

TR-IT-0006

自動翻訳電話研究所の対話データベースと
音声翻訳システム ASURA の音声認識用文法・
構文解析用文法間の日本語形態素情報の相違点
Differences in Japanese morphological information among
ATR Dialogue Database, the speech recognition grammar and
the parsing grammar in ATR speech translation system: ASURA

竹沢 寿幸 田代 敏久 保坂 順子 衛藤 純司†
Toshiyuki TAKEZAWA Toshihisa TASHIRO Junko HOSAKA Junji ETOH†

1993. 7. 29

内容梗概

ATR 自動翻訳電話研究所で開発された対話データベースと、音声翻訳システム ASURA で使われた音声認識用文法・構文解析用文法は、それぞれ別の目的のために独立に開発されてきたので、日本語形態素においてさまざまな差異(大きく分けて、品詞の設定の差異、語の分割の差異、表記の揺れの3種類)が存在している。新しい音声言語データベースにおける日本語形態素情報の体系を検討するために、それらの間の日本語形態素情報の差異を調査したので、その内容を述べる。

ATR 音声翻訳通信研究所
ATR Interpreting Telecommunications Research Laboratories
†日本アイアール株式会社

© 株式会社 エイ・ティ・アール 音声翻訳通信研究所
© 1993 by ATR Interpreting Telecommunications Research Laboratories

目次

1	まえがき	1
2	品詞・語彙区切りの差異	1
2.1	名詞	1
2.1.1	固有名詞	1
2.1.2	サ変名詞	3
2.1.3	形容名詞	4
2.1.4	普通名詞	4
2.1.5	数詞	5
2.1.6	代名詞	6
2.2	本動詞	7
2.3	補助動詞・助動詞	8
2.3.1	補助動詞	8
2.3.2	助動詞	10
2.3.3	長単位分割の補助動詞・助動詞	11
2.4	形容詞	13
2.5	副詞	13
2.6	連体詞	16
2.7	接続詞	16
2.8	感動詞・間投詞	16
2.9	助詞	17
2.10	接辞	19
2.10.1	接頭辞	19
2.10.2	接尾辞	19
3	表記の揺れ	19
3.1	同一単語の表記の揺れ	20
3.2	送り仮名	21
3.3	数詞	22
3.4	外来語・外国語固有名詞	22
4	むすび	23
	参考文献	24

1 まえがき

ATR 自動翻訳電話研究所で開発された対話データベースと、音声翻訳システム ASURA で使われた音声認識用文法・構文解析用文法はそれぞれ独立に開発されてきたので、形態素解析においてさまざまな差異がある。その差異のタイプは大きく分けて3つある。

1. 品詞の設定の差異
2. 語の分割の差異
3. 表記の揺れ

ATR 対話データベースの品詞は、大体において日本語学校文法に準拠している。構文解析用文法の品詞はそれよりも粒度が細かく、特に人名・地名などの固有名詞、副詞、助詞、助動詞などは、統語的・意味的に細分化されている。音声認識用文法の品詞は最も粒度が細かく、ほとんどすべての品詞にわたって詳細な接続制約が記述できるようになっている。

語の分割の差異は、品詞の設定の差異によって自ずから生じるものと、音声認識や構文解析の特殊性を考慮して例外的な処理をしているものがある。特に後者では、感動詞、助動詞・補助動詞、副詞などに、意味的なまとまりを考えて長単位登録しているものが多い。

表記は、いずれも原則として「新明解国語辞典」や「朝日新聞の用語の手引」に従っているが、音声認識と構文解析のあいだのインターフェイスの都合によって差異が生じているものがある。

以下では、「日本語形態素解析マニュアル」に沿いながら品詞の設定や語の分割の差異を説明し、最後に表記の揺れについて説明する。

2 品詞・語彙区切りの差異

2.1 名詞

2.1.1 固有名詞

ATR 対話データベースでは、固有名詞はすべて同じ品詞名 (固名詞) を与えられている。音声認識用文法・構文解析用文法では、固有名詞一般を指す品詞名はそれぞれ n-proper と n-prop であるが、人名・地名などでは細分化している。

人名は、ATR 対話データベースでは、姓・名・姓名のそれぞれを一語登録している。構文解析用文法では姓と名で別の品詞名を与えている。音声認識用文法では、それに加えて日本人の名前と外国人の名前を区別している。

人名に「様」や「先生」などの敬称がついたものは、ATR 対話データベース・音声認識用文法・構文解析用文法ともに人名と敬称に分割している。ただし、音声認識用文法と構文解析用文法では「長尾教授」だけは全体で一語登録している。

データベース		音声認識		構文解析	
固名詞	鈴木明 鈴木 明	n-nachname	鈴木	n-family	鈴木
		n-nachname-others	ブラウン		ブラウン
		n-vorname	明	n-first	明
		n-vorname-others	アダム		アダム
接尾語	様 教授	name-suf	様	n-suffix-noun	様
		n-nachname-pol	長尾教授	n-com	長尾教授

地名は、ATR 対話データベースでは、東京都・港区・元麻布などを一語として固名詞という品詞名を与えている。構文解析用文法と音声認識用文法では、東京／都・港／区・元麻布というように分割している。

丁目・番地・号は、ATR 対話データベース・音声認識用文法・構文解析用文法ともに、数詞と接尾語に分割している。

音声認識用文法では、地名であっても住所表示に使われない場合は、一般の固有名詞と同じ扱いをする。

データベース		音声認識		構文解析	
固名詞	東京都	ken-name-to	東京	n-prop	東京
		ken-name-suf-to	都	n-suffix-pref	都
	大阪府	ken-name-hu	大阪	n-prop	大阪
		ken-name-suf-hu	府	n-suffix-pref	府
	港区	si-name-ku	港	n-prop	港
		si-name-suf-ku	区	n-suffix-ward	区
	大阪市	si-name-si	大阪	n-prop	大阪
		si-name-suf-si	市	n-suffix-city	市
	東区	ku-name-ku	東	n-prop	東
		ku-name-suf-ku	区	n-suffix-ward	区
	元麻布	adre-tyo	元麻布	adre-tyo	元麻布
数詞	一	num-1	一	n-num	一
接尾語	丁目 番 号	num-suf-ch	丁目	n-suffix-chome	丁目
		num-suf-b	番	n-suffix-banchi	番
		num-suf-g	号	n-suffix-gou	号
固名詞	日本 京都	n-proper	日本 京都	n-prop	日本 京都

ATR 対話データベース・音声認識用文法・構文解析用文法ともに、固有名詞は長単位分割を基本としている。ただし、会社名・部課名・役職名が連なっているものは、ATR 対話データベースでは「…名」の単位で切るが、音声認識用文法と構文解析用文法では、名詞連続複合語の生成規則がないので、全体を一語登録することになる。

ATR 対話データベース: ATR 音声翻訳通信研究所／第一研究室／主幹研究員
 音声認識用文法: ATR 音声翻訳通信研究所第一研究室主幹研究員
 構文解析用文法: ATR 音声翻訳通信研究所第一研究室主幹研究員

学校の学科名、ツアー名、パンフレット名などは、ATR 対話データベースでは短単位分割している。音声認識用文法・構文解析用文法では、名詞連続複合語の生成規則がないので、全体を普通名詞として一語登録することになる。

データベース	音声認識	構文解析
普名詞 英文 接尾語 科	n-hutu 英文科	n-com 英文科
固名詞 イタリア 普名詞 ツアー	n-hutu イタリアツアー	n-com イタリアツアー

同様に、次のような語は、ATR 対話データベースでは短単位分割しているが、音声認識用文法・構文解析用文法では、長単位分割することになる。

データベース	音声認識	構文解析
固名詞 京 普名詞 料理	n-hutu 京料理	n-com 京料理
固名詞 関東 普名詞 地方	n-proper 関東地方	n-prop 関東地方
固名詞 キリスト 接尾語 教	n-proper キリスト教	n-prop キリスト教

2.1.2 サ変名詞

いわゆるサ変動詞を「サ変名詞 + 補助動詞」と分割するのは、ATR 対話データベース・音声認識用文法・構文解析用文法に共通である。

動詞の連用形から転用したものをサ変名詞とすることも共通である。

サ変名詞に敬語の「お・ご・御」がついたものは、ATR 対話データベースでは分割するが、音声認識用文法・構文解析用文法では長単位分割する。

データベース	音声認識	構文解析
サ名 参加 補動 し 助動詞 ます	n-sahen 参加 flex-sahen-renyo1 し masu-rentai ます	n-sahen 参加 auxv-npred し auxv-auxv-optt ます
サ名 申込み 補動 いたし 助動詞 ます	n-sahen 申込み dont-5dan-s いた flex-5-si し masu-rentai ます	n-sahen 申込み auxvstem-npred いた vinfl し auxv-optt ます
接頭 ご サ名 安心 補動 下さい	pre-v-sahen ご安心 dont-5dan-e-kougo 下さい	n-sahen ご安心 auxv-npred 下さい

「する」の前に助詞がくる時は、ATR 対話データベースでは普通名詞になる。音声認識用文法・構文解析用文法でも同様に普通名詞としている。

データベース		音声認識		構文解析	
普名詞	電話	n-hutu	電話	n-com	電話
格助詞	を	p-kaku-o	を	postp-oblg	を
補動	する	v-sa-rentai	する	vstem	する

＊改善意見(文献[1]に反映済)

「わたしが彼に電話をしましょう」のような文では、意味的に言って、「わたしが」「彼に」が「電話(する)」の補語だと考えた方がよいので、動作性がある場合はサ変名詞とした方がよい。

2.1.3 形容名詞

いわゆる形容動詞を「形容名詞 + 助動詞」と分割するのは、ATR 対話データベース・音声認識用文法・構文解析用文法に共通である。

「的」がつくものは、ATR 対話データベースでは「的」を含めて全体を形容名詞としている。これは、音声認識用文法・構文解析用文法でも同じである。

接頭語がつくものは、ATR 対話データベースでは、接頭語と名詞を切って「… な(名詞)」と置き替えて意味が通じたら切っている。音声認識用文法・構文解析用文法では、この種の接頭語は曖昧性の原因になるので、分割せずにすべて長単位登録している。

データベース		音声認識		構文解析	
形名	不便	kdo	不便	n-adj	不便
助動詞	です	flex-p-kdo-syusi	です	auxv-copl-adj	です
形名	具体的	kdo	具体的	n-adj	具体的
助動詞	な	flex-kdo-rentai	な	auxv-copl-adj	な
接頭	不	kdo	不明瞭	n-adj	不明瞭
形名	明瞭				
助動詞	な	flex-kdo-rentai	な	auxv-copl-adj	な
形名	不都合	kdo	不都合	n-adj	不都合
助動詞	な	flex-kdo-rentai	な	auxv-copl-adj	な

「形容名詞 + に」は、ATR 対話データベースでは、形容しているかどうかで、普通名詞か形容名詞かを判断している。音声認識用文法では、「形容名詞/に」と分割するのは ATR 対話データベースと同じである。構文解析用文法では、活用語の連用修飾の規則を作っていないので、全体を副詞として一語登録している。

データベース		音声認識		構文解析	
形名	具体的	kdo	具体的	adv-kernel	具体的に
助動詞	に	flex-kdo-renyo3	に		

2.1.4 普通名詞

構文解析用文法では、時間を表す名詞で格助詞を伴わずに連用修飾できるものと、常に連体修飾されて副詞的用法を持つ名詞に、それぞれ別の品詞名を与えている。音声認識用文

法では、それに加えて、形式名詞と、「いっぱい」が後続できる時詞にも、別の品詞名を与えている。

データベース	音声認識		構文解析	
普名詞 会議	n-hutu	会議	n-com	会議
こと	n-keisiki	こと	n-com	こと
先日	n-hutu-time	先日	n-temp	先日
今年	n-ippai	今年	n-temp	今年
際	n-adv	際	n-adv	際

ATR 対話データベースでは、普通名詞は意味を保つ限り小さく切るのを原則としている。これに対して、音声認識用文法・構文解析用文法では、接頭語・接尾語として登録されているものが少ないことと、名詞連続複合語の生成規則がないこととによって、限られたものを除いて大体長単位分割している。

データベース	音声認識		構文解析	
普名詞 案内 接尾語 状	n-hutu	案内状	n-com	案内状
固名詞 日本 接尾語 語	n-hutu	日本語	n-com	日本語
普名詞 自動 普名詞 翻訳 普名詞 電話	n-hutu	自動翻訳電話	n-com	自動翻訳電話

2.1.5 数詞

数詞および数量表現の扱いは、ATR 対話データベースでも音声認識用文法・構文解析用文法でも、たいへん複雑である。

まず数詞は、ATR 対話データベースでは長単位分割が原則である。これに対して、音声認識用文法と構文解析用文法では、桁を基準にして短単位分割している。また、音声認識用文法では、金額、年・月・日、人数などによって、品詞名も別々に与えている。

ただし、「数」「何」などの接頭語が数詞の前につく場合、ATR 対話データベースでは短単位分割している。音声認識用文法・構文解析用文法では、現在はまだ扱っていない。

また、電話番号に使われる棒読み型の数字は、ATR 対話データベースでは長単位分割しているが、音声認識用文法・構文解析用文法では短単位分割している。

データベース	音声認識	構文解析
数詞 三万五千 接尾語 円	money-num-man-1 三 money-man-suf 万 money-num-sen 五 money-sen-suf 千 money-suf 円	n-digit 三 digit-suffix-man 万 n-digit 五 digit-suffix-sen 千 n-quant-suffix 円
数詞 千九百 接尾語 年	anual-num-1000 千 anual-num-1 九 anual-num-100 百 anual-suf 年	n-num-sen 千 n-digit 九 digit-suffix-hyaku 百 n-temp-suffix 年
接頭語 数 数詞 万 接尾語 円	money-num-man-1 数? money-man-suf 万? money-suf 円	n-digit 数? digit-suffix-man 万? n-quant-suffix 円
数詞 三七二	tel-num 三 tel-num 七 tel-num 二	n-num-tel 三 n-num-tel 七 n-num-tel 二

数量表現は、接頭語・数詞・接尾語が接続して作られるが、ATR 対話データベースでは、すべて短単位分割している。これに対して、音声認識用文法・構文解析用文法では、接頭語・接尾語が限定されているので、短単位分割しているものもあれば、長単位分割しているものもあり、まちまちである。この部分が、ATR 対話データベース・音声認識用文法・構文解析用文法のあいだで、最もずれが大きい。

データベース	音声認識	構文解析
数詞 五 接尾語 人	human-num-1 五 count-human-suf 人	n-digit 五 n-quant-suffix 人
数詞 二 接尾語 時間	n-zahl 二時間	n-digit 二 n-quant-suffix 時間
数詞 一 接尾語 箇所 接尾語 所	n-zahal 一箇所	n-quant 一箇所
数詞 一 接尾語 週間 接尾語 間	n-zahl 一週間	n-digit 一 n-quant-suffix 週間
接頭語 第 数詞 二 接尾語 版	pre-suf-special 第二版	n-prefix 第 n-digit 二 n-quant-suffix 版

2.1.6 代名詞

音声認識用文法・構文解析用文法では、通常の代名詞と疑問代名詞を区別している。さらに、音声認識用文法では、通常の代名詞を「ども」を伴えるものと伴えないものとに分け、疑問代名詞を後続する助詞との接続関係を考慮して細分類している。

「わたしたち」は、ATR 対話データベースと音声認識用文法では短単位分割し、構文解析文法では長単位分割する。

データベース	音声認識		構文解析
代名詞 あなた	pro	あなた あれ	n-com あなた あれ
わたし	pro1	わたし	n-com わたし
誰	wh-pro-dare	誰	n-wh 誰
どなた	wh-pro-don	どなた	n-wh どなた
いくら	wh-pro-iku	いくら	n-wh いくら
代名詞 わたし	pro1	わたし	n-com わたしたち
接尾語 たち	pro-suf1	たち	

2.2 本動詞

ATR 対話データベースでは、語幹と活用語尾をあわせて「本動詞」と品詞づけしている。これに対して、音声認識用文法・構文解析用文法では、語幹と活用語尾に分割して、それぞれに品詞名を与えている。ただし、いくつかの動詞は、活用形を辞書登録してある。

また、音声認識用文法では、活用の型によって別々の品詞名を与えている。

データベース	音声認識		構文解析
本動詞 送る	v-5dan-r flex-5-ru	送 る	vstem 送 vinfl る
教える	v-1dan flex-1dan-rentai2	教え る	vstem 教え vinfl る
来る	v-rentai	来る	vstem 来る
する	v-sa-rentai	する	vstem する

複合動詞は、ATR 対話データベースでは、「新明解」に見出し語となっている場合はまとめて一語とし、見出し語でない場合は分割している。音声認識用文法・構文解析用文法では長単位分割している。

また、「電話をかけ(る)」「楽しみにする」などは、構文解析用文法では意味表現を考慮して一語登録している。「気に入る」は、音声認識用文法・構文解析用文法で一語登録している。

データベース	音声認識		構文解析
本動詞 書き直す	v-5dan-s flex-5-su	書き直 す	vstem 書き直 vinfl す
本動詞 書き 本動詞 添える	v-1dan flex-1dan-rentai2	書き添え る	vstem 書き添え vinfl る
接頭語 お 普名詞 礼 本動詞 申し上げる	v-1dan flex-1dan-rentai2	お礼申し上げ る	vstem お礼申し上げ vinfl る
普名詞 電話 格助詞 を 本動詞 かける	n-hutu p-kaku-o v-1dan flex-1dan-rentai2	電話 を かけ る	vstem 電話をかけ vinfl る
普名詞 気 格助詞 に 本動詞 入る	v-5dan-r flex-1dan-rentai2	気に入 る	vstem 気に入 vinfl る

2.3 補助動詞・助動詞

補助動詞と助動詞は文末表現を構成するものなので、ここでまとめて説明する。

本動詞と同様、ATR 対話データベースでは語幹と活用語尾をあわせて一語とするが、音声認識用文法と構文解析用文法では分割するのが基本である。ただし、いくつかの補助動詞・助動詞では語幹と活用語尾をあわせて一語登録しているものもある。

2.3.1 補助動詞

補助動詞には、次のような種類がある。

1. 「動詞＋テ」「助動詞＋テ」に続いて、アスペクトや授受などさまざまな意味を表すもの。

例: 機械翻訳という話題が案内書に載っていますが。
至急登録用紙を送ってください。

2. サ変名詞や動詞の連用形に続いて、可能、敬意などさまざまな意味を表すもの。

例: こちらでは専門的な質問にお答えできません。
先生もご出席になります。

3. 動詞の連用形について意味を添えるもの。

例: 本を読みはじめる。

4. 否定の助動詞などを伴って代動詞として用いられる「する」

例: そんなことは考えもしなかった。

「…テ」に続くものは、構文解析用文法では「テ＋補助動詞」を一語登録している。
また、「…テ」の後に係助詞の「は・も」が続き、さらに補助動詞が続くものも、構文解析用文法では全体を一語登録している。

データベース	音声認識		構文解析	
本動詞 載っ	v-5dan-r	載	vstem	載
接助詞 て	flx-r2-t-r	っ	vinfl	っ
補助動詞 いる	te-te	て	auxvstem-aspc	てい
	aspc-1dan	い	vinfl	る
	flex-1dan-rentai2	る		
本動詞 送っ	v-5dan-r	送	vstem	送
接助詞 て	flx-r2-t-r	っ	vinfl	っ
補助動詞 下さい	te-te	て	auxv-dont	て下さい
	dont-5dan-e-kougo	下さい		
本動詞 見	v-1dan	見	vstem	見
接助詞 て	te-te	て	auxvstem-aspc	てはい
係助詞 は	p-k-wa	は	vinfl	る
補助動詞 いる	aspc-1dan	い		
	flex-1dan-rentai2	る		

サ変名詞や動詞の連用形に続くものは、活用規則以外は、おおむね ATR 対話データベース・音声認識用文法・構文解析用文法で同じ分割である。ただ、敬意を表す「…にな(る)」は構文解析用文法では長単位分割している。

データベース	音声認識		構文解析	
サ名詞 参加	v-sahen	参加	n-sahen	参加
補助動詞 いたし	dont-5dan-s	いた	auxvstem-npred	いた
助動詞 ます	flex-d-5-si	し	vinfl	し
	masu-rentai	ます	auxv-optt	ます
接頭語 お	pre-v-sahen	お話し	n-sahen	お話し
本動詞 話し	p-kaku-ni	に	auxvstem-npred	にな
格助詞 に	dont-naru	な	vinfl	り
補助動詞 なり	flex-d-n-i	り	auxv-optt	ます
助動詞 ます	masu-rentai	ます		

*将来への展望

動詞の連用形について意味を添えるものは、音声認識用文法・構文解析用文法ではまだ現れていない。もし現れたとしても、連用形を項にもつ構文規則は副作用が大きいので、一語登録することになるだろう。

データベース	音声認識		構文解析	
本動詞 読み	v-1dan	読み始め	vstem	読み始め
補助動詞 始める	flex-1dan-rentai2	る	vinfl	る

***将来への展望**

否定の助動詞を伴って代動詞として用いられる「する」は、音声認識用文法・構文解析用文法ではまだ現れていない。もし現れたら、本動詞の「する」と同じ扱いをするだろう。

データベース	音声認識		構文解析	
本動詞 送り	v-5dan-r	送	n-sahen	送
係助詞 は	flex-5-ri	り	vinfl	り
補動詞 し	p-k-wa	は	postp-contr	は
助動詞 ない	v-sa-mizen1	し	vstem	し
	negt-kyo	な	auxv-negt	ない
	flex-kyo-rentai	い		

2.3.2 助動詞

ATR 対話データベースと音声認識用文法では、形態素分割はだいたい一致する。構文解析用文法では、語幹と活用語尾の扱いがまちまちである。

データベース	音声認識		構文解析	
助動詞 せる	caus-seru-rentai	せる	auxvstem-caus	せ
させる	caus-saseru-rentai	させる	vinfl	る
			auxvstm-caus	させ
			vinfl	る
れる	deac-reru-rentai	れる	auxvstem-deac	れ
られる	deac-rareru-rentai	られる	vinfl	る
			auxvstem-deac	られ
			vinfl	る
ます	masu-rentai	ます	auxv-optt	ます
たい	optt-rentai	たい	auxvstem-optt	た
			vinfl	い
らしい	evid-rentai	らしい	auxvstem-evid	らし
			vinfl	い
だ	cop-syusi	だ	auxv-copl-com	だ
です	v-cop-rentai	です	auxv-copl-com	です
そうだ	evid-rentai	そうだ	auxv-evid	そうだ
そうです	evid-rentai	そうです	auxv-evid	そうです
ようだ	conj-rentaif	ようだ	auxv-evid	ようだ
ようです	conj-rentai	ようです	auxv-evid	ようです
みたいだ	conj-rentai	みたいだ	auxv-evid	みたいだ
みたいです	conj-rentai	みたいです	auxv-evid	みたいです
た	past-ta	た	auxv-tense	た
だ	past-da	だ	auxv-tense	だ
う	intn	う	auxv-tense	う
よう	intn-you	よう	auxv-tense	よう
ない	negt-kyo	な	auxvstem-negt	な
	flex-kyo-rentai	い	vinfl	い

2.3.3 長単位分割の補助動詞・助動詞

構文解析用文法には、許可・禁止のようなムードを表す一群の表現や、推量・否定を表す表現で、長単位分割しているものがある。これらは、ATR 対話データベース・音声認識用文法では短単位分割しているが、音声認識用文法では将来的には構文解析用文法のように長単位分割してもよいかもしれない。

データベース		音声認識		構文解析	
接助詞	たら	sp-renyo-tara	たら	auxvstem-optt	たらい
形容詞	いい	kyo-ii	い	vinfl	い
		flex-kyo-rentai	い		
接助詞	ば	sp-katei-ba	ば	auxvstem-optt	ばい
形容詞	いい	kyo-ii	い	vinfl	い
		flex-kyo-rentai	い		
接助詞	ても	sp-renyo-temo	ても	auxvstem-optt	てもい
形容詞	いい	kyo-ii	い	vinfl	い
		flex-kyo-rentai	い		
接助詞	と	sp-rentai-to	と	auxvstem-optt	とよ
形容詞	よい	kyo-yoi	よ	vinfl	い
		flex-kyo-rentai	い		
助動詞	なく	negt-kyo	な	auxvstem-optt	なくてはいけな
接助詞	て	flex-kyo-renyo	く	vinfl	い
係助詞	は	sp-renyo-te	て		
		p-k-wa	は		
形容詞	いけない	kyo-a	いけな		
		flex-kyo-rentai	い		
助動詞	なけれ	negt-kyo	な	auxv-optt	なければなりま
接助詞	ば	flex-kyo-katei	けれ		せん
本動詞	なり	sp-katei-ba	ば		
助動詞	ませ	v-5dan-r	な		
助動詞	ん	flex-5-ri	り		
		masu-mizen1	ませ		
		nai-n-rentai	ん		
助動詞	ない	negt-kyo	な	auxv-optt	ないでください
接助詞	で	flex-kyo-rentai	い		
補動詞	ください	te-de	で		
		dont-5dan-e-kougo	くだ さい		
普名詞	もの	n-keisiki	もの	auxv-optt	ものだ
助動詞	だ	cop-syusi	だ		
格助詞	で	p-kaku-de	で	auxvstem-optt	でよろし
形容詞	よろしい	kyo	よろし	vinfl	い
		flex-kyo-rentai	い		

データベース	音声認識		構文解析	
終助詞 か	evid-rentai	かもしれ	auxvstem-evid	かもしれ
係助詞 も		ません		ません
本動詞 しれ				
助動詞 ませ				
助動詞 ん				
助動詞 に	φ		auxvstem-evid	にちがいな
形容詞 ちがいな				
助動詞 い			vinfl	い
助動詞 で	cop-rentai1	で	auxv-copl-adj	で
係助詞 は	p-k-wa	は	auxv-negt	はありません
助動詞 あり	negt-rentai-kyo	ありま		
助動詞 ませ		せん		
助動詞 ん				
普名詞 ところ	n-hutu	ところ	auxv-evid	ところです
助動詞 です	polt-aux-rentai	です		
普名詞 つもり	n-hutu	つもり	auxv-evid	つもりです
助動詞 です	polt-aux-rentai	です		
普名詞 はず	n-hutu	はず	auxv-evid	はずです
助動詞 です	polt-aux-rentai	です		
普名詞 ばかり	n-hutu	ばかり	auxv-evid	ばかりです
助動詞 です	polt-aux-rentai	です		

2.4 形容詞

ATR 対話データベースでは語幹と活用語尾をあわせて一語とするが、音声認識用文法と構文解析用文法では語幹と活用語尾を分割する。

音声認識用文法では、活用の仕方に応じて品詞を細分化している。

データベース	音声認識		構文解析	
形容詞 新しい	kyo	新し	vstem	新し
	flex-kyo-rentai	い	vinfl	い
近い	kyo-a	近	vstem	近
	flex-kyo-rentai	い	vinfl	い
いい	kyo-ii	い	vstem	い
	flex-kyo-rentai	い	vinfl	い
良い	kyo-yoi	良	vstem	良
	flex-kyo-rentai	い	vinfl	い

2.5 副詞

副詞は定義するのがむずかしい品詞である。したがって、語によっては、ATR 対話データベース・音声認識用文法・構文解析用文法で、扱いがまちまちである。

音声認識用文法では、後続する助詞によって細分化している。また、構文解析用文法では、修飾する述語階層によって細分化している。

「に」のつく副詞は、ATR 対話データベース・音声認識用文法では短単位分割しているが、構文解析用文法では「に」を含めて長単位分割している。

「と」のつく副詞は、ATR 対話データベースでは短単位分割しているが、音声認識用文法・構文解析用文法では「と」を含めて長単位分割している。

データベース		音声認識		構文解析	
副詞	まだ	adv-b	まだ	adv-kernel	まだ
	そう	adv-k-s	そう	adv-kernel	そう
	どう	adv-k-d	どう	adv-kernel	どう
	かなり	adv-num	かなり	adv-degr	かなり
	たぶん	adv-b	たぶん	adv-sent	たぶん
副詞	すぐ	adv-ni	すぐ	adv-kernel	すぐに
格助詞	に	p-kaku-ni	に		
副詞	ゆっくり	adv-b	ゆっくりと	adv-kernel	ゆっくりと
格助詞	と				

形容詞や形容動詞の連用形が連用修飾する場合、ATR 対話データベースでは「形容詞」「形容名詞 + 助動詞」となるが、音声認識用文法・構文解析用文法では副詞としている。

データベース		音声認識		構文解析	
形容詞	早く	adv-b	早く	adv-kernel	早く
形名詞	簡単	adv-b	簡単に	adv-kernel	簡単に
助動詞	に				

その他に、音声認識用文法や構文解析用文法では、連用修飾的な語句を副詞として長単位分割しているものがある。慣用的な語句や言い回しに多い。ただ、その認定基準は曖昧で、かなり恣意的である。

データベース		音声認識		構文解析
代名詞	これ	pro	これ	adv-kernel これから
格助詞	から	p-kaku-kara	から	
代名詞	そこ	pro	そこ	adv-kernel そこまで
格助詞	まで	p-kaku-made	まで	
副詞	どう	adv-b	どうしても	adv-kernel どうしても
本動詞	し			
接助詞	ても			
副詞	どう	adv-k-d	どう	adv-kernel どうやって
本動詞	やっ	v-5dan-r	や	
接助詞	て	flex-d-r2-t-r	っ	
		sp-renyo-te	て	
副詞	どう	adv-s1	どうすれば	adv-kernel どう vstem すれ fadv ば
本動詞	すれ			
接助詞	ば			
連体詞	どの	rentai	どの	adv-kernel どのように
助動詞	ように	meta-renyo	ように	
普名詞	今	n-hutu-time	今	adv-kernel 今のところ
格助詞	の	p-kaku-no	の	
普名詞	ところ	n-adv	ところ	
普名詞	念	n-hutu	念	adv-kernel 念のため
格助詞	の	p-kaku-no	の	
普名詞	ため	n-adv	ため	
副詞	もう	adv-b	もう	adv-kernel もう一度
副詞	一度	adv-b	一度	
形名詞	失礼	kdo	失礼	adv-conj 失礼ですが
助動詞	です	flex-p-kdo-syusi	です	
接助詞	が	sp-rentai-both+f	が	
形名詞	残念	kdo	残念	adv-sent 残念ですが
助動詞	です	flex-p-kdo-syusi	です	
接助詞	が	sp-rentai-both+f	が	
接頭語	お	kdo	お気の毒	adv-sent お気の毒ですが
形名詞	気の毒	flex-p-kdo-syusi	です	
助動詞	です	sp-rentai-both+f	が	
接助詞	が			
普名詞	申し訳	n-exc	申し訳	adv-sent 申し訳ありませんが
本動詞	あり	v-arui	あ	
助動詞	ませ	flex-arui-i	り	
助動詞	ん	masu-mizen1	ませ	
接助詞	が	nai-n-rentai	ん	
		sp-rentai-bothf	が	

2.6 連体詞

ATR 対話データベースで「連体修飾にならない連体詞」「『こう、そう、ああ、どう』のつく連体詞」として短単位分割されているものは、音声認識用文法や構文解析用文法では、大体、長単位分割される。

「大きな」「小さな」は、ATR 対話データベースと音声認識用文法では長単位分割するが、構文解析用文法では短単位分割する。

データベース		音声認識		構文解析	
連体詞	どの	rentai	どの	adn	どのような
助動詞	ような	meta-rentai	ような		
副詞	そう	rentai	そういう	adn	そういう
本動詞	いう				
連体詞	大きな	rentai	大きな	vstem	大きな
				vinfl	な
連体詞	小さな	rentai	小さな	vstem	小さな
				vinfl	な

2.7 接続詞

ATR 対話データベース・音声認識用文法・構文解析用文法で、あまり差異がない。ただ、「そうしたら」は ATR 対話データベースでは短単位分割するが、音声認識用文法・構文解析用文法では長単位分割する。

データベース		音声認識		構文解析	
副詞	そう	conj	そうしたら	adv-conj	そうしたら
本動詞	し				
接助詞	たら				

2.8 感動詞・間投詞

ATR 対話データベースでは感動詞と間投詞に別々の品詞を与えているが、音声認識用文法と構文解析用文法では同じ品詞を与えている。

感動詞の認定は、ATR 対話データベース・音声認識用文法・構文解析用文法ともに、大体において、「日本語形態素解析マニュアル」に従っているが、活用語を含むものは、若干の差異がある。

データベース	音声認識	構文解析
感動詞 ありがとう	kyo-renyo3 ありがとう	v-interj ありがとう
補動詞 ございます	polt-v-renyo1 ござい masu-rentai ます	ございます
副詞 そう	interj そうですねえ	v-interj そうですねえ
助動詞 です		
終助詞 ねえ		
形容詞 いい	interj いいですよ	v-interj いいですよ
助動詞 です		
終助詞 よ		
名詞 申し訳	interj 申し訳ありま	v-interj 申し訳ありま
本動詞 あり	せん	せん
助動詞 ませ		
助動詞 ん		

N.B. 「そうですねえ」「申し訳ありません」は、音声認識用文法では短単位分割もされるようになっており、問題である。

2.9 助詞

ATR 対話データベースと構文解析用文法は、大体において同じ品詞分類に従っている。音声認識用文法は、一つの助詞に対して一つの品詞を与えるというように極端に細分化している。

分割の仕方の問題になるのは、いくつかの助詞が接続する場合と、いわゆる複合格助詞の場合である。

ATR 対話データベースと音声認識用文法では、「までに」は「まで」と「に」に分割されるが、構文解析用文法では一語登録されている。

複合格助詞は、ATR 対話データベースでは次のもののみを認めている。

として、について、において、における、にとって、にしろ、にせよ、でもって、をもって

音声認識用文法では、次のものを複合格助詞としている。

として、について、において、にとって、に関して、によって、と共に、に関する、による、に対する、についての

構文解析用文法では、次のものを複合格助詞としている。

として、において、にとって、に関して、に関する、によって、による、と共に、に対して、に対する、と一緒に、を通して、と比べて、に比べて、につき、にかけて、という

これらに対応する丁寧体は、ATR 対話データベースでは短単位分割するが、音声認識用文法・構文解析用文法では長単位分割する。

データベース		音声認識		構文解析
格助詞	まで	p-kaku-made	まで	postp-optn までに
格助詞	に	p-kaku-ni	に	
格助詞	に	p-kaku-nikanshite	に関して	postp-optn に関して
本動詞	関し			
接助詞	て			
格助詞	に	p-special	に関する	fadn に関する
本動詞	関する			
格助詞	と	p-kaku-to	と	postp-optn と一緒に
形名詞	一緒	kdo	一緒	
助動詞	に	flex-kdo-renyo3	に	

引用の格助詞は、ATR 対話データベースでは「と、って」、音声認識用文法では「と、って、とか」、構文解析用文法では「と、って、とか、か、かどうか、ように」を認定している。

データベース		音声認識		構文解析
格助詞	と	p-i-toka	とか	postp-quote とか
副助詞	か			
副助詞	か	sfp-1	か	postp-quote かどうか
副詞	どう	adv-k-d	どう	
副助詞	か	p-f-ka	か	

副助詞・接続助詞・準体助詞、終助詞は、ATR 対話データベース・音声認識用文法・構文解析用文法で、分割の仕方に差異がない。

並列助詞は、ATR 対話データベースでは次のもののみを認めている。

と、や、なり、たり、だり、の、たの、だの、とか

音声認識用文法では次のもののみを並列助詞としている。

と、や、とか

構文解析用文法では通常の並列助詞の他に、本来は接続詞であったり複合句であったりするが、文中で名詞と名詞を結び付けて名詞句を形成しているものも並列助詞としている。

と、や、も、か、とか、から、および、または、ないし、あるいは、もしくは、とかいう

データベース		音声認識		構文解析
接続詞	ないし	conj2	ないし	fadn ないし
格助詞	と	p-i-toka	とか	fadn とかいう
副助詞	か	v-5dan-w	い	
本動詞	いう	flex-5-wu	う	

2.10 接辞

接辞は、ATR 対話データベースでは原則としてすべて短単位分割するが、音声認識用文法・構文解析用文法では、そのうちごく少数のものしか認定していないので、それ以外のものはすべて長単位分割する。

2.10.1 接頭辞

音声認識用文法では、次のものを接頭辞として認めている。

約

構文解析用文法では、次のものを接頭辞として認めている。

お(御)、ご(御)、約、第、各、全、およそ

2.10.2 接尾辞

音声認識用文法では、次のものを接尾辞として認めている。

都、府、市、区、丁目、番、号

年、月、日、時、分

万、千、百、円

様、さん、先生、たち、ども、人、名

構文解析用文法では、次のものを接尾辞として認めている。

都、府、県、市、区、丁目、番、号

年、日、時、分、週間、週、カ月、時間、晩、前、後、ごろ、以後、おき

万、千、百、十、円、ドル

様、さま、さん、先生、人、名

回、カ国語、語、件、種類、字、通、種類、枚、ページ、箇所、つ、以上、版

中、あて

3 表記の揺れ

表記の揺れには、代表的なものとして次の4つがある。

1. 同一単語の表記の揺れ

例: 私共、私ども、わた(く)しども

2. 送り仮名の揺れ

例: 申(し)込(み)

3. 数詞の表記の揺れ

例: 五千、5千、五〇〇〇、5000

4. 外来語・外国語固有名詞のカタカナ表記の揺れ

例: コンピューター、コンピュータ
バイオリン、ヴァイオリン

音声認識用文法・構文解析用文法では、これらについて以下のような原則を立てている。

3.1 同一単語の表記の揺れ

漢字を使うことによって意味が明瞭になるものは漢字で、漢字があまり意味を持っていないものは平仮名で表記する。

この原則に従うと、名詞・動詞・形容詞のように、漢字がその語の意味を明かにする品詞は漢字で表記する。その他の、感動詞・副詞・連体詞・代名詞・助動詞・補助動詞・助詞のように、漢字があまり意味を持たない品詞は原則として平仮名で表記する。

名詞: 昨日 (×きのう、さくじつ)
今日 (×きょう、こんにち)
明日 (×あす、あした、みょうにち)
あたり (×辺り)

代名詞: これ (×此)
いつ (×何時)

動詞: 済む (×すむ)
来る (×くる)
ございます (×御座います)
さしあげる (×差し上げる)
できる (×出来る)
なさる (×為さる)

形容詞: 難しい (×むずかしい)
欲しい (×ほしい)
ありがたい (×有り難い)
おもしろい (×面白い)

副詞: おそらく (×恐らく)
たくさん (×沢山)
ちょうど (×丁度)
以前 (×いぜん)
一応 (×いちおう)
先ほど (×先程、さきほど)

連体詞: いろんな (×いろんな)
この (×此の)
大きな (×おおきな)

感動詞: ありがとう (×有り難う)
すみません (×済みません)
申し訳ありません (×もうしわけありません)

接続詞: および (×及び)
また (×又)

補動詞: いたす (×致す)
(ても／ば) よい (×良い、善い)
願う (×ねがう)

助詞: まで (×迄)
くらい (×位)
において (×に於いて)
に関して (×にかんして)
に対して (×にたいして)

形式名詞: 「こと」「もの」は、実質的な意義を持つ場合は漢字で表記する。
また、人を表す「かた」は漢字で、方向や比較の対象を表す「ほう」は仮名で表記する。

お聞きしたい事があります。
京都へ行ったことがありますか。

いい物をあげよう。
ツアーに参加する者は集まれ。
あきらめたものと見える。

ツアーに参加する方は受付にお集まり下さい。
プリンスホテルのほうが安い。

3.2 送り仮名

活用のある語は活用語尾を送る。ただし、一段活用の動詞は「-i」「-e」の部分も送る。
また、活用形によって同形異義になることのあるものは、互いに区別できるように送る。

言う
過ぎる
当てる

行(い)った
行(おこ)なった (×行った)

長い
早い

語幹が「し」で終わる形容詞は「し」から送る。

忙しい
涼しい

活用語尾以外の部分に他の語を含む語は、含まれている語の送り仮名の付け方によって送る。

済ます (← 済む)
確かめる (← 確かだ)
重たい (← 重い)

複合語は、複合語を構成する各語の送り仮名を送る。ただし、特定の領域で慣用が固定しているものは、慣用に従って送り仮名をつけない。

手書き (×手書)
書き方 (×書方)
振り込み (×振込、振込み)

肩書 (×肩書き)
銀行振込 (×銀行振り込み、銀行振込み)

3.3 数詞

数詞には、漢数字による表記、算用数字による表記、両者の混ぜ書きによる表記がある。さらに、桁を表す単位を表記するものとししないものや、コンマを含むものと含まないものがある。また、桁読みと棒読みの違いもある。

漢数字: (桁表記) 三万五千
(桁無表記) 三五〇〇〇／三五,〇〇〇
算用数字: 35000／35,000
混ぜ書き: 3万5千

音声認識用文法・構文解析用文法では、現在は、漢数字の桁表記のみを許している。

3.4 外来語・外国語固有名詞

外来語の表記の揺れが生じる主な原因は、日本語の音韻体系にない音を含むこと、導入された時期や分野によって表記の仕方が異なることなどである。

日本語の音韻体系にない音には次のようなものがある。これらは、音声翻訳システム ASURA の音声認識部では、音素モデルとして用意されていない。

ツア: モーツァルト
ツイ: ソルジェニーツイン
ツエ: フィレンツェ
ツオ: カンツォーネ
テイ: テーパーティー
テュ: テュービンゲン
デイ: ビルディング
デュ: プロデューサー
ファ: ファックス
フィ: フィルム
フェ: エッフェル塔
フォ: フォーク
フュ: フュージョン

イエ: イエルサレム (エルサレム)
ウイ: ウイスキー (ウイスキー)
ウエ: ウェールズ (ウエールズ)
ウオ: ウォール街 (ウォール街)
クア: クアルテット (クアルテット、カルテット)
グア: グアテマラ (グアテマラ、ガテマラ)
クイ: クインテット (クインテット)
クエ: クエート (クエート)
クオ: クォーク (クォーク)
トゥ: トゥールーズ (ツールーズ)
ドゥ: ヒンドゥー教 (ヒンズー教)
ヴァ: ヴァイオリン (バイオリン)
ヴィ: ヴィーナス (ビーナス)
ヴ: ヴラマンク (ブラマンク)
ヴェ: ヴェルサイユ (ベルサイユ)
ヴォ: ヴォルテール (ボルテール)
ヴュ: インタビュー (インタビュー)

長音記号も外来語に特有な表記であり、とくに語末の長音を表記するか否かで揺れることが多い。これなどは、専門分野によって慣用が違うだろう。音声認識用文法・構文解析用文法では、長音を表記している。

一: コンピューター/コンピュータ

4 むすび

この報告書に記されている差異の調査に基づき、ATR 音声翻訳通信研究所における音声言語データベースの日本語形態素情報の体系は新たに設計され直している [1]。したがって、文献 [1] を読む際には、本書を見比べながら読み進めるとよい。また、本報告書より以前にも一度同じような調査が行なわれている [2]。本報告書は、文献 [2] をもとに、自動翻訳

電話研究所において開発された音声翻訳システム ASURA で採用されている最終版の文法に対して調査を発展させたものである。

参考文献

- [1] 浦谷則好, 竹沢寿幸, 田代敏久, 衛藤純司: “音声言語データベースのための日本語形態素情報と表記の体系”, *ATR Technical Report*, TR-IT-0003 (1993-07).
- [2] 永田昌明, 保坂順子, 鈴木雅実: “音声認識部、形態素解析部、構文解析部の品詞・語彙区切りの相違点の調査”, *ATR Technical Report*, TR-I-0245 (1992-02).
- [3] 東洋情報システム: 「日本語形態素解析マニュアル」 (1993).
- [4] 日本放送出版協会: 「新用字用語辞典」 (1986).
- [5] 朝日新聞社: 「朝日新聞の用語の手引き」 (1989).