Internal Use Only（非公開）

TR-IT-0354

TDMTとCHATRのインタフェース：言語フィルター
Interface between TDMT and CHATR: language filter

下村　賢*　　キャンベル　ニック
Ken Shimomura *　　Nick Campbell

2000.2.1

内容梗概

TDMT出力から言語ごとに異なる形式の「CHATR入力情報」を生成する言語フィルターに関して各言語の入出力サンプルを例に挙げてその機能を説明する。

ATR音声翻訳通信研究所
ATR Interpreting Telecommunications Research Laboratories
© 株式会社 エイ・ティ・アール音声翻訳通信研究所
© 2000 by ATR Interpreting Telecommunications Research Laboratories
* 株式会社東洋情報システム

TDMT と CHATR のインターフェイス: 言語フィルター

目次

1 更新履歴 **3**

2 言語フィルターとは… **3**

3 言語フィルター **4**

3.1 IJCAI '97 デモシステムの構成 4
3.2 MATRIX での言語フィルターの位置付け 4
3.3 MATRIX での言語フィルターの機能 5
3.3.1 言語ごとの CHATR 入力レベル 5
3.3.2 部分翻訳の扱い 5
3.3.3 韓国語文に含まれるアラビア数字 5
3.3.4 中国語の BI 6
3.3.5 翻訳結果中の不明な記述 6
3.3.6 MATRIX 使用話者一覧 6
3.3.7 C-STAR II 使用話者一覧 6
3.4 言語フィルターの構成 7
3.5 言語フィルターへの入力データ 8
3.6 各言語の入力サンプルデータ 8
3.7 言語フィルターの日本語処理 10
3.7.1 mor2bou 入出力例 10
3.7.2 mor2bou 入力フォーマット 10
3.7.3 mor2bou への入力 12
3.7.4 kan2rom 14
3.7.5 無声化フィルタの追加 14
3.7.6 無声化フィルタの実行例 14
3.8 言語フィルターの多言語処理 15
3.8.1 言語フィルターの英語処理 15
3.8.2 言語フィルターの韓国語処理 15
3.8.3 言語フィルターの中国語処理 15
3.8.4 言語フィルターのドイツ語処理 16
3.9 言語フィルターのコンパイル 18
3.9.1 ディレクトリ構成 18
3.9.2 ソースのチェックアウト 19
3.9.3 morph のコンパイル 19
3.9.4 gfilter のコンパイル 19
3.10 言語フィルターの起動方法 19
3.10.1 gfilter 実行 19
3.10.2 言語フィルターの動作環境 19

4 問題点 **20**

4.1 英数字対応 20

5 言語フィルターの改善 **21**
5.1 言語フィルターの改善に伴う 形態素区切りの変更点 21
5.1.1 TDMT 日本語辞書の種類 21
5.1.2 差分辞書の修正 22
5.1.3 生成辞書の修正 23
5.2 言語フィルターの改善に伴う gfilter の変更点 24
5.2.1 付属語処理の追加 24
5.2.2 入力エントリの分解 28
5.2.3 ＢＩの変更 28
5.2.4 無声化処理の切り替え 28
5.3 言語フィルターの改善で修正したファイル 29
5.3.1 TDMT 辞書 29
5.3.2 gfilter 29
5.3.3 リカバリ機能 29
5.3.4 関数の追加 29

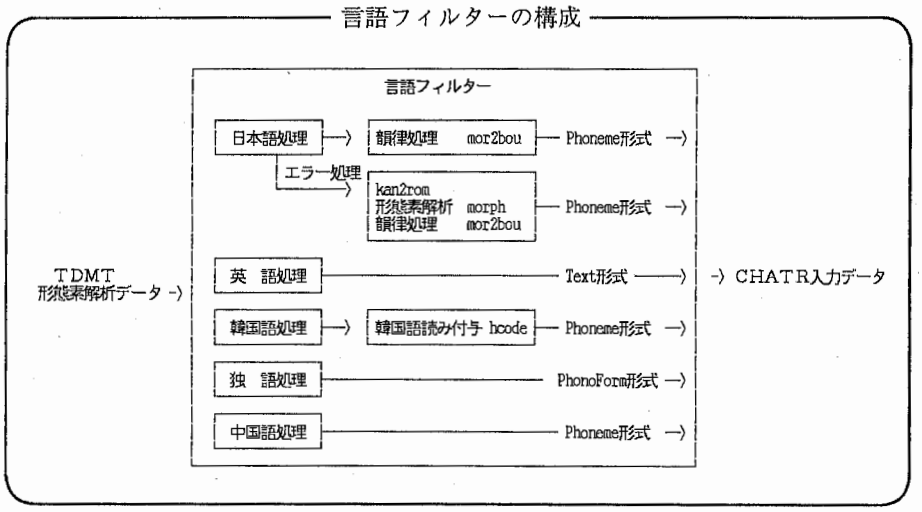
6 APPENDIX **30**
6.1 TDMT データサンプル 30
6.2 言語フィルターの改善で使用した TDMT データサンプル 30
6.3 gfilter 関数 31
6.4 プログラム関数一覧 32
6.5 ドイツ語 PhonoForm 処理参考資料 34

1 更新履歴

- 1997 年 12 月 19 日作成 「IJCAI デモシステムにおける言語フィルターについて」
- 1998 年 5 月 29 日更新 「Chat Translation における言語フィルター」
- 1998 年 11 月 11 日更新 「MATRIX における言語フィルター」
  日本語部分翻訳 (エラー処理) 対応, JC(日中)-MATRIX 対応
- 1999 年 3 月 5 日更新 「言語フィルターの改善」 [3] を追加
- 2000 年 1 月 28 日更新 「C-STAR II 使用話者一覧」を追加

2 言語フィルターとは…

TDMT 出力の言語別形態素解析データから CHATR に必要な入力情報を生成するのが言語フィルターの役割です。

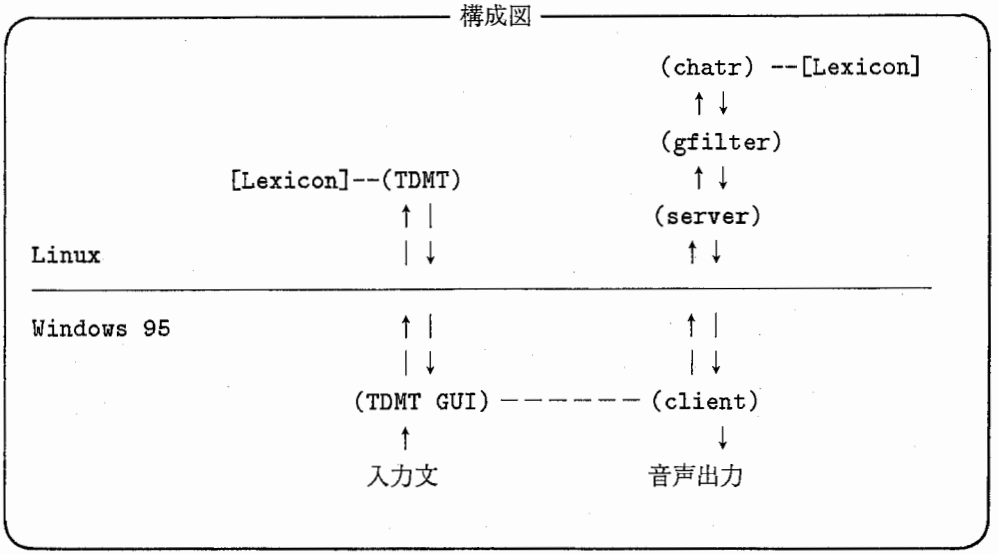


3 言語フィルター

IJCAI で発表された Chat Translation で使われた言語フィルターが MATRIX に組み込まれるまでに更新された内容やその状況などに付いて説明する。

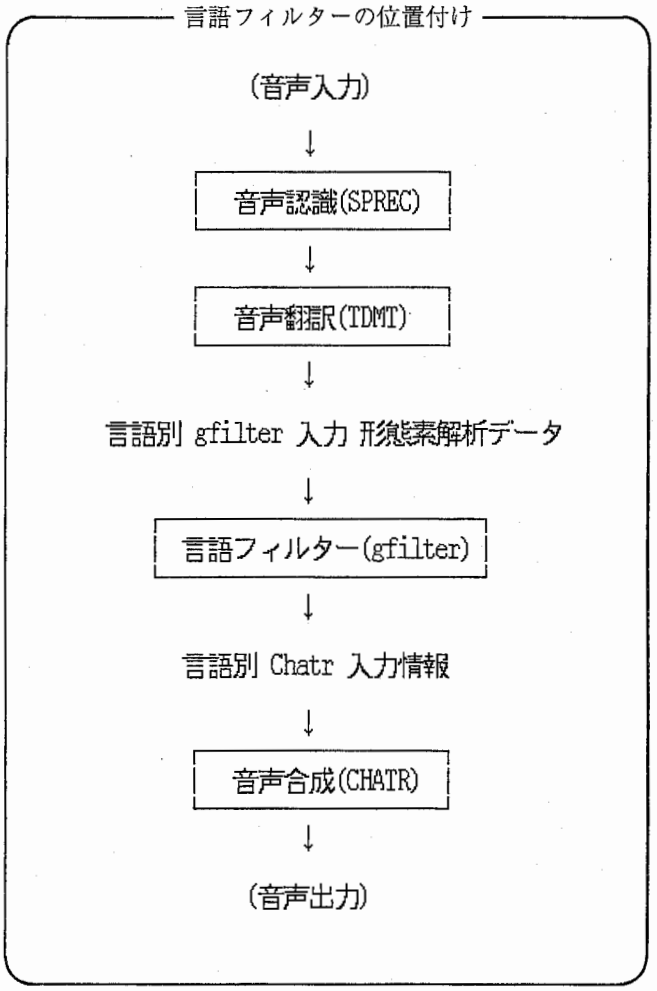
3.1 IJCAI '97 デモシステムの構成

構成図



3.2 MATRIX での言語フィルターの位置付け

言語フィルターの位置付け



3.3 MATRIX での言語フィルターの機能

現在の gfilter の機能を以下に示す。TDMT 出力から言語ごとに異なる形式の「CHATR 入力情報」を作成する。

3.3.1 言語ごとの CHATR 入力レベル

言語	日本語	英語	ドイツ語	韓国語	中国語
	↓	↓	↓	↓	↓
Text	↓	○	↓	○	↓
Phoneme	○		↓		○
PhonoForm			○		

3.3.2 部分翻訳の扱い

・TDMT の部分翻訳は、

入力データのうち、「入力言語の言い回し」として
解析することができなかった部分。

を翻訳対象から外す機能であり、部分翻訳結果は音声合成対象では
ないので、MATRIX の制御部で、
「音声合成対象」と「部分翻訳結果（= 合成対象でない）」とを
切り分けて、「音声合成対象」のみを音声合成部（gfilter + CHATR）へ
渡すものとする。

→日、中については、「||3|」の形式で TDMT から出力される。
→韓、独、英についても、基本的には、
日、中と同じ形式で出力する方針だが、問題あれば、関係者で相談する。

→音声合成処理のタイムラグがあるので、
「音声合成対象」ごとに音声合成部（gfilter + CHATR）を
呼び出せば、「部分翻訳」のためにわざわざポーズを
設定しないでも、勝手にある程度の無音が生じると思われる。

3.3.3 韓国語文に含まれるアラビア数字

・hcode は、韓国語の文字しか処理できないので、
数字も、アラビア数字でなく韓国語の数字で hcode に渡す必要がある。
→ MATRIX の処理では、TDMT から出力される韓国語の数字は、
アラビア数字でなく韓国語の文字で出力されるので MATRIX では問題がない。
但し、Chat Translation においてはその限りではない。

3.3.4 中国語の BI

中国語対応では、TDMT 出力の読み情報から CHATR の Phoneme を作成している為ポーズを挿入する場合や部分翻訳の場合にも同様に読み情報へ Break index の値を設定する。

- BI の種類

  3 研側との調整結果、BI(Break Index) は以下の通りの形式で TDMT から出力するようにする。

BI の種類

句点（。）　：「||5|」
読点（、）　：「||3|」
その他（，・）：「||3|」

3.3.5 翻訳結果中の不明な記述

8／26 時点の翻訳結果中に含まれていた不明な記述の内容は、以下のとおり。

不明な記述	説明
ng	単語を生成するための辞書のエントリはあるが、 「単語に対応するヨミ」のエントリが、ヨミ付与用の辞書にない場合に NG が発生する。
NIL	間投詞については MATRIX で前処理し、翻訳処理に渡さないため翻訳結果がない。 → MATRIX で合成処理に送らないようにする。
###########	TDMT から「NIL」を渡されたときに、 MATRIX で「NIL」を「########」に変換して表示していた。
...	部分翻訳の結果として、TDMT から出力する。 → MATRIX で合成処理に送らないようにする。
---	不要 → TDMT から出力しないようにする。

3.3.6 MATRIX 使用話者一覧

言語	男性	女性
日本語	MYS	FMP
英 語	nes2	mlp450
ドイツ語	kko	rtd
韓国語	mcu	son_m
中国語	cs4m	cs9f

※ 1998/11/06 現在

3.3.7 C-STAR II 使用話者一覧

言語	男性	女性
日本語	MYS/MYA	FMP/FOM

※ 1999/07/22 現在

3.4 言語フィルターの構成

言語フィルター (以下 gfilter) は、 TDMT の出力を基に CHATR での音声合成に必要な情報を作成します。gfilter は、以下の処理から出来ています。

- 日本語処理
  TDMT の出力から、 CHATR の Utterance Phoneme 形式のアクセント付きローマ字表記を作成する。
  アクセント処理、音素表記変換は mor2bou[1] で行う。

- 英語処理
  TDMT の出力から、 CHATR の Utterance Text 形式の表記を作成する。
  アクセント処理、音素表記変換は CHATR 内部で行う。

- ドイツ語処理
  TDMT の出力から、 CHATR の Utterance PhonoForm 形式の表記を作成する。
  アクセント処理、音素表記変換は gfilter で行う。

- 韓国語処理
  TDMT の出力から、CHATR の Utterance Phoneme 形式の表記を作成する。
  アクセント処理、音素表記変換は hcode で行う。

- 中国語処理
  TDMT の出力から、 CHATR の Utterance Phoneme 形式の表記を作成する。

3.5 言語フィルターへの入力データ

server から gfilter へ渡される データフォーマットは以下の通りです。

server → gfilter

```
言語情報 話者情報 表層 | 品詞 | 読み | アクセント情報 \n
表層 | 品詞 | 読み | アクセント情報 \n
 :      :   :          :
 :      :   :          :
表層 | 品詞 | 読み | アクセント情報 \n
```

最初に言語情報と話者情報、以降が言語別翻訳結果の情報が格納されてます。
言語情報 [J,E,G,K,C] J: 日本語 E: 英語 G: ドイツ語 K: 韓国語 C: 中国語
話者情報 [M,F] M: 男性 F: 女性

3.6 各言語の入力サンプルデータ

- 日本語 (Shift JIS code)

日本語サンプル

```
JM スイートは | 副助詞 | すいーと は |4000001000000020 0
一 | 普通名詞 |1|1200001000000020
泊 | 普通名詞 | はく |9100001000002020
七百 | 数詞 |700|1200001000000011
ドル | 接尾辞 | どる |9100001000003210
です | 助動詞 | です |0
、 | 特殊 | |
ツインルームは | 副助詞 | ついんるーむ は |1000001000000040 0
一 | 普通名詞 |1|1200001000000020
泊 | 普通名詞 | はく |9100001000002020
三百 | 数詞 |300|1200001000000011
ドル | 接尾辞 | どる |9100001000003210
です | 助動詞 | です |0
。 | 特殊 | |
```

- 英語 (MULE internal code)

英語サンプル

```
EMI'd|PRON||
like|PREP||
to|PREP||
have|V||
a|DET||
reservation|CN||
for|PREP||
the|DET||
hotel|CN||
```

- ドイツ語 (MULE internal code)

ドイツ語サンプル

```
GMes|SATZ:SUBJECT:PERSONALPRONOMEN|@.s.|0
gibt|SATZ:V-FIN:VERB|g.I.b.t.|1
eine apotheke|SATZ:OBJECT:DETERMINATIV-INDEF NOMEN|aI.-n.@. a.-p.o:.-t.e:.-k.@.|10 0010
nahe|SATZ:F-MIDDLE:ADJEKTIV|n.a:.-@.|10
.|SATZ:BOUNDARY-END:INTERPUNKTION|4|0
sie|SATZ:SUBJECT:PERSONALPRONOMEN|z.i:.|1
ist|SATZ:V-FIN:HILFSVERB|I.s.t.|1
neun uhr|SATZ:TIME:KARDINALZAHL NOMEN|n.OY.n. u:.6.|1 1
geoeffnet|SATZ:V-INF:VERB|g.@.-9.f.-n.@.t.|010
.|SATZ:BOUNDARY-END:INTERPUNKTION|4|0
```

- 韓国語 (EUC code)

韓国語サンプル

```
KM호텔|NCOMM||
예약|NACTV||
부탁|NACTV||
하|XSUFFVERB||
고 싶|VAUX1||
은데|VAUX2TEND||
요|VAUX2SEND||
```

- 中国語 (EUC china code)

中国語サンプル

```
CM您|先頭|ning2|
好|叹|hao3|
，|符|3|
这里|地|zhe4 li3|
是|判|shi4|
半岛酒店|地|ban4 dao3 fan4 dian4|
。|符|5|
```

3.7 言語フィルターの日本語処理

TDMT 側の情報から 韻律処理 (以下 mor2bou[1]) への入力データを作成し、
mor2bou の処理結果を gfilter の出力とする。
但し、TDMT 側の情報に不備があった際には、mor2bou ではなく kan2rom[2] に処理を移行する。

3.7.1 mor2bou 入出力例

mor2bou 入力例

```
スイートは スイート ハ 0 4000001000000020
一 1 0 0 1200001000000020
泊 ハク 0 0 9100001000002020
七百 700 0 0 1200001000000011
ドルです ドル デス 0 9100001000003210
、 0 0 0 d000001000000001
ツインルームは ツインルーム ハ 0 1000001000000040
一 1 0 0 1200001000000020
泊 ハク 0 0 9100001000002020
三百 300 0 0 1200001000000011
ドルです ドル デス 0 9100001000003210
。 0 0 0 0000001000000001
```

mor2bou 出力例

```
sui'itowa 1 ippaku' 1 nanahyaku'dorudesu 1 tsuiNru'umuwa 2 ippaku' 1 saNbyaku'dorudesu 5
```

3.7.2 mor2bou 入力フォーマット

形態素解析処理データ		
フィールド名	バイト	備考
漢字	可変長	全角
フィールドの区切り	1	半角スペース
漢字の読み	可変長	半角カタカナ
フィールドの区切り	1	半角スペース
付属語	可変長	半角カタカナ
フィールドの区切り	1	半角スペース
最後の付属語の分類	可変長	該当するものがない場合は "0"
フィールドの区切り	1	半角スペース
品詞コード	3	
派生情報	1	
複合語情報	2	
実効文節数	1	
語構成	1	
分離情報	1	
語句の区切り	1	
連濁情報	1	
アクセント接続情報	2	
アクセント情報	2	
音訓情報	1	
読み選択情報	可変長	

品詞コードの抜粋	
0: 単漢字	000 単漢一般 010 固有名詞一般
1: 名詞	100 一般名詞 110 固有名詞一般 120 数字
2: 動詞	200 カ五 290 上一下一 2a0 カ変 2a1 カ変 2a2 カ変 2b0 サ変 a 2b1 サ変 b 2b2 サ変 c 2b3 サ変 d 2b4 サ変 e 2b5 サ変 f
3: 形容詞	300 形容詞一般
4: 形容動詞	400 形動幹 a 410 形動幹 b
5: 副詞	500 副詞一般
6: 連体詞	600 連体詞
7: 接続詞・感動詞	700 接続詞 710 感動詞
8: 接頭語	800 接頭語一般
9: 接尾語	900 接尾語一般 910 助数詞
a: ひらがな・カタカナ未知語	a**
d: 記号	d00 一般記号
e: 連用	e** 名詞化連用形
f: サ変	f** サ変化連用形

詳しくは、韻律処理マニュアル [1] 参照。

3.7.3 mor2bou への入力

TDMT 側の情報から mor2bou への入力の際、以下の条件での変換を行う。
変換後、 1 音節内に 2 つ以上のアクセント情報及び読みがあった際は 1 つのアクセント及び読みに変換する。
但し、アクセント情報 0 の組合せは除く。

- 読点情報の挿入
  先行品詞が終助詞＋続く品詞が特殊でない場合は、 [、 0 0 0 d000001000000001] を挿入する。

- 促音処理
  先行読み情報が「っ」で終了している場合は、 続く音節を促音の付属語扱いとする。

例)TDMT 出力

```
待っ | 本動詞 | まっ |2350001000000011
て下さい | 助動詞 | てください |0
```

例) 変換後 mor2bou 入力

```
待って下さい マッ テクダサイ 0 2350001000000011
```

- 日付処理
  先行アクセント情報の先頭が 120(数詞)+ 続く読みが「にち」で始まる場合は、下記の様に変換を行う。

例)TDMT 出力

```
十 | 数詞 |10|1200001000000010
日 | 接尾辞 | にち |9000001000000000
```

例) 変換後 mor2bou 入力

```
十日の トウカ 0 0 9000001000000000
```

```
 2 + にち ----- 「ふつか」  3 + にち ----- 「みっか」
 4 + にち ----- 「よっか」  5 + にち ----- 「いつか」
 6 + にち ----- 「むいか」  7 + にち ----- 「なぬか」
 8 + にち ----- 「ようか」  9 + にち ----- 「ここのか」
10 + にち ----- 「とうか」 20 + にち ----- 「はつか」
```

- 音節の統合
  1. 先行アクセント情報が 0(終助詞ではない)+ 続くアクセント情報が 0(終助詞含む) の場合
  2. 先行アクセント情報の 1 バイト目が 1(名詞)+ 続くアクセント情報が 0(終助詞含む) の場合
  3. 先行アクセント情報の 1 バイト目が 9(接尾辞)+ 続くアクセント情報が 0(終助詞含む) の場合

  上記いずれかの場合、2 つの音節を 1 音節に統合にする。

例 1)TDMT 出力

```
て頂け | 助動詞 | ていただけ |0
ます | 助動詞 | ます |0
か | 終助詞 | か |0
```

例 1) 変換後 mor2bou 入力

```
て頂けますか テイタダケマスカ 0 0 0000000000000000
```

例 2)TDMT 出力

```
時間 | 普通名詞 | じかん |1000001000000000
の | 連体助詞 | の |0
```

例 2) 変換後 mor2bou 入力

```
時間の ジカン ノ 0 1000001000000000
```

例 3)TDMT 出力

```
十五 | 数詞 |15|1200001000000011
分 | 接尾辞 | ふん |9100001000000020
で | 助動詞 | で |0
```

例 3) 変換後 mor2bou 入力

```
十五 15 0 0 1200001000000011
分で フン デ 0 9100001000000020
```

- 中点処理
  表層が「・」の場合、読点「、」扱いとする。

- 読点処理
  表層が「、」の場合、[、 0 0 0 d000001000000001] とする。

- 句点処理
  表層が「。」の場合、[。 0 0 0 0000001000000001] とする。

- 読み＋アクセント情報の統合 1 つの文節内にアクセント情報 0 が続く場合、またはアクセント情報 1 が続く場合
  アクセント情報の統合を行う。

3.7.4 kan2rom

TDMT 側から読み情報の欠落などの不備があった際に kan2rom へ処理を移行する。
TDMT 側の情報から表層部分を取り出し /tmp/ch_k2r.txt を作成 kan2rom.gfilter の結果を gfilter の出力とする。
※ shell/kan2rom.gfilter を参照して下さい。

3.7.5 無声化フィルタの追加

mou2bou/kan2rom の結果に日本語無声化処理を加える。
日本語無声化フィルタは、gfilter から popen 形式で読み書きしています。

ファイル名	内容	備考
UNVOICE.pm	肝心の「無声化フィルタ」	Perl のパッケージにしてある
unvoice	UNVOICE.pm のドライバ部	Perl で記述
LOGGING.pm	中間的なログの出力機能	Perl のパッケージにしてある

注意) Perl (Ver.5.003) のパスが通っていることを確認してください。

3.7.6 無声化フィルタの実行例

mou2bou 出力結果

```
sui'itowa 1 ippaku' 1 nanahyaku'dorudesu 1 tsuiNru'umuwa 2 ippaku' 1 saNbyaku'dorudesu 5
```

無声化フィルタ実行結果

```
sui'itowa 1 ippaku' 1 nanahyaku'dorudesu 1 tsuiNru'umuwa 2 ippaku' 1 saNbyaku'dorudesU 5
```

3.8 言語フィルターの多言語処理

3.8.1 言語フィルターの英語処理

TDMT 側の情報から表層部分を取り出し下記の処理の後 gfilter の出力とする。

言語フィルターへの入力例

```
EMI'd|PRON||
like|PREP||
to|PREP||
have|V||
a|DET||
reservation|CN||
for|PREP||
the|DET||
hotel|CN||
```

言語フィルターからの出力例

```
I'd like to have a reservation for the hotel
```

the(The) に続く音素が次の音素の場合に限り thi(Thi) と処理する。
'a' 'e' 'i' 'o' 'u' 'A' 'E' 'I' 'O' 'U'

3.8.2 言語フィルターの韓国語処理

TDMT 側の情報から表層部分を取り出し /tmp/hcode.txt を作成、 hcode の結果を gfilter の出力とする。

言語フィルターへの入力例

```
KM호텔|NCOMM||
예약|NACTV||
부탁|NACTV||
하|XSUFFVERB||
고 싶|VAUX1||
은데|VAUX2TEND||
요|VAUX2SEND||
```

言語フィルターからの出力例

```
hoteLyeyaGbutaGhago siPUNdeyo
```

3.8.3 言語フィルターの中国語処理

TDMT 側の情報から読み部分を取り出し gfilter の出力とする。

言語フィルターへの入力例

```
CM您|先頭|ning2|
好|叹|hao3|
，|符|3|
这里|地|zhe4 li3|
是|判|shi4|
半岛酒店|地|ban4 dao3 fan4 dian4|
。|符|5|
```

言語フィルターからの出力例

```
ning2 hao3 3 zhe4 li3 shi4 ban4 dao3 fan4 dian4 5
```

3.8.4 言語フィルターのドイツ語処理

TDMT 側からの情報から、 CHATR の Utterance PhonoForm 形式の表記を作成する。
詳細は、最後に添付した資料を参考にして下さい。

言語フィルターへの入力例

```
GMes|SATZ:SUBJECT:PERSONALPRONOMEN|@.s.|0
gibt|SATZ:V-FIN:VERB|g.I.b.t.|1
eine apotheke|SATZ:OBJECT:DETERMINATIV-INDEF NOMEN|aI.-n.@. a.-p.o:.-t.e:.-k.@.|10 0010
nahe|SATZ:F-MIDDLE:ADJEKTIV|n.a:.-@.|10
.|SATZ:BOUNDARY-END:INTERPUNKTION|4|0
sie|SATZ:SUBJECT:PERSONALPRONOMEN|z.i:.|1
ist|SATZ:V-FIN:HILFSVERB|I.s.t.|1
neun uhr|SATZ:TIME:KARDINALZAHL NOMEN|n.OY.n. u:.6.|1 1
geoeffnet|SATZ:V-INF:VERB|g.@.-9.f.-n.@.t.|010
.|SATZ:BOUNDARY-END:INTERPUNKTION|4|0
```

言語フィルターからの出力例

```
:D nil (:S nil (:C nil (:P nil
   (Word ich nil
     (Syl I.C. ()
       (Phoneme I 0 0 ((0 0)))
       (Phoneme C 0 0 ((0 0)))))
   (Word moechte nil
     (Syl m.9.C. ()
       (Phoneme m 0 0 ((0 0)))
       (Phoneme 9 0 0 ((0 0)))
       (Phoneme C 0 0 ((0 0))))
     (Syl t.@. ()
       (Phoneme t 0 0 ((0 0)))
       (Phoneme @ 0 0 ((0 0)))))
   (Word die nil
     (Syl d.i:. ()
       (Phoneme d 0 0 ((0 0)))
       (Phoneme i: 0 0 ((0 0)))))
   (Word reservierung nil
     (Syl r.e:. ()
       (Phoneme r 0 0 ((0 0)))
       (Phoneme e: 0 0 ((0 0))))
     (Syl z.E6. ()
       (Phoneme z 0 0 ((0 0)))
       (Phoneme E6 0 0 ((0 0))))
     (Syl v.i:. ((Stress 1) (Intones H*))
       (Phoneme v 0 0 ((0 0)))
       (Phoneme +i: 0 0 ((0 0))))
     (Syl r.U.N. ((Intones L-))
       (Phoneme r 0 0 ((0 0)))
       (Phoneme U 0 0 ((0 0)))
       (Phoneme N 0 0 ((0 0))))))
 (:P nil
   (Word fuer nil
     (Syl f.y:6. ()
       (Phoneme f 0 0 ((0 0)))
       (Phoneme y:6 0 0 ((0 0)))))
   (Word das nil
     (Syl d.a:.s. ()
       (Phoneme d 0 0 ((0 0)))
       (Phoneme a: 0 0 ((0 0)))
       (Phoneme s 0 0 ((0 0)))))
   (Word hotel nil
     (Syl h.o:. ()
       (Phoneme h 0 0 ((0 0)))
       (Phoneme o: 0 0 ((0 0))))
     (Syl t.E.l. ((Stress 1) (Intones H+L* L-L%%))
       (Phoneme t 0 0 ((0 0)))
       (Phoneme +E 0 0 ((0 0)))
       (Phoneme l 0 0 ((0 0)))
       (Phoneme # 0 0 ((0 0))))))))
```

3.9 言語フィルターのコンパイル

3.9.1 ディレクトリ構成

gfilter

```
gfilter -+-- src     ------- gfilter_test.c       (gfilter 単体テスト用 main)
         |                   gfilter.c            (gfilter)(english filter)(korean filter)
         |                   tdmt_in.c            (tdmt+mor2bou data input)
         |                   tdmt_mor2bou.c       (tdmt+mor2bou mor2bou.c)
         |                   tdmt_inritu.c        (tdmt+mor2bou inritu.c)
         |                   tdmt_tougokyoukai.c  (tdmt+mor2bou tougokyoukai.c)
         |                   tdmt_hfr.c           (tdmt+mor2bou hfr.c)
         |                   zenkaku2hankaku.c
         |                   chinese_filter.c
         |                   m2b_ver3.c           (german filter)
         |                   tdmt_in.h
         |                   type.h
         |                   zenkaku2hankaku.h
         |
         |                   makefile
         |                   main.c               (hcode main)
         |                   codeconv.c           (hcode Code Conversion)
         |                   hantable.h
         |
         |                   unvoice              (無声化フィルタドライバ部)
         |                   UNVOICE.pm           (無声化フィルタ)
         |                   LOGGING.pm           (無声化フィルタログ出力部)
         |
         |
         +-- morph           morph, hmorph, mor2bou
         |
         +-- dic     ------- Japanese Dictionary
         |
         +-- doc     ------- gfilter.tex  (gfilter document)
         |                   korean_txt.ps
         |
         +-- lib     ------- libgfilter.a (gfilter library)
         |
         +-- shell   ------- kan2rom.gfilter
         |                   makelib_gfilter
         +-- README
         |
         +-- 1998.04.16.html
```

3.9.2 ソースのチェックアウト

- gfilter
  setenv CVSROOT /DB/PI/CVSROOT
  cvs checkout gfilter

- morph
  setenv CVSROOT /DB/PI/CVSROOT
  cd gfilter
  cvs checkout morph

- unvoice
  setenv CVSROOT /DB/PI/CVSROOT
  cvs checkout unvoice

3.9.3 morph のコンパイル

gfilter をコンパイルする前に mor2bou のコンパイルが必要です。

- cd gfilter/morph
- makemorph.sh

注 1) OS によって morph/PROSODY/makefile.option のオプションのチェックが必要です
注 2) 付属語エラーを解消するには morph/PROSODY/srh_ftbl/srh_ftbl.c に IGNORE_FUZOKUGO_ERROR 宣言が必要です。

3.9.4 gfilter のコンパイル

- cd gfilter/src
- make

3.10 言語フィルターの起動方法

3.10.1 gfilter 実行

使用方法	gfilter infile .....
オプション	infile TDMT からのデータファイル (複数指定可)

3.10.2 言語フィルターの動作環境

machine:	itlpc96
OS :	Linux 2.0.30.
make :	/usr/local/gnu/bin/make
gcc :	/usr/local/gnu/gcc-2.7.2.1/bin/gcc

4 問題点

「Chat Translationで発生している問題」への2研の対処状況
1998/4/16での最新版は、「1998.04.16.html」としてCVS管理下に置きます。
「MATRIXで発生している問題」への2研の対処状況
1998/09/31での最新版は、「investigate+fix.shtml」としてCVS管理下に置きます。

尚、最新情報は `http://www.itl.atr.co.jp/~kasimura/chatr/chat_trans`
`http://itlpc60/~kasimura/MATRIX/problem/latest/index.html`
に随時更新されます。

4.1 英数字対応

TDMTで読めない英数字は、そのままgfilterへと受け渡される。
但し、日本語に関しては英数字の読み情報は付加せず、緊急処置としてのkan2romで全文処理を行う。

言語	英数字処理	エラー	備考
日本語	kan2rom処理	「1998.04.16.html」参照	英数字の読み情報なし
英 語	CHATR Text入力	「1998.04.16.html」参照	
ドイツ語	CHATR PhonoForm入力		
韓国語	無視		英数字はhcodeで読めない

注意）パンクチュエーションなどの記号の削除について。
英語・ドイツ語の場合のみ、暫定的に`gfilter`内で以下の記号は削除しています。
`( ) [ ] { }` の 6種類

5 言語フィルターの改善

TDMT と CHATR のインターフェイスをより良くする為に以下のような改善を行った。

- TDMT 辞書における形態素区切りの変更
  今までの辞書に較べてより KDD morph の形態素区切りに近いものに変更したため，アクセント推定の精度が高い
- 形態素区切りの変更に伴う gfilter ソースの修正
  新しい形態素区切りに対応するための修正
- 付属語検出エラーの修正
  mor2bou ではもともと付属語接続がおかしいときに，その後の出力をやめてしまう不具合がありました．gfilter では mor2bou で付属語の接続がおかしいときには Fuzokugo detection error ... recovered と表示して読みだけでも出力する機能をいれていましたが，サ変名詞のときに付属語検出が正しく行われないバグがあり，gfilter をコンパイルするときのみ -DIGNORE_FUZOKUGO_ERROR として mor2bou をコンパイルしていました．酒井氏によりこのバグが修正されたので，-DIGNORE_FUZOKUGO_ERROR のコンパイルオプションは廃止しました．

5.1 言語フィルターの改善に伴う 形態素区切りの変更点

言語フィルターの改善に伴う形態素区切りの変更ポイントを下記に説明します。

5.1.1 TDMT 日本語辞書の種類

TDMT の辞書には形態素解析辞書と生成辞書の 2 種類がある。
形態素解析辞書のエントリで生成辞書に含まれないものは、差分辞書に記述されている。
生成辞書と差分辞書を修正対象として修正を行った。

TDMT 生成辞書	
ファイル名	内容
dohsi.dic	TDMT 側の品詞で本動詞のもの
keiyou.dic	TDMT 側の品詞で形容詞、形容名詞のもの
fukushi.dic	TDMT 側の品詞で副詞、感動詞、接続詞のもの
jodoshi.dic	TDMT 側の品詞で助動詞のもの
meishi.dic	TDMT 側の品詞で固有名詞、普通名詞、日時のもの
rentaishi.dic	TDMT 側の品詞で連体詞のもの
setsuji.dic	TDMT 側の品詞で接頭辞、接尾辞のもの
joshi.dic	TDMT 側の品詞で助詞一般のもの
TDMT 差分辞書	
ファイル名	内容
jg.dic	形態素解析辞書に登録されていて、生成辞書に登録されていないもの

5.1.2 差分辞書の修正

- 読みが誤っているもの

修正例

（修正前）サ変名詞 "直行" 読み "なおゆき" アクセント "1120001000000021"
（修正後）サ変名詞 "直行" 読み "ちょっこう" アクセント "1002421000000000"

- 区切りが誤っているもの

修正例

（修正前）普通名詞 "アンチェインドメロディー" 読み "あん" アクセント "1120001000000010" 読み "ちぇ" アクセント "a000001000000010" 読み "いんど" アクセント "1130001000000010" 読み "めろでぃー" アクセント "1000001000000010"
（修正後）普通名詞 "アンチェインドメロディー" 読み "あんちぇいんどめろでぃー" アクセント "1100001000000070"

- アクセント位置が正しくないもの

修正例

（修正前）普通名詞 "てんどん" 読み "てんどん" アクセント "1000001000000031"
（修正後）普通名詞 "てんどん" 読み "てんどん" アクセント "1000001000000001"

- 未知語のもの

修正例

（修正前）普通名詞 "にぎり寿司" 読み "にぎり" アクセント "a000001000000010" 読み "すし" アクセント "1000101000110020"
（修正後）普通名詞 "にぎり寿司" 読み "にぎりずし" アクセント "1000001000000030"

- 結合し、長い単語にしたもの

修正例

（修正前）普通名詞 "クリアランスセール" 読み "くりあ" アクセント "4002431000000020" 読み "らんす" アクセント "a000001000000010" 読み "せーる" アクセント "1000001000000010"
（修正後）普通名詞 "クリアランスセール" 読み "くりあらんすせーる" アクセント "1000001000000070"

5.1.3 生成辞書の修正

- doshi.dic の修正

修正例

（修正前）本動詞 "見せる" 読み "みせる" アクセント "2900001000000011"
（修正後）本動詞 "見せる" 読み "みせ" アクセント "2900001000000011" 読み "る" アクセント "0"

- keiyou.dic の修正

修正例

（修正前）形容詞 "美味しい" 読み "おいしい" アクセント "3000001000000001"
（修正後）形容詞 "美味しい" 読み "おいし" アクセント "3000001000000031" 読み "い" アクセント "0"

- jg.dic の修正

修正例

（修正前）形容詞 "なまぬるい" 読み "なまぬるい" アクセント "3000001000000041"
（修正後）形容詞 "なまぬるい" 読み "なまぬる" アクセント "3000001000000041" 読み "い" アクセント "0"

- jodoshi.dic の修正

修正例

（修正前）助動詞 "願う" 読み "ねがう" アクセント "2800001000000021" 読み "う" アクセント "0"
（修正後）助動詞 "願う" 読み "ねが" アクセント "2800001000000021" 読み "う" アクセント "0"

5.2 言語フィルターの改善に伴う gfilter の変更点

形態素区切りの変更に伴う gfilter のソースの修正ポイントを下記に説明します。

5.2.1 付属語処理の追加

- 先行の品詞が動詞 (アクセント情報の 1 バイト目が 2) であり、
  続くエントリの品詞が助詞 (アクセント情報が 0) であれば、
  助詞を動詞の付属語として結合させる。
  動詞に付属語がなければ、その助詞を付属語とする。

例 1) TDMT 出力

```
開い | 本動詞 | あ い |2000001000000001 0
てい | 助動詞 | てい |0
ます | 助動詞 | ます |0
```

例 1) 改良前 gfilter

mor2bou 入力

```
開い ア イ 0 2000001000000001
ています テイマス 0 0 0000000000000000
```

mor2bou 出力

```
ai 1 teemasU
```

例 1) 改良後 gfilter

mor2bou 入力

```
開いています ア イテイマス 0 2000001000000001
```

mor2bou 出力

```
aite imasu
```

例 2) TDMT 出力

```
見 | 本動詞 | み |2900001000000011
まし | 助動詞 | まし |0
た | 助動詞 | た |0
か | 終助詞 | か |0
```

例 2) 改良前 gfilter

mor2bou 入力

```
見 ミ 0 0 2900001000000011
ましたか マシタカ 0 0 0000000000000000
```

gfilter 改良前 mor2bou 出力

```
mi' 2 mashItaka
```

例 2) 改良後 gfilter

mor2bou 入力

```
見ましたか ミ マシタカ 0 2900001000000011
```

mor2bou 出力

```
mima'shitaka
```

- 先行の品詞が形容詞 (アクセント情報の 1 バイト目が 3) であり、
  続くエントリの品詞が助詞 (アクセント情報が 0) であれば
  動詞の場合と同様に付属語化させる。

TDMT 出力

```
嬉しい | 形容詞 | うれし い |3000001000000021 0
です | 助動詞 | です |0
```

改良前 gfilter

mor2bou 入力

```
嬉しい ウレシ イ 0 3000001000000021
です デス 0 0 0000000000000000
```

mor2bou 出力

```
ureshi'i desu
```

改良後 gfilter

mor2bou 入力

```
嬉しいです ウレシ イデス 0 3000001000000021
```

mor2bou 出力

```
ureshi'idesu
```

- 先行の品詞が形容動詞 (アクセント情報の 1 バイト目が 4) であり、
  続くエントリの品詞が助詞 (アクセント情報が 0) であれば
  動詞の場合と同様に付属語化させる。

TDMT 出力

```
確か | 形容名詞 | たしか |4000001000000011
です | 助動詞 | です |0
```

改良前 gfilter

mor2bou 入力

```
確か タシカ 0 0 4000001000000011
です デス 0 0 0000000000000000
```

mor2bou 出力

```
tashIkadesU
```

改良後 gfilter

mor2bou 入力

```
確かです タシカ デス 0 4000001000000011
```

mor2bou 出力

```
ta'shikadesu 5
```

- 語幹の無いカ変動詞 (アクセント情報の先頭 2 バイトが 2a) の場合
  自立語と付属語を結合させたものを付属語とする。

TDMT 出力

```
来 | 本動詞 | き |2a00001000000011
てい | 助動詞 | てい |0
まし | 助動詞 | まし |0
た | 助動詞 | た |0
```

改良前 gfilter

mor2bou 入力

```
来 キ 0 0 2a00001000000011
ていました テイマシタ 0 0 0000000000000000
```

mor2bou 出力

```
ki' 1 teemashIta
```

改良後 gfilter

mor2bou 入力

```
来ていました キ キテイマシタ 0 2a00001000000011
```

mor2bou 出力

```
ki'te imashita
```

- 語幹の無いサ変動詞 (アクセント情報の先頭 3 バイトが 2b3) の場合
  自立語と付属語を結合させたものを付属語とする。

TDMT 出力

```
し | 本動詞 | し |2b30001000000001
ます | 助動詞 | ます |0
```

改良前 gfilter

mor2bou 入力

```
し シ 0 0 2b30001000000001
ます マス 0 0 0000000000000000
```

mor2bou 出力

```
shi 1 masU 5
```

改良後 gfilter

mor2bou 入力

```
します シ シマス 0 2b30001000000001
```

mor2bou 出力

```
shima'su 5
```

5.2.2 入力エントリの分解

- エントリ１行に助詞 (アクセント情報が 0) を除いた品詞が複数混在している場合
  ２つ目以降を次のエントリとして分解し２行にする。

※ 表層部分の分解は困難である為 mor2bou での韻律処理に影響が無い事から分解後も同じ表層とする。

TDMT 出力

```
レストランに関する | 格助詞 | れすとらん に かん する |1000001000000010 0
 2b10001000000010 0
```

改良前 gfilter

mor2bou 入力

```
レストランに関する レストランニカンスル 0 0 1000001000000010
```

mor2bou 出力

```
resUtoraNnikaNsuru
```

改良後 gfilter

mor2bou 入力

```
レストランに関する レストラン ニ 0 1000001000000010
レストランに関する カン スル 0 2b10001000000010
```

mor2bou 出力

```
re'sutoraNni kaNsu'ru
```

5.2.3 ＢＩの変更

BI(Break Index) を伴わない空白区切りが存在した場合、強制的に BI 1 を挿入するように変更された。但し、BI が余分に出力される場合は、最も深い統語境界をもつ BI を出力する。

5.2.4 無声化処理の切り替え

無声化処理を行うには、gfilter から Perl 版無声化フィルタを使用していたが
C版 無声化フィルタの mor2bou への実装に伴い、これを削除した。

5.3 言語フィルターの改善で修正したファイル

5.3.1 TDMT 辞書

gfilter/dic/japanese/dic/jg.dic [1.10]
gfilter/dic/japanese/dic/doshi.dic [1.5]
gfilter/dic/japanese/dic/keiyou.dic [1.2]
gfilter/dic/japanese/dic/jodoshi.dic [1.2]
gfilter/dic/japanese/dic/setsuji.dic [1.4]
gfilter/dic/japanese/dic/fukushi.dic [1.4]
gfilter/dic/japanese/dic/meishi.dic [1.2]

5.3.2 gfilter

gfilter/src/tdmt_in.c [1.4]

5.3.3 リカバリ機能

morph/PROSODY/cal_accbou/inritu.c [1.8]
morph/PROSODY/srh_ftbl/srh_ftbl.c [1.8]

5.3.4 関数の追加

関数名	戻り値	ファイル名	備考
bunkai	int: 分解した個数	tdmt_in.c	gfilter
presearchftbl()	int: 最長一致文字列の全角文字数 int: 一つもつがらない時:0 int: 一つもつがらない:-1	srh_ftbl.c	mor2bou [1]

詳しくは、TR-IT-0290「言語フィルターの改善」[3] を参照。

6 APPENDIX

6.1 TDMT データサンプル

/home/as53/simomura/dept3/chat_translation_sample_data

```
---+-- J --- ej001.sjis ~ ej161.sjis          (Japanease)
   |         kj001.sjis ~ kj028.sjis
   +-- E --- je001.internal ~ je165.internal  (English)
   +-- G --- jg001.internal ~ jg045.internal  (German)
   +-- K --- jk001.internal ~ jk050.internal  (Korean)
```

/home/as53/simomura/dept3/text

```
text  ---+-- TAS12008 --- TAS12008_01 ~ TAS12008_24 (China)
         +-- TAS12010 --- TAS12010_01 ~ TAS12010_25 (China)
         +-- TAS13005 --- TAS13005_01 ~ TAS13005_18 (China)
         +-- TAS13009 --- TAS13009_01 ~ TAS13009_40 (China)
         +-- TAS22001 --- TAS22001_01 ~ TAS22001_30 (China)
         +-- TAS23001 --- TAS23001_01 ~ TAS23001_37 (China)
         +-- TAS32002 --- TAS32002_01 ~ TAS32002_33 (China)
         +-- TAS33001 --- TAS33001_01 ~ TAS33001_41 (China)
         +-- TAS33011 --- TAS33011_01 ~ TAS33011_41 (China)
```

6.2 言語フィルターの改善で使用した TDMT データサンプル

/home/as67/xssakai/simura2/Ssample

```
---+-- J --- ej001 ~ ej161          (Japanease)
```

6.3 gfilter 関数

gfilter(char *TDMT からのデータ, char *CHATR への出力データ)

char *CHATR への出力データのサイズに関しては、入力されるデータサイズをもとに確保しておいて下さい。
特にドイツ語処理の場合は、品詞の数だけ詳細な情報が出力されますので注意が必要です。
現在 server 側では、 50000 で定義しているようです。
尚、サイズ確保が十分で無い場合の gfilter の動作は保証されておりません。

※ gfilter のライブラリ化は、 shell/make_gfilter を参照して下さい。

shell/make_gfilter

```
ar rv lib/libgfilter.a  \
                        morph/PROSODY/cal_accbou/a2k.o \
                        morph/PROSODY/cal_accbou/accent_alpha.o \
                        morph/PROSODY/cal_accbou/accent_number.o \
                        morph/PROSODY/cal_accbou/calaccent.o \
                        morph/PROSODY/cal_accbou/conv.o \
                        morph/PROSODY/cal_accbou/getsuffixtype.o \
                        morph/PROSODY/cal_accbou/inritu.o \
                        morph/PROSODY/cal_accbou/num_kana.o \
                        morph/PROSODY/cal_accbou/strdup.o \
                        morph/PROSODY/cal_accbou/tougokyoukai.o \
                        morph/PROSODY/srh_ftbl/srh_ftbl.o \
                        morph/PROSODY/srh_ftbl/ftbl_lib.o \
                        src/gfilter.o \
                        src/tdmt_in.o \
                        src/tdmt_hfr.o \
                        src/tdmt_mor2bou.o \
                        src/tdmt_inritu.o \
                        src/tdmt_tougokyoukai.o \
                        src/zenkaku2hankaku.o \
                        src/m2b_ver3.o \
                        src/chinese_filter.o

ranlib lib/libgfilter.a
```

6.4 プログラム関数一覧

関数名	戻り値	ファイル名	備考
gfilter	void	gfilter.c	言語フィルター
japanese_filter	void	gfilter.c	日本語フィルター
mor2bou	int: 正常:0 異常:-1	tdmt_mor2bou.c	gfilter
tdmt_accent_kyoukai	int: 正常:0 異常:-1	tdmt_inritu.c	gfilter
tdmt_input	int: 正常:0 異常:-1	tdmt_in.c	gfilter
datasize_check	int: 入力データ行数	tdmt_in.c	gfilter
memory_alloc	int: 正常:0 異常:-1	tdmt_in.c	gfilter
set_tdmtdata	int: 異常:-1 状況コード:0-9	tdmt_in.c	gfilter
set_yomiaccent	int: 異常:-1 それ以外: 件数	tdmt_in.c	gfilter
mor2bou_in	int: 正常:0 異常:-1	tdmt_in.c	gfilter
zen2han	int: 異常:-1 それ以外: 件数	zenkaku2hankaku.c	gfilter
trans_zen2han	int: 正常:0 異常:-1	zenkaku2hankaku.c	gfilter
mor2bou_text_in	int	inritu.c	mor2bou [1]
tdmt_mor2bou_main	int: 正常:0	tdmt_inritu.c	gfilter
irt_st	int	inritu.c	mor2bou [1]
rendaku	int	calaccent.c	mor2bou [1]
num_alp_yomi	int	num_kana.c	mor2bou [1]
kyoukai_tokusyu	int	tougokyoukai.c	mor2bou [1]
syurui_yomi	int	num_kana.c	mor2bou [1]
fkg_acc	int	calaccent.c	mor2bou [1]
acc_fz	int	calaccent.c	mor2bou [1]
acc_mark	int	calaccent.c	mor2bou [1]
tdmt_kyokai_write	int: 正常:0	tdmt_tougokyoukai.c	mor2bou [1]
mor2bou_text_out	int: 正常:0	tdmt_mor2bou.c	mor2bou [1]
conv	int	conv.c	mor2bou [1]
hfr	void	tdmt_hfr.c	mor2bou [1]
special_write_outbuf	int	tdmt_hfr.c	mor2bou [1]
write_outbuf	void	tdmt_hfr.c	mor2bou [1]
kyokai_bunrui	int	tougokyoukai.c	mor2bou [1]
chk_cmpw	int	tougokyoukai.c	mor2bou [1]
kf_cat	int	tougokyoukai.c	mor2bou [1]
mor2bou_init_free	int	inritu.	mor2bou [1]

関数名	戻り値	ファイル名	備考
english_filter	void	gfilter.c	英語フィルター
E_vowel	int: YES:1 NO:0	gfilter.c	
german_filter	void	m2b_ver3.c	ドイツ語フィルター
make_phonoform	int: 正常:0	m2b_ver3.c	
put_wrd	int: 処理中のテキスト文字数	m2b_ver3.c	
put_syl	int: 処理中のテキスト文字数	m2b_ver3.c	
put_ph	int: 処理中のテキスト文字数	m2b_ver3.c	
stressable	int: 品詞情報のチェック YES:1 NO:0	m2b_ver3.c	
vowel	int: YES:1 NO:0	m2b_ver3.c	
pos	int: INTERPUNKTION のチェック YES:-1 NO:0	m2b_ver3.c	
skip_syl	int: skip 処理	m2b_ver3.c	
set_tdmtdata_german		m2b_ver3.c	
korean_filter	void	gfilter.c	韓国語フィルター
chinese_filter	void	gfilter.c	中国語フィルター

6.5 ドイツ語 PhonoForm 処理参考資料

```
From: Michael Paul <paul@itl.atr.co.jp>

we added additionional linguistic information to the G-FILTER input
for German, which (hopefully) can be used in the demo.
The main format (4 fields with | separator) didn't change, but
we split the second field into 3 sub-fields
(sentence-type, topological field type, part-of-speeches).

You can find the examples in the modified format in:

~doi/work/test/IJCAI/G-FILTER/G/

The format of the G-FILTER input is as follows:

FORMAT: 4 fields per line seperated by '|'

FIELD1: surface words
        -> list of words seperated by spaces

FIELD2: linguistic information
        -> consists of 3 sub-fields (seperator ':')
           + SUB-FIELD1: sentence-type
           + SUB-FIELD2: topological field of sentence structure
           + SUB-FIELD3: list of part-of-speeches seperated by spaces
                         (cooresponding to parts of FIELD1)

           => see below detailed description of the content of the SUB-FIELDS

FIELD3: syllabified phonetic information (syllable marker '-')
        -> list of syllabified phones seperated by spaces
          (cooresponding to parts of FIELD1)

FIELD4: stress pattern information (0-no stress, 1-stress)
        -> list of stress pattern seperated by space
          (cooresponding to parts of FIELD1)

***-------------------------------------------------------------------------***
*** SUB-FIELD1: sentence types
***-------------------------------------------------------------------------***

  ;------------------------- main clauses ------------------------------;

+ SATZ                  - Aussagesatz (平叙文  ヘイジョブン)

                        + V-pos:        verb-second
                        + V-mood:       INDIKATIV / KONJUNKTIV
                                        (never: IMPERATIV !)
                        + intonation:   end of sentence: down
                        + punctuation:  "."
```

```
+ SATZ+                  - Aussagesatz (平叙文  ヘイジョブン)

                         + V-pos:        verb-second
                         + V-mood:       INDIKATIV / KONJUNKTIV
                                         (never: IMPERATIV !)
                         + intonation:   end of sentence: down
                         + punctuation:  ""

+ SATZ+verb-first        - Aussagesatz with preceding subord-clause

                         + V-pos:        verb-first
                         + V-mood:       INDIKATIV / KONJUNKTIV
                                         (never: IMPERATIV !)
                         + intonation:   end of sentence: down
                         + punctuation:  "."
                         + others:       - preceding subord before main-clause

+ FRAGE mit WH           - Ergaenzungsfrage (捕捉質問  ホソクシツモン)
  [WH-Q]
                         + V-pos:        verb-second; somtimes: verb-final
                                         (never: verb-first !)
                         + V-mood:       INDIKATIV / KONJUNKTIV
                                         (never: IMPERATIV !)
                         + intonation:   end of sentence: as well down, as up
                         + punctuation:  "?"
                         + others:       - characteristic particle
                                           [bloss,wohl,denn,eigentlich]
                                         - intro: WHPRON / WHADV
                                                  [wann,wo]

+ FRAGE ohne WH          - Entscheidungsfragesatz (決定疑問  ケッテイギモン)
  [YN-Q]
                         + V-pos:        verb-first
                         + V-mood:       INDIKATIV / KONJUNKTIV
                                         (never: IMPERATIV !)
                         + intonation:   end of sentence: up
                         + punctuation:  "?"
                         + others:       - characteristic particle
                                           [etwa,mal,wohl,eigentlich,auch,denn]

+ FRAGE_verb-first       - question with preceding subord-clause

+ IMP-S-I                - Aufforderungssatz (命令文  メイレイブン)

                         + V-pos:        verb-first
                         + V-mood:       IMPERATIV
                         + intonation:   end of sentence: down
                         + punctuation:  "."
                         + others:       - characteristic particle
                                           [doch,bloss,eben,einfach,halt,ja,
                                            ruhig,schon]
```

```
+ IMP-S-II              - Aufforderungssatz (命令文  メイレイブン)

                        + V-pos:        verb-first
                        + V-mood:       IMPERATIV
                        + intonation:   end of sentence: down
                        + punctuation:  "!"
                        + others:       - characteristic particle
                                          [doch,bloss,eben,einfach,halt,ja,
                                           ruhig,schon]

+ DES-S-I               - Wunschsatz Typ I (願望文  ガンボウブン)

                        + V-pos:        verb-second (person: 3)
                        + V-mood:       KONJUNKTIV I
                        + intonation:   end of sentence: down
                        + punctuation:  "." / "!"

+ DES-S-II              - Wunschsatz Typ II (願望文  ガンボウブン)

                        + V-pos:        verb-first / verb-final
                        + V-mood:       KONJUNKTIV II
                        + intonation:   end of sentence: down
                        + punctuation:  "!"
                        + others:       - characteristic particle
                                          [bloss,nur,doch]

  ;---------------- subordinate clauses (副詞  フクシ) ------------------;

+ REL-S                 - Relativsatz (関係文  カンケイブン)

                        + intro:        - :RELATIV- / :INTERROGATIVPRONOMEN
                                          [der,die,das,welcher,welche,welches]
                                        - :ADVERB (Pronomialadverb)
                                          [wo,wie,wohin,wodurch]
                        + pronoun-pos:  - first:  [<REL-PRON> ...]
                                        - second: [<PREP> <REL-PRON> ...]
                        + others:       - near to reference word

+ KAU-S                 - Kausalsatz (原因文  ゲンインブン)
                        + intro:        [da,weil;weswegen,weshalb; zumal(da);
                                         dafuer_,_dass]

+ KONS-S                - Konsekutivsatz (結果文  ケッカブン)
                        + intro:        [dass,so dass, so...dass,um...zu,zu,
                                         als dass]

+ KONZ-S                - Konzessivsatz (認容文  ニンヨウブン)
                        + intro:        [obgleich,obwohl,obschon,wenn auch,
                                         wenngleich,wennschon]
```

```
+ TEM-S                  - Temporalsatz (時の副文  トキのフクブン)
                         + intro:
                           * simultan:  [waehrend,indem,indes,indessen,solange,
                                         sobald, sowie,sooft,als,wie,wenn]
                           * previous:  [nachdem,als,seit.seitdem,sobald,sowie,
                                         wenn]
                           * posterior: [bis,bevor,ehe]

+ KOND-S                 - Konditionalsatz (条件文  ジョウケンブン)
                         + intro: [wenn,wofern,sofern,falls]

+ FIN-S                  - Finalsatz (目的文  モクテキブン)
                         + intro:        [damit,dass,auf dass,um...zu]

+ MOD-S                  - Modalsatz (状況文  ジョウキョウブン)
                         + intro:        [indem,dardurch_,_dass,so...dass,]

+ IND-R                  - Indirekte Rede (間接話法  カンセツワホウ)
                           + V-mood:     KONJUNKTIV

+ INH-S - Inhaltssatz (内容文  ナイヨウブン)

          * INH-S-DASS  - intro: "dass"
                        - V-pos: verb-final
            [es hat mich gefreut, dass er kommt]

          * INH-S-ZU    - intro: -*-
                        - V-pos: INFINITIV with/without "zu" + verb-final
            [es hat mich gefreut, sie gesehen zu haben]

          * INH-S-WENN  - intro: "wenn" (not konditional / temporal)
                        - V-pos: verb-final
            [es wuerde mich freuen, wenn ich ihn sehen wuerde]

          * INH-S-ALS   - intro: "als" (not temporal,no compare,not "als ob")
                        - V-pos: verb-final
            [als er ploetzlich auftauchte, freute mich das]

          * INH-S-ALSOB - intro: "als"       / "als ob"
                        - V-pos: verb-second / verb-final
            [mir schien, als wuesste er nicht weiter]
            [mir schien, als ob er nicht weiter wuesste]

          * INH-S-WIE   - intro: "wie"
                        - V-pos: verb-final
            [ich merkte, wie mein Kraefte nachliessen]

          * INH-S-ANF   - intro: -*- (reported speech)
                        - V-pos: verb-second
            [er sagte, er sei selbst schuld]
```

```
* INH-S-OB     - intro: "ob"
               - V-pos: verb-final
  [er fragte, ob ich kaeme]

* INH-S-WH     - intro: WH conjunction (reported speech)
               - V-pos: verb-final
  [sie fragte, wann er kommt]

***-------------------------------------------------------------------------***
*** SUB-FIELD 2: topo-fields
***-------------------------------------------------------------------------***

INTRO          - introduction field (eg. relative pronoun in a relative clause)
F-PRE          - pre-field (coord/subordinated clauses preceding main clause)
F-MIDDLE       - middle field (parts without top-field specification, eg. NP)
F-POST         - post-field (coord/subordinated clauses succeeding main clause, punctuations)
V-FIN          - finite predicate part
V-INF          - infinite predicate part
VP-ADVERB      - adverbial non-verbal parts of predicate
VP-NON-VERB    - other non-verbal parts of predicate
REF-VERB-PRON  - reflexive pronoun
SUBJECT        - subject
IND-OBJECT     - dative object
OBJECT         - accusative object
TIME           - temporal expressions
PLACE          - locative expressions
NEGATION       - sentence negation
BOUNDARY       - punctuation
BOUNDARY-END   - punctuation (marking end of a (sub-)sentence)
BOUNDARY-START - punctuation (marking start of a (sub-)sentence)

***-------------------------------------------------------------------------***
*** SUB-FIELD 3: part-of-speeches
***-------------------------------------------------------------------------***

  + VERB                  - verb
  + HILFSVERB             - auxiliary verb
  + MODALVERB             - modal verb
  + VERBZUSATZ            - (separable) verb prefix
  + NOMEN                 - noun
  + ADJEKTIV              - adjective
  + ATTRIBUTIV-GEBRAUCHTES-PARTIZIP-PERFEKT  - attributive used participle perfect
  + ATTRIBUTIV-GEBRAUCHTES-PARTIZIP-PRAESENS - attributive used present participle
  + POSSESSIVPRONOMEN     - possessive pronoun
  + PERSONALPRONOMEN      - personal pronoun
  + REFLEXIVPRONOMEN      - reflexive pronoun
  + RELATIVPRONOMEN       - relative pronoun
  + INTERROGATIVPRONOMEN  - interrogative pronoun
  + FRAGEADVERB           - interrogative pronoun
  + DETERMINATIV          - determiner
  + DETERMINATIV-INDEF    - indefinite determiner
```

```
+ KARDINALZAHL          - cardinal number
+ ORDINALZAHL           - ordinal number
+ ADVERB                - adverb
+ PARTIKEL              - particle
+ PRAEPOSITION          - preposition
+ SUBORD-KONJUNKTION    - subord-conjunction
+ KOORD-KONJUNKTION     - coord-conjunction
+ STOP-WORD             - word without any morphological information (proper nouns)
+ INTERPUNKTION         - punctuation
+ BUCHSTABE             - letter
+ INITIAL               - inital
+ UNIT                  - unit
+ FIX-EXP               - fixed expression (no morph.analysis)
+ FIX-CAP               - capitalized fixed expression (eg. proper nouns)
+ FIX-UP                - up-cased fixed expression
+ FIX-INTRO             - fixed expression occurring at the sentence beginning
+ FIX-END               - fixed expression occurring at the sentence end
;-----------------------------------------------------------------------;
```

From: Kazuyuki Ashimura (1391) <kasimura@itl.atr.co.jp>

芦村です。

「チャット・トランンスレーション」のドイツ語出力に関する、
ポールさんのメールに補足します。
→この理解で間違ってたら、教えてください、ポールさん。

８／８最新版のドイツ語出力形式は、以下のとおりに変更されています。
ただし、読みから、「.」および「-」を除けば、今までのフォーマット
（単語｜品詞｜読み｜アクセント）として、g-filter で処理することが可能です。

＜ドイツ語新フォーマット＞

```
単語１ 単語２| 文タイプ: トポロジ: 品詞１ 品詞２| 読み１ 読み２| アクセント１ アクセント２
+---------+ +---------------------------+ +---------+ +---------------------+
  ↓           ↓                             ↓          ↓
単語フィールド  品詞フィールド              読みフィールド  アクセントフィールド
```

[凡例]

「|」は、フィールド区切り
「:」は、第二フィールド内での、サブフィールド区切り

※注意：

1. 第二フィールド（品詞）が、３つのサブフィールドに細分された。

番号	内容

1　　文タイプ（疑問文か、平叙文か）
2　　トポロジフィールド（エントリの文中での位置、意味）
3　　単語の品詞（複数あるときは、半角スペース区切り）
---

2. 第三フィールド（読み）に、音節および音素ごとの区切りが追加された。
・単語区切りは、「 」（半角スペース）
・音節区切りは、「-」
・音素区切りは、「.」

3. 第四フィールド（アクセント）については以下の通り。
・１単語を、１音節を１桁とする数字で表す。
・各桁の数字は、音節中のアクセント位置
（アクセントないときは「0」）。

例）
im reisefuehrer|REL-S:F-MIDDLE:PRAEPOSITION NOMEN|I.m. r.aI.-z.@.-f.y:.-r.@.r.|0 1000

```
第一フィールド  im reisefuehrer
                -- ------------
                ↓    ↓
               単語１ 単語２

第二フィールド  REL-S:F-MIDDLE:PRAEPOSITION NOMEN
                ----- -------- ------------ -----
                ↓    ↓       ↓            ↓
               文タイプ トポロジ  imの品詞  reisefuehrerの品詞

第三フィールド  I.m. r.aI.-z.@.-f.y:.-r.@.r.
                ==== ======================== （単語）
                ---- ----- ---- ----- ------ （音節）
                - -  - --  - -  - --  - - - （音素）

第四フィールド  0 1000
                - ----
                ↓  ↓
     imのアクセント  reisefuehrerのアクセント
```

From: Michael Paul <paul@itl.atr.co.jp>

Ashimura-san,

> ８／８最新版のドイツ語出力形式は、以下のとおりに変更されています。
> ただし、読みから、「.」および「-」を除けば、今までのフォーマット
> （単語｜品詞｜読み｜アクセント）として、g-filterで処理することが可能です。

前のフォーマットに戻ること二ついて：

前のフォーマットで一つの線では一つの単語が入りました。今の新フォーマット
では文の句が入りますのでもし前のフォーマットに戻りたい時は半角スペースも
消さないといけないですよ（前の方は読みのフィールドだけ使ったから他の

フィールドは別に変換しなくてもいいです）。

※注意：

生成した文の CHATR の pause の音素は変わりました。

前音素:　「 2 」「 3 」「 4 」
新音素:　「2」　「3」　「4」

しかし、前のフォーマットに変換する時 pause の回りで半角スペース「 」を
入れないといけないんですけど。
でも pause の音素は音素区切りを入りませんから認定は問題ないと思います。

参考文献

[1] システム科学研究所 「韻律処理マニュアル」 1994 年 2 月 1 日

[2] 国際電信電話株式会社研究所 「形態素解析処理技術の研究 1993 年 3 月

[3] 酒井 慎介 藤澤 謙 Nick Campbell 「言語フィルターの改善」 1999 年 2 月 26 日