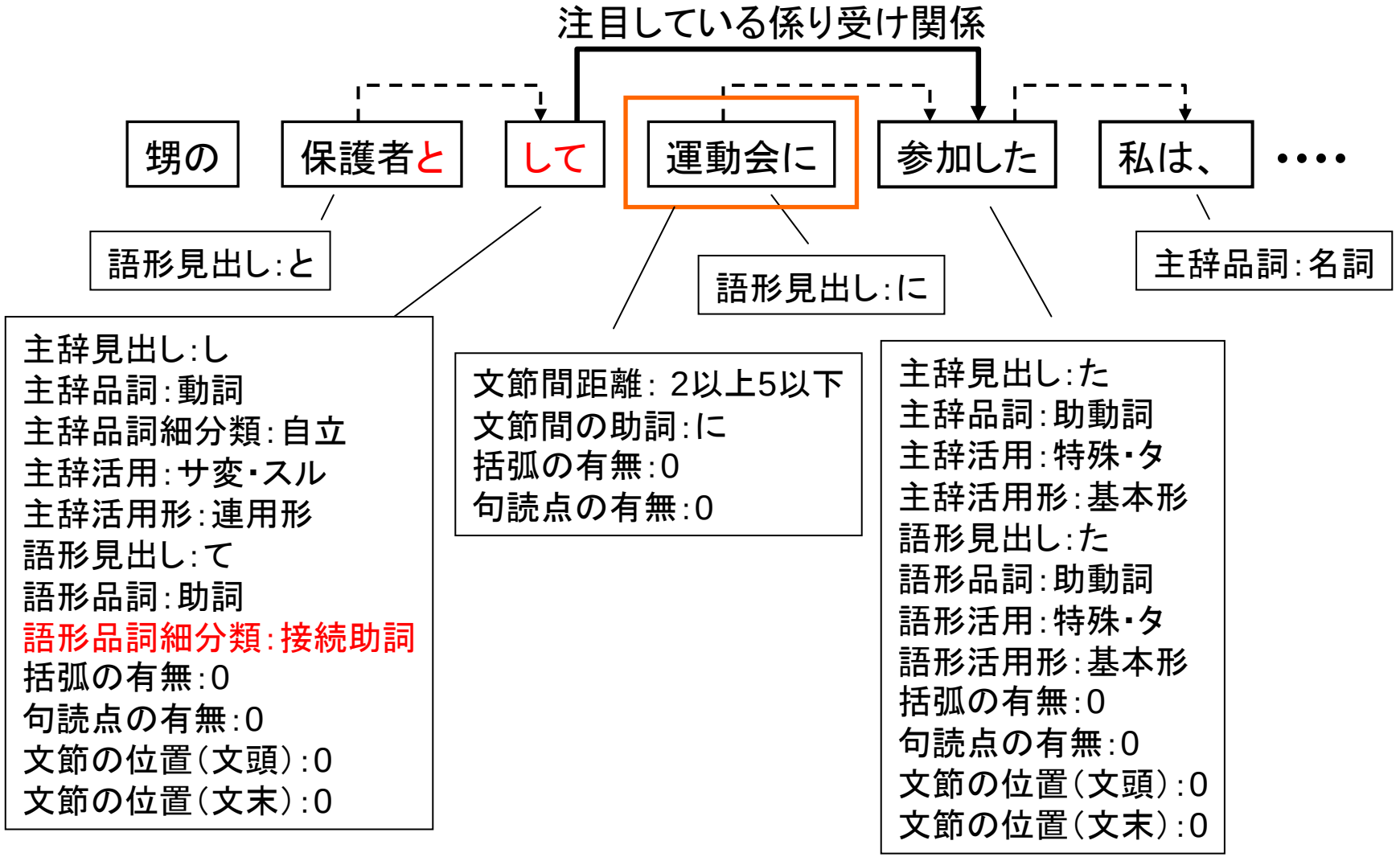
- SVMを用いた統計的係り受け解析手法を利用「工藤2002」
  - ツールとしてはCaboChaを使用
- 京都コーパス38400文に対して、機能表現の情報を付与
- 上のデータを基に機能表現を考慮した係り受け解析用の訓練データを作成
- 上の訓練データで係り受け解析器の学習を行うことによって、機能表現を考慮した係り受け解析器を実現

学習・解析に使用する素性の変化



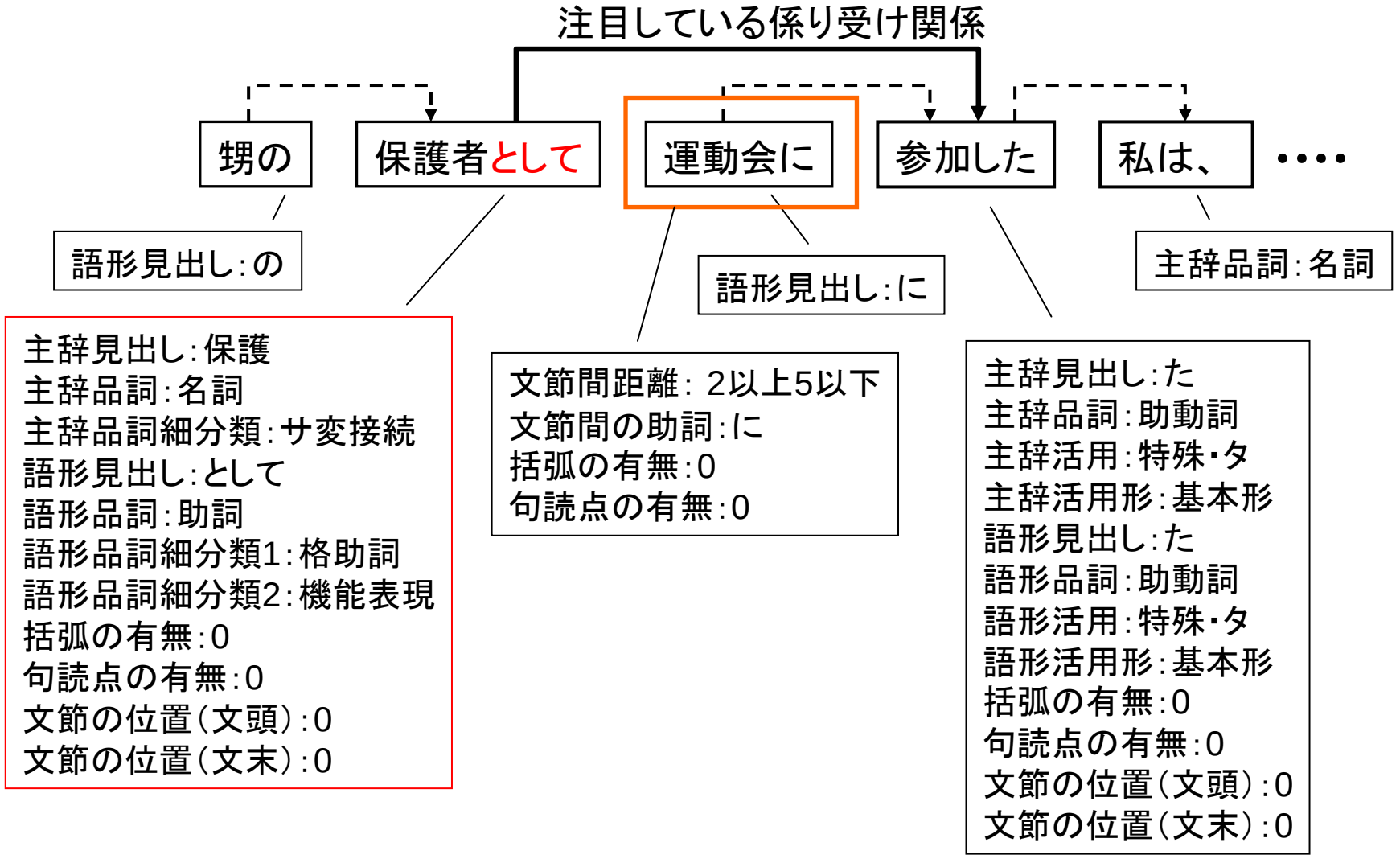
主辞・・・品詞が特殊、助詞、接尾辞となるものを除き一番文節末に近い形態素
語形・・・品詞が特殊となるものを除き一番文節末に近い形態素

学習・解析に使用する素性の変化



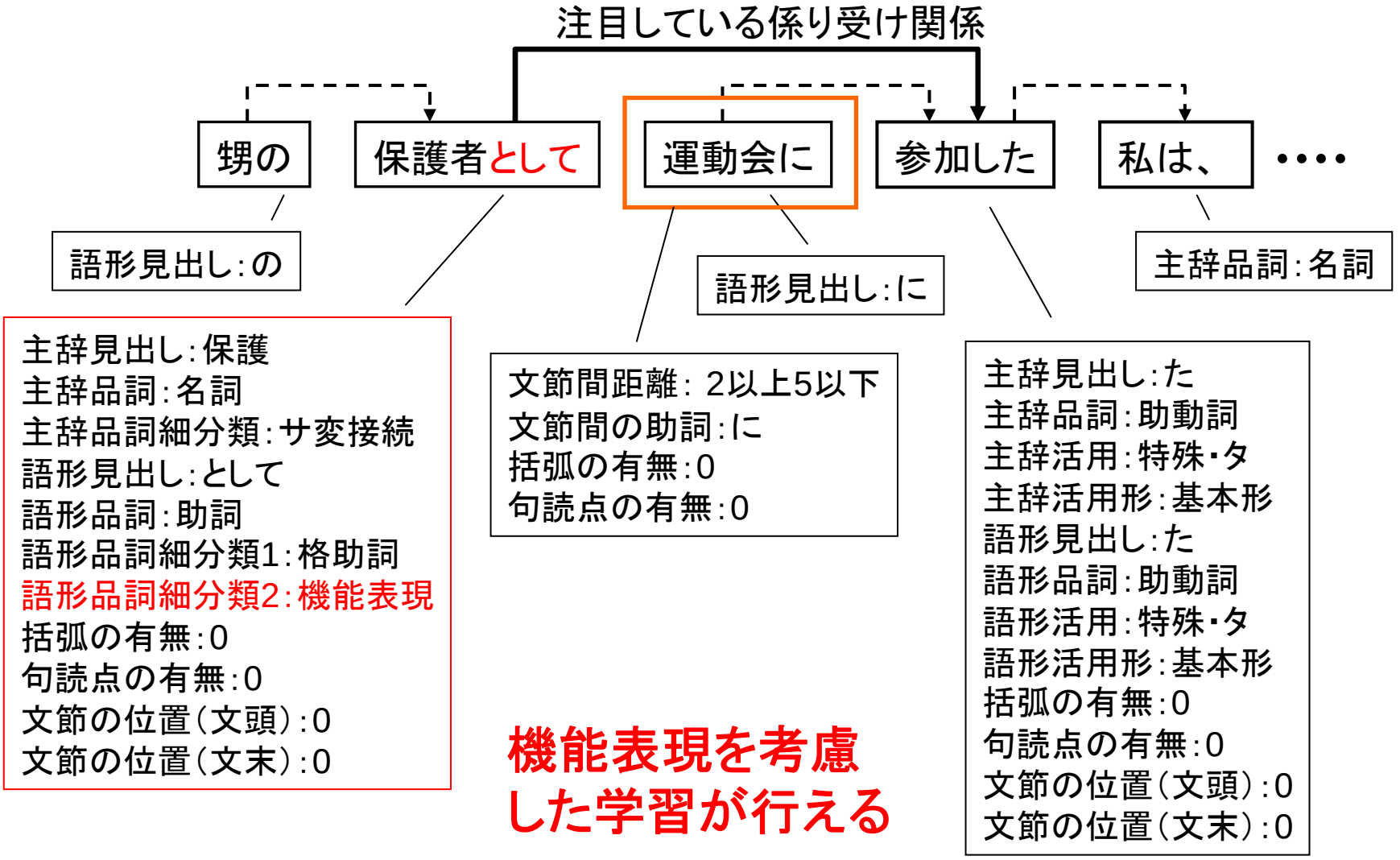
主辞・・・品詞が特殊、助詞、接尾辞となるものを除き一番文節末に近い形態素
語形・・・品詞が特殊となるものを除き一番文節末に近い形態素

学習・解析に使用する素性の変化



主辞・・・品詞が特殊、助詞、接尾辞となるものを除き一番文節末に近い形態素
語形・・・品詞が特殊となるものを除き一番文節末に近い形態素

学習・解析に使用する素性の変化



主辞・・・品詞が特殊、助詞、接尾辞となるものを除き一番文節末に近い形態素
語形・・・品詞が特殊となるものを除き一番文節末に近い形態素

係り受け解析の評価実験

- 機能表現を考慮した係り受け解析と以下のベースラインの解析性能の比較評価を行った
- 京都テキストコーパスを訓練・評価データとして10分割交差検定を行う
- 使用した機能表現検出器
  - 先の実験における検出器を使用
  - 形態素素性、チャンク素性、チャンク文脈素性で学習

ベースライン

- **CaboCha（機能表現抜き）**
  - 対象の表現が形態素解析辞書に登録されている場合、その表現を構成している形態素単位に分割して学習をし直している。
- **CaboCha（オリジナル）**
  - 従来のCaboCha

評価尺度（係り受け解析）

$$\text{係り先精度} = \frac{\text{係り先を正しく同定できたFE文節数}}{\text{FE文節数}}$$

$$\text{係り元精度} = \frac{\text{係り元を正しく同定できたFE文節数}}{\text{FE文節数}}$$

FE文節 ＝ 機能表現を含む文節

評価結果

各モデルでの同定精度（%）

		係り先精度	係り元精度
ベースライン	CaboCha（機能表現抜き）	88.0	72.5
	CaboCha（オリジナル）	87.6	73.9
提案手法	検出器使用*1	88.0	74.0
	正解使用* 2	88.1	74.4

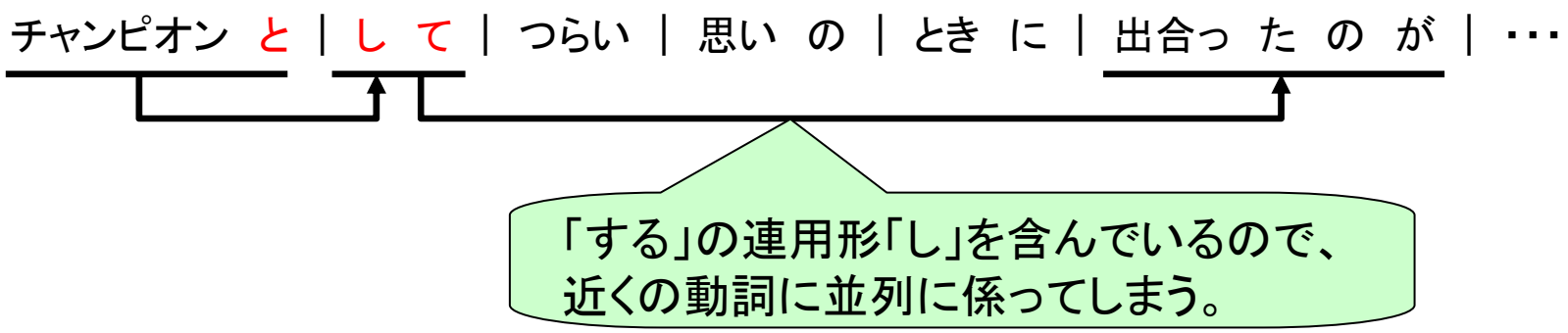
機能表現を考慮すること、正しく機能表現を検出することは係り元の推定に効果的

＊1・・・機能表現検出に機能表現検出器を使用
＊2・・・機能表現検出に正解データを使用

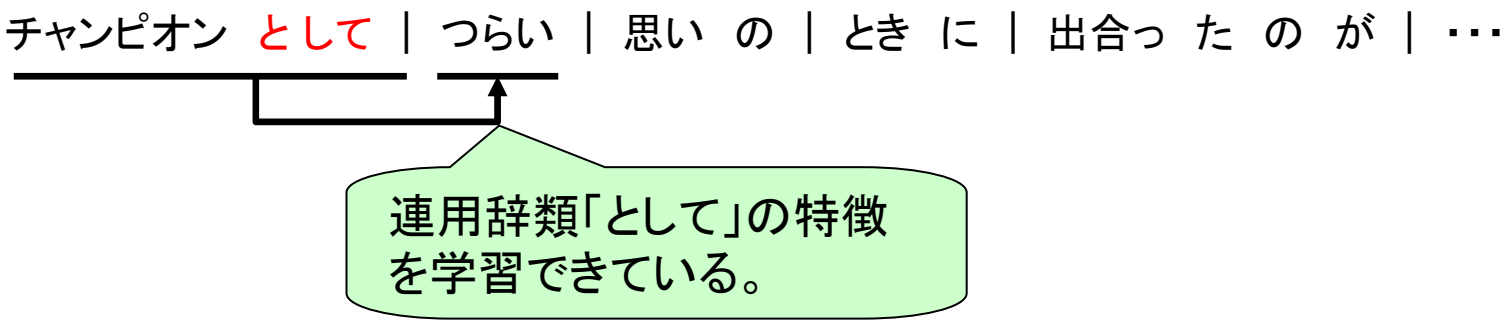
係り先同定の改善例（連用辞類）

A62-1000（として）：助詞型-連用辞類

・機能表現を考慮しない係り受け解析



・機能表現を考慮した係り受け解析

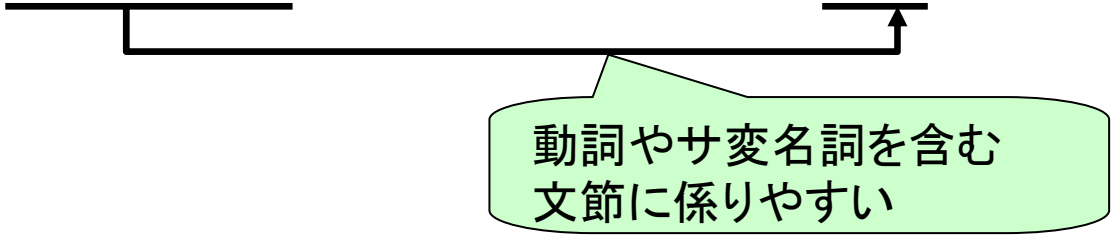


係り元同定の改善例（連用辞類）

A62-1000（として）：助詞型-連用辞類

・機能表現を考慮しない係り受け解析

・・・ | ロシア　軍　の | チェチェン　進行　を | 東欧　諸国　の | 首脳　と | し　て | ・・・ | 批判　。



・機能表現を考慮した係り受け解析

・・・ | ロシア　軍　の | チェチェン　進行　を | 東欧　諸国　の | 首脳　として | ・・・ | 批判　。

まとめと今後の課題

- まとめ
  - 従来手法より高性能な機能表現検出器の構築
  - 機能表現検出器を使い、機能表現を考慮することにより、従来手法より高性能な係り受け解析を実現

- 今後の課題
  - 対象とする表現の拡張
  - 格解析との統合
  - 応用への展開

周辺話題

17,000全機能表現の検出・・・
派生的表現を代表的表現で代用する方式の確立

- 対象とする表現の規模を数千のオーダへ:
  助詞型・接続詞型他3000表現
  ⇒従来どおり用例を作成して検出器を学習する
  　代表表現を立てる.
  　　現在，L^3 (意味・派生が異なる551分類)について50用例作成済
  助動詞型:14000表現⇒下記による

- 機能的用法・自立的用法の区別が必要な
  表現数の推定:全17000中，約5000，新聞記事中では2600～3000

- 代表的表現についての訓練コーパスのみを用いて、
  派生的表現の検出を実現
  – 機能表現階層分類 [松吉 2007] のL^3 (意味・派生が異なる551分類)の単位ごとに一括処理

文型検索ツールの開発
（川村先生との共同研究）

辞書ツール：　単語と辞書情報をリンク

↓

文型検索ツール

・文中の機能表現を自動的に検出するシステム

（土屋・宇津呂ほか2005）

機能表現＝いくつかの語が複合してひとまとまりの句となって付属語的な役割を果たしている語

＝文型

文型検索ツールの開発
（川村先生との共同研究）

文型検索ツール

・入力された文章を形態素解析

・形態素列パターンに基づくパターンマッチング

・本文中の文型を抽出

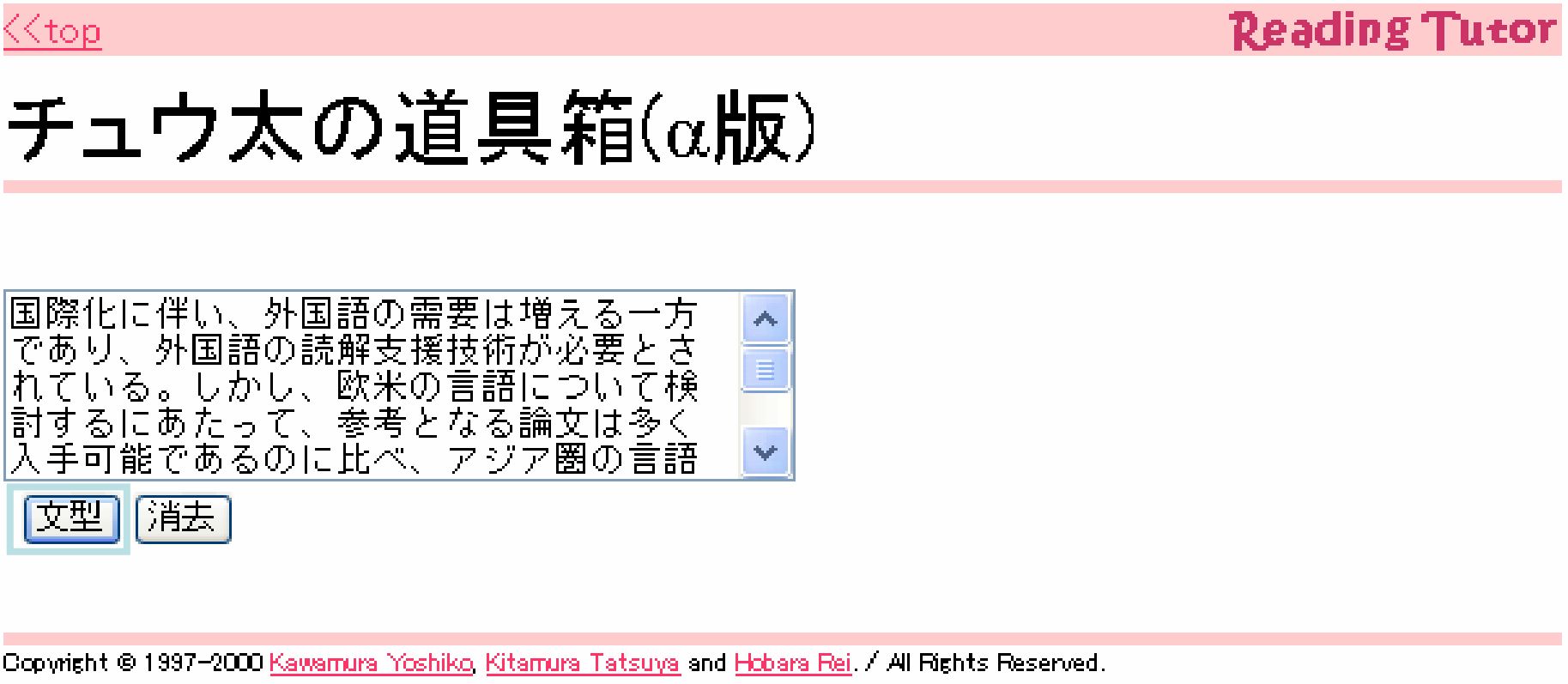
↓

・辞書情報とリンクして表示

グループ・ジャマシイ『日本語文型辞典』

文型検索ツールの入力画面

文型検索ツールの入力画面



文型検索ツールの結果画面

Reading Tutor

リーディング チュウ太

日本語 / English / Deutch

入力された文章

分からない文型をクリックしてください。意味が右に表示されます。

国際化にともない、外国語の需要は増える一方であり、外国語の読解支援技術が必要とされている。しかし、欧米の言語について検討するにあたって、参考となる論文は多く入手可能であるのに対して、日本語に関する論文を入手することは容易ではない。

のに対して [のにたいして]

対比的なふたつのことがらを並べて示すのに用いる。

にあたって [に当って / に当たって]

名詞や動詞の辞書形を受けて、「ものごとの節目となるような重要な時機にさしかかって」という意味を表す。...にさいして。式辞や礼状などの形式張った表現として用いることが多い。さらに形式張ったものとして「にあたり(まして)」を用いることもある。名詞を修飾する場合は(4)(5)のように「...にあたってのN」という形になる。

に関する

A46-1011

について

文型検索ツール：現状と計画

- 現状
  - 「日本語文型辞典」の表記と
    MUST（337表記）のAND（166表記）に対して
    「日本語文型辞典」のエントリを提示

- 計画
  - 階層的分類[松吉2007]の表記と
    「日本語文型辞典」のエントリを対応付け
  - 評価用用例集作成：川村先生・科研費