非公開

TR-S-0001

データベース仕様書（第１版）

Specifications for the Database(Ver.1)

中嶋秀治 Hideharu Nakajima 熊野正 Tadashi Kumano　松井知子
Tomoko Matsui - 山本博史 Hirofumi Yamamoto - 隅田英一郎
Eiichiro Sumita - 柏岡秀紀 Hideki Kashioka - 竹澤寿幸
Toshiyuki Takezawa - マルコフ コンスタンティン Konstantin
Markov

2000.8.15

目次

1　はじめに

ATR 音声言語通信研究所で構築するデータベースの仕様を述べる。ATR 音声言語通信研究所では、旅行会話などの対話音声や、講演等の独話音声のデータベースも構築する。また、逐次通訳や後付けの翻訳や同時通訳のデータベースも構築する。本稿では、さまざまな研究目的の全てのデータベースに共通する仕様について記述する。また、構築するデータベースでは、日本語のみならず英語や中国語などの多言語データが収集の対象であるが、本稿では、多言語共通の仕様と、日本語および英語に関する固有の仕様を述べる。

2　ファイル構成と各ファイルの概要

本データベースは、収録環境のデータ、音声波形データ、書き起しテキスト、及び、形態素データ等からなる。 このようなファイル構成と各ファイルの形式を述べる。

2.1　データベースの構成

データベースを構成するファイルを表 1 に示す。

表 1: ファイル種別とファイル名

ファイル種別	ファイル名
環境ファイル	識別番号.env
物理ファイル	識別番号.拡張子.aud または識別番号.拡張子.img
波形情報ファイル	識別番号.拡張子.wif
書き起しファイル	識別番号.拡張子.txt
音素転記ファイル	識別番号.拡張子.trs
形態素情報ファイル	識別番号.拡張子.mor

データベースの環境ファイルには、一連の収録（任意チャネル数による独話、対話、通訳付き独話／対話）に対して、8 桁の識別番号（例えば、00123456）を付与する。なお、識別番号は ATR 音声言語通信研究所内の全データベースでユニーク であればよい。番号付与の詳細は別途指定する。すなわち、表の「識別番号」が通し番号となる。
拡張子の部分には、環境ファイルで指定された拡張子が入る。
ファイル種別は、ファイル名の末尾の env、aud、img、wif、txt、trs、mor などのファイル識別子（小文字）で表現する。

各ファイルの概要説明は以下の通りである。（各ファイルの記述内容の詳細は以後の各節、または、別途述べる。）

環境ファイル(.env)

通訳行為以外の収録環境を再現できる情報を記載するファイル。さらに、このファイルは、波形情報ファイル、物理ファイル、書き起こしファイル、音素転記ファイルとの対応関係を示すファイルである。
記載内容は、以後の 3 章で説明する。

波形情報ファイル(.wif)

波形または画像の別、さらに格納時のサンプリング周波数、バイトオーダ、コーディング種別などの情報を記述するファイル。
詳細は以後の 3 章で説明する。

物理ファイル(.aud または.img)

上記の環境の下で収録され、別に決めたフォーマットで、音声波形または画像を格納するファイル。

書き起しファイル(.txt)

上記の音声波形ファイルをテキスト化した結果が格納されるファイル。 テキスト化する内容は音を文字化したもので、単語、[ ]つき間投詞、{ }を使った音の挿入／削除／置換の記録、文末としての句点と/、書けないけれども何か音があったことを示す記号とする。
詳細は以後の 5 章で説明する。

音素転記ファイル(.trs)

書き起しファイルの内容を音素表記したものを時刻情報（開始時刻と終了時刻）とともに記載したファイル。 咳払いなどを音素として扱う場合にはそれらも書く。
詳細は以後の 6 章で説明する。

形態素情報ファイル(.mor)

1) ATR 音声言語通信研究所が定義する形態素体系に添った形態素を、一行ごとに記述するファイルで、 上記の書き起しファイルに含まれる形態素の、開始時刻、終了時刻、品詞 などの詳細情報を記載したファイル。
詳細は以後の 7 章で説明する。

2.2 ファイルの作成順序の概略

以上のファイル作成の大まかな順序は以下の通りである。

1) 収録時に「環境ファイル」を作成する

2) 環境ファイルに従って収録したデータを「物理ファイル」に格納する。並行して、「波形情報ファイル」を作成する
3) 音声波形の場合、機械的に「物理ファイル」を分割し、その単位毎に人手で「書き起しファイル」 を作成する
4) 「書き起しファイル」を入力として、機械と人手で「形態素情報ファイル」を作成する

各ステップの詳細は本稿の以後の章節で説明、または、別途指定する。

音素転記ファイルは基本的に自動作成するが、自動作成されたファイルのチェックは専門作業者が目視で行なう。なお、その音素転記ファイルに誤りがあった場合には、その箇所をリストアップする。誤りは即座に手で直すのではなく、

1) 辞書を確認する
2) 書き起こしテキストを確認する

の順に確認を行い、誤りの原因を探し、各ファイル間の整合性を保ちつつ、誤りの訂正を行っていく。

以後、各ファイルについて簡単にまとめる。詳細は 3 章から説明する。

2.3 ファイルの形式

環境ファイル(.env)

ファイルの定義

通訳行為以外の収録環境を再現できる情報を記載するファイル。 さらに、このファイルは、波形情報ファイル、物理ファイル、書き起こしファイル、音素転記ファイルとの対応関係を示すファイルである。

ファイル名

識別番号.env

一連の収録ごとに、識別子 env の環境ファイルを用意する。

形式

フォーマットは以後の 3 章で詳しく説明する。

波形情報ファイル(.wif)

ファイルの定義

波形または画像の別、さらに格納時のサンプリング周波数、バイトオーダ、 コーディング種別などの情報を記述するファイル。

ファイル名

識別番号.拡張子.wif

env と同じ識別番号を用いる。 拡張子の部分は環境ファイル（.env）で指定される。

形式

フォーマットは以後の 4 章で詳しく説明する。

物理ファイル(.aud または .img)

ファイルの定義

上記の環境の下で収録され、別に決めたフォーマットで、音声波形または画像を格納するファイル。

ファイル名

音声波形の場合、ファイル名は、

識別番号.拡張子.aud

をデフォルトとする。
画像の場合、ファイル名は、

識別番号.拡張子.img

をデフォルトとする。 env と同じ識別番号を用いる。 拡張子の部分は環境ファイル（.env）で指定される。

物理ファイルの収録環境の例は 3 章に示す。

書き起しテキスト(.txt)

ファイルの定義

上記の音声波形ファイルをテキスト化した結果が格納されるファイル。テキスト化する内容は音を文字化したもので、単語、[ ]つき間投詞、{ }を使った音の挿入／削除／置換の記録、文末としての句点と / 、 書けないけれども何か音があったことを示す記号とする。

ファイル名

識別番号.拡張子.txt

env と同じ識別番号を用いる。拡張子の部分は環境ファイル（.env）で指定される。

形式

上のような形式で記述する。 読点「、」は文の見やすさを考慮して付与するが、特に制限は与えない。 また音響的特徴との関連性は考慮されない。記述法の詳細は 5 章で説明する。

音素転記ファイル(.trs)

ファイルの定義

書き起しファイルの内容を音素表記したものを 時刻情報（開始時刻と終了時刻）とともに記載したファイル。咳払いなどを音素として扱う場合にはそれらも書く。

ファイル名

識別番号.拡張子.trs

env と同じ識別番号を用いる。 拡張子の部分は環境ファイル（.env）で指定される。

形式

開始時刻 終了時刻 音素記号列

の順に書く。 時刻の単位は''秒(sec)''である。

形態素ファイル(.mor)

記述内容の定義

ATR 音声言語通信研究所が定義する形態素体系に添った形態素を、一行ごとに記述するファイルで、上記の書き起しファイルに含まれる形態素の、開始時刻、終了時刻、品詞などの詳細情報を記載したファイル。

ファイル名

識別番号.拡張子.mor

env と同じ識別番号を用いる。拡張子の部分は環境ファイル（.env）で指定される。

形式

- mor のファイルは、１行１形態素とする。
- 各項目は「/」で区切られる。項目内容が無くても「/」は省略しない。
- 1 行の記述順序は LEX, POS, PRO, MRK, SEQ, TIM, MSC の順である。
  LEX、POS は機械付与、人手で CHECK。
  PRO、MRK、SEQ、MSC,TIM は完全機械付与。
- merge-list や replace-word を廃止し、mor に埋め込む。
- 以下の各項目内での小項目間の区切りには「|」を利用する。

    - LEX = 「 出現形|出現基本形|統一形|統一基本形」
      出現形：TEXT のままの活用形
      出現基本形：TEXT のままの終止形
      統一形：表記の揺れを吸収した活用形
      統一基本形：表記の揺れを吸収した終止形
      ;;; 多少冗長だが、安全のため上のように記述する。
      ;;; 機械処理できる見込み。英語も同様。

    - POS = 品詞情報
      = 「 品詞|活用型|活用形」 ... 日本語の場合
      = 「 品詞|意味活用|一致活用」 ... 英語の場合

    - PRO = 「読み」
      ・現在は日本語のみ、英語は現在は空列。

    - MRK = L or A or 空列
      ・L は言語専用、A は音声専用、空列は共用。

    - SEQ = 「 *sentenceID*|*utterenceID*|*turnID*|*sourceID*」
      ・SEQ の情報は上位からもらう。
      ・sourceID は発話者の識別用に env で定義される。

    - TIM = 当該形態素の開始時刻と終了時刻。

    - MSC = その他の情報。
      MSC の活用は公開するか、作業用の完全に LOCAL なデータとする。

- 記号「、」や「・」は mor に残す。
- チャネルは env ファイルに記述。

※ 詳細は 7 章に従うものとする。

3　環境ファイル（.env）の表記法

3.1 ファイル名の決め方

一連の収録（任意チャネル数による独話、対話、通訳付き独話／対話）に 対して、8桁の識別（通し）番号（例えば、00123456）を付与する。
なお、識別番号は ATR 音声言語通信研究所内の 全データベースでユニークであればよい。
一連の収録ごとに、拡張子 env の環境ファイルを用意する。（例えば、00123456.env）。

3.2 環境ファイルの項目と書き方の例

00123456.env
-- 8< -- 8< -- 8< --env ここから-- 8< -- 8< -- 8< --

```
------------------------------------------------------------------------
ファイル ID:    00123456
収録名:         NIL                  # NIL or <string>: 番組 ID や会話 ID
収録回:         3                    # NIL or <int>: 収録回
収録年月日:     21/2/2000            # NIL or <int/int/int>: 日/月/年
収録場所:       NIL                  # NIL or <string>: 場所名
収録機関:       インタグループ       # NIL or <string>: 機関名
ドメイン:       ニュース経済         # NIL or <string>: ドメイン名（管理法は別途）
タイプ:         NIL                  # NIL or "対面対話" or "非対面対話" or "独話"
チャネル数:     2                    # マイク数に概ね対応
チャネル ID:    1                    # NIL or <int>: 物理チャネルの ID 番号
                                     # (NIL: わからない、0: 音がない)
  wif 拡張子:   001                  # NIL or <string,string,...>: wif 拡張子のリスト
  マイク:       MU-2C                # NIL or <string>: マイク名
  入力機器:     PCM2700A,AMX3032     # NIL or <string,string,...>: 機種名のリスト
  サンプリング周波数: 48kHz          # NIL or <string>
  量子化精度:   16bit,linear         # NIL or <string,string,...>
チャネル ID:    2
   :
   :
ソース数:       2                    # <int>
  ソース ID:    1                    # NIL or <int>: ソースは書き起こしや形態素
                                     # 解析の処理単位に相当。実際には物理チャネル
```

```
                                    # に対応することが多い。
  拡張子: {001,AAA,AAA,AAA},        # <{string,string,string,string},
                                    # {string,string,string,string},...>:
                                    # {wif 拡張子,txt 拡張子,mor 拡張子,trs 拡張
子}
                                    # のリスト
  発話スタイル: 自由                # NIL or <string>: "孤立単語" or "文節" or
                                    # "朗読" or "自由"
  役割: NIL                         # NIL or <string>: 言語特性
                                    # "顧客" or "担当者" or "同時通訳者 s" or
                                    # "逐次通訳者 s" or "翻訳者 s" (s: ソース
ID)
  言語: jp,el                       # NIL or <string,string,...>: 言語のリスト
                                    # "jp" and/or "el" and/or "cn" ...
  話者: {山田・太,男性,30,東京,30/8/1970,アナウンサー,jp},
        {山本・洋,男性,35,大阪,27/4/1965,NIL,jp},
        other
                                    # NIL or "other" or
<{string,string,int,string,
                                    # int/int/int,string,string},...>:
                                    # 話者の情報 {名前,性別,年齢,出身地,生年月
                                    # 日,職業,母語,備考}
                                    # (NIL: 話者なし、other: 不特定)
                                    # (話者数が複数の場合は話者を区別した処理を
                                    # しない)
  ソース ID: 2
     :
     :
コメント:     NIL
-------------------------------------------------------------------------
-
# 注 1) {string}は','で区切られた string 列
# 注 2) 拡張子はデフォルトとして 3 桁とする
# 注 3) ある物理チャネルで複数の話者が発声し、それぞれを扱いたい時には、ソー
スを
#       分けて記述する。
# 注 4) あるソースに複数の言語が含まれる場合も許す。(言語: jp,en)
#
# <file ID (6 桁)>.<拡張子 (ENV ファイルで指定) >.<ファイル識別子
 (wif/txt/trs/mor) >
  → 00123456.001.wif
    00123456.AAA.txt
    00123456.AAA.mor
    00123456.AAA.trs

-- 8< -- 8< -- 8< -- ここまでが env -- 8< -- 8< -- 8< --
```

3.3 env ファイルの典型例

環境ファイルの典型的な例を独話と対話の場合について示す。

- 独話： 一人の話者が発声する音声を 2 本のマイクで収録する場合

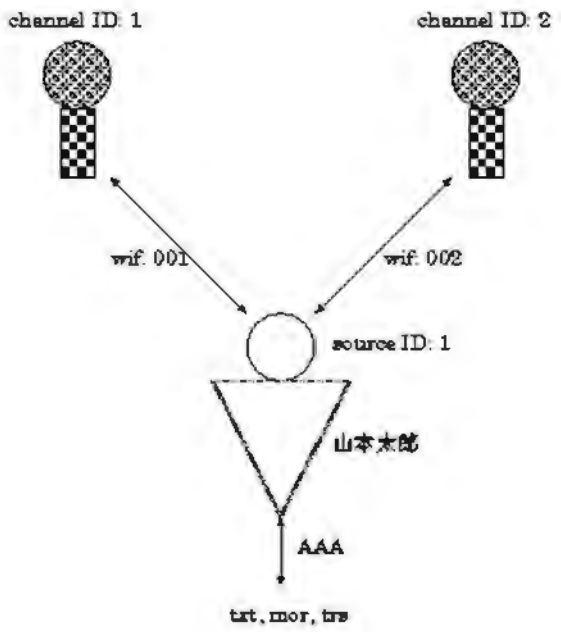

図 1: 一人の話者が発声する音声を 2 本のマイクで収録する場合

00111111.env

ファイル ID:　00111111
収録名:　NIL
収録回:　1
収録年月日:　25/4/2000
収録場所:　NIL
収録機関:　NIL
ドメイン:　音声認識
タイプ:　独話
チャネル数:　2
チャネル ID:　1
　wif 拡張子：　001,
　マイク:　MU-2C
　入力機器:　PCM2700A,AMX3032
　サンプリング周波数: 48kHz
　量子化精度:　16bit,linear

チャネル ID:　2
　wif 拡張子：　002,
　マイク:　　SONY-C350
　入力機器:　　PCM2700A,AMX3032
　サンプリング周波数: 48kHz
　量子化精度:　16bit,linear
ソース数:　　1
　ソース ID: 1
　拡張子: {001,AAA, AAA, AAA},{002,AAA,AAA,AAA}
　発話スタイル: 自由
　役割: NIL
　言語: jp
　話者: {山本・太郎,男性,35,大阪,27/4/1965,NIL}
コメント:　　NIL

→ 00111111.001.wif, 00111111.002.wif
　00111111.AAA.txt
　00111111.AAA.mor
　00111111.AAA.trs

- 対話： 二人の話者の対話を 2 本のマイクで収録する場合

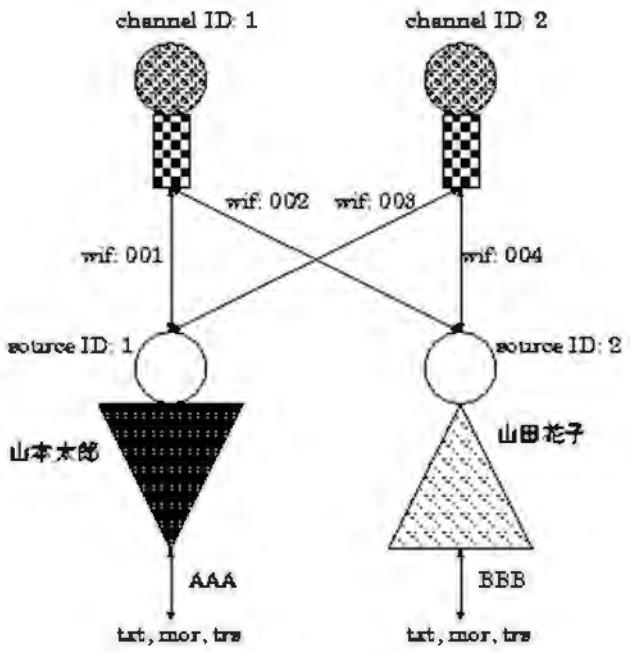

図 2: 二人の話者の対話を 2 本のマイクで収録する場合

00222222.env

ファイル ID:　00222222
収録名:　NIL
収録回:　1
収録年月日:　25/4/2000
収録場所:　NIL
収録機関:　NIL
ドメイン:　ホテル予約
タイプ:　対面対話
チャネル数:　2　#
チャネル ID:　1
　wif 拡張子：　001,002
　マイク:　MU-2C
　入力機器:　PCM2700A,AMX3032
　サンプリング周波数: 48kHz
　量子化精度:　16bit,linear
チャネル ID:　2
　wif 拡張子：　003,004
　マイク:　MU-2C
　入力機器:　PCM2700A,AMX3032

サンプリング周波数: 48kHz
量子化精度:　16bit,linear
ソース数:　2
ソース ID: 1
拡張子: {001,AAA,AAA,AAA},{003,AAA,AAA,AAA}
発話スタイル: 自由
役割: 顧客
言語: jp
話者: {山本・太郎,男性,35,大阪,27/4/1965,NIL}
ソース ID: 2
拡張子: {002,BBB,BBB,BBB},{004,BBB,BBB,BBB}
発話スタイル: 自由
役割: 担当者
言語: jp
話者: {山田・花子,女性,35,大阪,27/4/1965,NIL}
コメント:　NIL

→ 00222222.001.wif, 00222222.002.wif, 00222222.003.wif, 00222222.004.wif
00222222.AAA.txt, 00222222.BBB.txt
00222222.AAA.mor, 00222222.BBB.mor
00222222.AAA.trs, 00222222.BBB.trs

- ATR 言語データベース (LDB) の場合
-
-
- 00123456.env
- ----------
- ファイル ID:　00123456
- 収録名:　NIL　# NIL or <string>: 番組 ID や会話 ID
- 収録回:　NIL　# NIL or <int>: 収録回
- 収録年月日:　22/12/1994 # NIL or <int/int/int>: 日/月/年
- 収録場所:　NIL　# NIL or <string>: 場所名
- 収録機関:　日本アイアール株式会社　# NIL or <string>: 機関名
- ドメイン:　旅行　# NIL or <string>: ドメイン名
- タイプ:　非対面対話　# NIL or "対面対話" or "非対面対話" or "独話"
- チャネル数:　0　#
- ソース数:　4　# <int>
- ソース ID:　1　# NIL or <int>:
- 拡張子: {NIL,010,010,NIL},　# <{string,string,string,string},..>
- 発話スタイル: NIL　# NIL or <string>
- 役割: ?　# NIL or <string>

- 言語: el # NIL or <string,string,...>
- 話者: {ブーン・リア,女性,28,東京,NIL,会社員,el},
- ソース ID: 2 # NIL or <int>:
- 拡張子: {NIL,011,011,NIL}, # <{string,string,string,string},..>
- 発話スタイル: NIL # NIL or <string>
- 役割: 逐次通訳者 1 # NIL or <string>
- 言語: jp # NIL or <string,string,...>
- 話者: {高柳・智子,女性,29,東京,NIL,通訳,jp},
- ソース ID: 3 # NIL or <int>:
- 拡張子: {NIL,020,020,NIL}, # <{string,string,string,string},..>
- 発話スタイル: NIL # NIL or <string>
- 役割: ? # NIL or <string>
- 言語: jp # NIL or <string,string,...>
- 話者: {平川・加奈,女性,31,大阪,NIL,主婦,jp},
- ソース ID: 4 # NIL or <int>:
- 拡張子: {NIL,021,021,NIL}, # <{string,string,string,string},..>
- 発話スタイル: NIL # NIL or <string>
- 役割: 逐次通訳者 3 # NIL or <string>
- 言語: el # NIL or <string,string,...>
- 話者: {竹内・里佳,女性,29,神奈川,NIL,通訳,el},
- コメント: /DB/LDB/JE/ENV/DEN94B/13014107.ST, ノイズレベル:＜オフィスルーム＞,マイクロホン:ＥＣＭ－Ｔ１４０＜ＳＯＮＹ＞,ＤＡＴの機種:ＴＣＤ－Ｄ７＜ＳＯＮＹ＞,ミキサーの機種:NIL,サンプリング周波数:48kHz,発話形態:自由会話,トピック:レストランの予約,言語パターン:英語－通訳者－日本語,発声方法:文発声
- ----------
- 表現集の場合
-
- 00222222.env
- ----------
- ファイル ID: 00222222
- 収録名: NIL
- 収録回: NIL
- 収録年月日: NIL
- 収録場所: NIL
- 収録機関: NIL
- ドメイン: 旅行
- タイプ: NIL
- チャネル数: 0

- ソース数: 2
- ソース ID: 1 # NIL or <int>:
- 拡張子: {NIL,010,010,NIL}, # <{string,string,string,string},..>
- 発話スタイル: NIL # NIL or <string>
- 役割: NIL # NIL or <string>
- 言語: el # NIL or <string,string,...>
- 話者: NIL
- ソース ID: 2 # NIL or <int>:
- 拡張子: {NIL,020,020,NIL}, # <{string,string,string,string},..>
- 発話スタイル: NIL # NIL or <string>
- 役割: 翻訳者 1 # NIL or <string>
- 言語: jp # NIL or <string,string,...>
- 話者: NIL
- コメント: NIL
- ----------
- 同時通訳の場合
  第１の人が行なっている独話を第２の人が聞き、他言語に 翻訳して発声を行なう場合。
-
- 00222222.env
- ----------
- ファイル ID: 00222222
- 収録名: NIL
- 収録回: 1
- 収録年月日: 25/4/2000
- 収録場所: ATR 言語研 仮 打合せ室
- 収録機関: ATR 言語研
- ドメイン: 講演
- タイプ: 独話
- チャネル数: 2
- チャネル ID: 1
- wif 拡張子: 001
- マイク: ライン入力（DV テープより）
- 入力機器: DAT
- サンプリング周波数: 48kHz
- 量子化精度: 16bit,linear
- チャネル ID: 2
- wif 拡張子： 002
- マイク: SONY C-355
- 入力機器: PCM2700A,AMX3032
- サンプリング周波数: 48kHz
- 量子化精度: 16bit,linear

- ソース数:　　2
- 　ソース ID: 1
- 　拡張子: {001,AAA,AAA,AAA}
- 　発話スタイル: 自由
- 　役割: 演者
- 　言語: jp
- 　話者: {山本・太郎,男性,35,大阪,27/4/1965,NIL,jp}
- 　ソース ID: 2
- 　拡張子: {002,BBB,BBB,BBB}
- 　発話スタイル: 同時通訳
- 　役割: 通訳者
- 　言語: en
- 　話者: {山田・花子,女性,35,大阪,27/4/1965,NIL,jp}
- コメント: NHK「明日を読む」title:「インターネットが変える経済」
-
- ----------
-
- → 00222222.001.wif, 00222222.002.wif
- 　00222222.AAA.txt, 00222222.BBB.txt
- 　00222222.AAA.mor, 00222222.BBB.mor
- 　00222222.AAA.trs, 00222222.BBB.trs

4　波形情報ファイル(.wif)の表記法

4.1 ファイル名の決め方

環境ファイル（.env） に従って定める。

4.2 書き方の例

```
------------------------------------------------------------------------
  kind:                 audio     # <string>: wif ファイルの種類 "audio"
                                  # or "image"など

  sampling_rate:        16000     # <int>: サンプリング周波数 (Hz)
  n_bytes_in_sample:    2         # <int>: 1 サンプル当りのバイト数 (byte)
  byte_order:           BigEndian # <string>: "BigEndian" or "LittleEndian"
  coding:               pcm       # <string,string,...>:
                                  # "pcm", "ulaw", "shorten"など
  n_samples:            10        # <int>: このファイルで扱う物理ファイル数
  wav_file: 0.00        100.00 1 /DB/SDB/ALL/SPH/WAV/JAPANESE/xxx/yyy.aud
                         # <float> <float> <int> <string>:
                         # <開始時刻> <終了時刻> <トラック番号> <物理ファイル名>
                         # 時刻は''秒(sec)''単位で。
    :
    :

  sampling_rate:        12000
  wav_file: 2000.00     2500.00 1 /DB/SDB/ALL/SPH/WAV/JAPANESE/xxx/zzz.aud
    :
    :
  ------------------------------------------------------------------------
```

5　書き起こしテキスト(.txt)の表記法

書き起しの表記の詳細を本節で定める。

5.1 多言語共通項目

ここでは、全言語に共通の表記法についてまとめる。

入力コード

ユニコードを採用する。詳細は別途指定する。

文頭と文末

- 文末および行末には、半角の / を必ず入れる。
- 間投詞や言い誤りの後でも、音声が切れている時には / を入れる。
- 通常英文等では文頭の単語の先頭の文字は大文字で表記されるが、これを行なわない。

間投詞と言い誤りなど

- 間投詞は半角の[ ]で囲む。 １つの間投詞に１対の[ ]を使用し、間投詞が連続する場合には、それぞれを [ ]でくくる。[ええまあ]ではなく、[ええ][まあ]と書く。 間投詞一語のみの直後に「。」が来る場合もあり得る。
- 言い誤りなどの音の挿入、削除、および置換部分は半角の ``{''、``}''、および、``|'' で記述する。
- 聞こえた通りに書く。 よくありがちな表現はそのまま書き起す。
  英語についても、できるだけ言った通りに書く。簡単に気づくところは、例えば３単現の ｓ などは、{ }を使って書くが、受動と能動の態の修正のような複雑なものは要らない。
- ``['' 、``]''、``{''、および、 ``}'' の括弧で囲む場合は、それらの括弧の数が少なく、かつ、囲まれる文字数も少なくなるように囲む。単語の内部であることが明らかである箇所には境界を置かない。
- ''[``、'']''、''{``、および、''}'' を削除した書き起しテキストが正常なテキストになることを基本に作成していく。
- 言い誤り部分の書き起しについて述べる。
  表記法は { }の内外で同じである。
  聞こえた音自体は各言語に存在する単語の正しい音ではないが、聞こえた音からその単語の正しい音が分かる場合には、話者が言おうとした音の表記と実際に聞こえた発声内容の表記とを行なう。すなわち、

  {言おうとした表記|そのかな表記|実際の発声内容}

  の順に記述する。

  例は以下の通りである（ここでは日本語の例を示す）。

  置換誤りの場合（「いち」が「いつ」に）

  [実際の発声：]明日、だいいつ会議室で行ないます。
  [書き起こし：]明日、第{一|いち|いつ}会議室で行ないます。

  実際に聞こえたが、挿入誤りに当る場合には、

  {||実際の発声内容}

のように記述する。例は以下の通りである。

挿入誤りの場合（「て」の挿入）

[実際の発声：]今アナウンスメントを発送てしてるところです。
[書き起こし：]今アナウンスメントを発送{||て}してるところです。

実際に聞こえなかったが、正しくはそこに音が入り得る削除誤りの場合には、

{言おうとした表記 |そのかな表記|}

のように記述する。例は以下の通りである。

削除誤りの場合（会議の「会」の削除誤り）
[実際の発声：]明日、第一議室で行ないます。
[書き起こし：]明日、第一{会|かい|}議室で行ないます。

- 漢字を使わない言い誤りの時は、真ん中の項は空とする。例えば、

  {です||れす}

- 方言は言い誤りに入らない。
- 間投詞は半角の[ ]で囲むが、前の語の語尾を伸ばしたものとは区別する。[ ]の内部に聞き取れないことを示す記号を埋め込むことは禁止する。 すなわち、その内部では後述する@_uk_@は使用してはならない。 聞き取れない場合は、すべて{・・・||@_uk_@}のように記述する。

雑音などの表記

書き起こしにおいては、目的ソースの音源以外（例えばA さん以外）から発せられる音は全て雑音と見なし、別のソースとして書き起こす。書き起こすべき雑音の種類はATR 音声言語通信研究所第1研究室から改めて指定する。目的ソースが音声の場合、各境界設定区間とテキスト欄に記述する表記との対応を表2に従って書き起す。特に、聞き取れない場合には半角の``@_uk_@''を記述する。その他、必要に応じて表2の記号を利用する。全て半角文字で、英字は小文字とする

表2: 各境界設定区間とテキスト欄表記

境界設定区間	テキスト欄表記
non-speech	@_ns_@
連続有声	日本語文断片

書き起し出来ない連続有声	@_uk_@
咳の音	@_cg_@
笑いの音	@_lg_@
鼻をすする音	@_sz_@
息つぎ音	@_br_@
本文から外れた発声(ワキセリフ)	@_ng_@
リップノイズ	@_ls_@

ただし、リップノイズの定義は、「話はじめに唇から出るペチャッという音で、続く音声と同じくらいの音」とする。

“@_ng_@”は書き起しにない本文から外れた失敗の発声部分に用いる。 その失敗の発声中に咳や笑いの音があれば、それらも含めて全体 で “@_ng_@”とする。つまり、その場合の咳や笑いは、“@_cg_@”, “@_lg_@”としない。

他には、例えば、２人会話の収録時に、話者が相手に対して話すのではなく、話相手以外の第３者に話すようなワキゼリフは “@_ng_@”で記述する。

5.2 日本語書き起しファイルの入力仕様（記述方法について）

ここでは日本語の書き起しテキストファイル全般の書き方について述べる。表記は全般的に 5.1 節と 5.3 節に従う。

読点

読点「、」は書き起こしの見易さを考慮して付与するが、付与に関しての制限を特に与えない。また、音響的な特徴との関連性は考慮されずに付与してよい。

漢字の表記

漢字の表記、特に送り仮名については、5.3 節の表記の基準に従う。データベース作成時には漢字の読みは考慮しない。

略号

略号は全角のカタカナで記述する。

- １ｓｔ　ファースト
- ＪＲ　　ジェーアール or ジェイアール

- ＡＴＲ エーティーアール or エイティーアール

アルファベット、外来語

アルファベット、外来語は全角カタカナで表記する。

- Ｋ　ケー or ケイ
- ｓｗｅｅｔ　スイート or スウィート
- ｏｋ　オーケー or オーケイ or オッケー or オッケイ

人名

漢字での表記が特定できる人名は、漢字とする。姓名の間に「・」を付ける。

- オザワセイジ　小澤・征爾
- タカノユウイチ　タカノ・ユウイチ
- ヘルマンヘッセ　ヘルマン・ヘッセ

電話番号、小数点

電話番号には「，」、「ー（長音記号）」および「－（マイナス）」等の記号は用 いない。

- 零六, 九四九の一八三零です。

小数点は「点」で表記する。

- マイナス二点五パーセントです。

言い誤りの記述の詳細

言い誤りの部分は、{ | }で囲む。 囲まれる文字は、日本語の書き起しの場合、漢字、ひらがな、または、 カタカナとする。すなわち、表記法は { }の内外で同じである。 また囲まれる範囲が最も短く、括弧の数も少なくなるようにする。

実際に聞こえたが、文法的に正しくない挿入誤りに当る場合には、

{||実際の発声内容}

のように記述する。例は以下の通り。

挿入誤りの場合（「て」の挿入）
[実際の発声：]今アナウンスメントを発送てしてるところです。
[書き起こし：]今アナウンスメントを発送{||て}してるところです。

置換誤りの箇所は、言い誤っているが言い直しをしていない箇所に当る。聞こえた音自体は各言語に存在する単語の正しい音ではないが、聞こえた音からその単語の正しい音が分かる場合には、 話者が言おうとした表記、そのかな表記、および、実際に聞こえた発声内容の表記とを行なう。すなわち、

{言おうとした表記|そのかな表記|実際の発声内容}

の順に記述する。例は以下の通りである。

置換誤りの場合（「いち」が「いつ」に）
[実際の発声：]明日、だいいつ会議室で行ないます。
[書き起こし：]明日、第{一|いち|いつ}会議室で行ないます。

実際に聞こえなかったが、正しくはそこに音が入り得る削除誤りの場合には、

{言おうとした表記|そのかな表記|}

のように記述する。例は以下の通りである。

削除誤りの場合（会議の「会」の削除誤り）
[実際の発声：]明日、第一議室で行ないます。
[書き起こし：]明日、第一{会|かい|}議室で行ないます。

{ }を使って不必要な部分を囲む例を幾つか挙げると以下の通りである。

[実際の発声：]あなたただったらどうしますか。
[書き起こし：]あなた{||た}だったらどうしますか。

[実際の発声：]金曜は一時か昼からは空いてます。
[書き起こし：]金曜は一時{||か}昼からは空いてます。

[実際の発声：]サービス料込みでにひゃ。
[書き起こし：]サービス料込みで{||にひゃ}。

[実際の発声：]サーサービス料込みの値段です。
[書き起こし：]{||サー}サービス料込みの値段です。

[実際の発声：]タナカ・カカズコです。
[書き起こし：]タナカ・{||カ}カズコです。

間投詞

間投詞は半角コードの[ ]で囲む。

ただし、前の語の語尾を伸ばしたものとは区別する。[ ]で囲む範囲はなるべく最小にする。 [ ]の内部に聞き取れないことを示す記号を埋め込むことは禁止する。すなわち、その内部では@_uk_@は使用してはならない。聞き取れない場合は、 すべて{・・・|・・・|@_uk_@}のように記述する。

[実際の発声：]ですからー ・・・・
[書き起こし：]ですから ・・・・

[実際の発声：]ですから、あー ・・・・
[書き起こし：]ですから[あー] ・・・・

5.3 日本語テキストファイル表記の基準

1) 「数詞」は原則として漢数字表記とする。

例

三千二百五十七	(金額などの場合)
六の五四九零	(電話番号などの場合)
	(注: 「ゼロ」「レイ」は「零」と表記する。)
五〇六	(部屋番号や電話番号など、「ゼロ」を「マル」と発音している 場合)
	(注 1: 〇＝全角モード→z→c)
一つ	”ひとつ””ひとり”など数字の和読みも漢字で表記する。
	ひとつ 一つ
	ふたつ 二つ
	ひとり 一人
	ふたり 二人

2) 単語は以下を使わずに書く。

- 単独の濁点、半濁点
- 「～」
- １つ以上の連続した「ー」
- 「・・・」
- 「―」（マイナスやダッシュ）

3) 送りがな
送りがなについては長めを優先させる。

- おこなった　行なった
- ふりこみ　振り込み

名詞と動詞に使い分ける語は品詞によって送りがなをかえる。

- 名詞：送りがな無し が入ってきた。
- 動詞：送りがな有り 稲妻が。

5.4 英語書き起しファイルの入力仕様（記述方法について）

ここでは英語の書き起しテキストファイル全般の書き方について述べる。全般的に5.5節の表記の基準に従う。

文頭

文頭は大文字にせず、小文字で始める。ただし、代名詞 `I’や固有名詞、略語など、元々大文字で表記されるものはこの限りではない

固有名詞、地名、略号、個別のアルファベット読み

LDCの ``Proper names’’、 ``Acronyms IとII’’、 ``Individual letters’’ の表記をATRのデータベースから生成できるようにするため、 固有名詞、地名、略号、個別のアルファベットなどの表記には、 半角の ``、 ``、 `@’を使用する。詳細は以下で指定する。

略号

基本的に略号は用いない。スペルアウトする。

- 1st　first
- No.　number
- O.K., OK　okay
- alright　all right

- Y= yen

  ※ `yen’は大文字で始めない。
  ※ `all right’, `all righty’は `alright’, `alrighty’を 用いない。

ただし、それ以上スペルアウトできないものは可とする。

- a.m. / p.m.

文末とカンマ(,)、ピリオド(.)

文末には文末記号(/)をつける。 カンマ(,)やピリオド(.)は、必要に応じて適切に使用する。

言い誤りの記述の詳細

言い誤りの語句などは、{ | }で囲む。 囲まれる語句などは、発音に即した単語列で表記する。 { | }で囲む範囲は、できるだけ括弧の数を少なくするようにして、最も短い範囲にする。文境界を越えないような連続する複数の言い誤りなどの語句は 1 つの{ | }で囲む。

{ | }内の語句などは、辞書に存在する単語、および、 [ ] で囲まれた間投詞(後述)、 発音から辞書に存在する単語への対応をつけることができない発話断片（**表 2：各境界設定区間とテキスト欄**の表記法にしたがって@_uk_@ などと記す)からなること。 単語の表記法は{ | }の外と同一にする。{ | }で単語を分断してはならない。

× per{||@_uk_@}mit
○ {permit||@_uk_@}

挿入誤り、置換誤り、削除誤りの各表記法については、5 章に従う。次の３つは挿入誤りの例である。

- I want to go by {||bus} bus. /
- {||could} would you please ...
- the train {||if five} leave at five. /

{ | }内では文末記号(/)を使用しない。すなわち、{ | }は文境界をまたいではならない。

{ | }は通常の単語と同様に扱う。 すなわち、前後の単語とはスペースなどのデリミタで適切に区切り、また{ | }の前にカンマ(,)を打つときにはカンマと{ | }との間にスペースを入れ、{ | }の後にカンマ(,)やピリオド(.)を打つときには間は空けない。

語や句などが 2 回繰り返されている場合、言い直しか強調かをイントネーションで判断し、強調であれば{ | }でくくらない。

- 言い直し　I {||really} really like it. /
- 強調　　I really really like it. /

{ | }は連続して使用しない。

× {||the}{||there is}
○ {||the there is}

間投詞

間投詞は[ ]で囲み、[ ]内はすべて小文字を使用する。

[ ]内ではカンマ(,)やクエスチョンマーク(?)などの記号を使用しない。また、文末記号(/)も使用しない。すなわち、[ ]は文境界をまたいではならない。

[ ]で単語を分断してはならない。

× per [ah] mit
○ {permit||@_uk_@ [ah] @_uk_@} または {permit||per [ah] mit}

[ ]は通常の単語と同様に扱う。すなわち、前後の単語とはスペースなどのデリミタで適切に区切り、 また[ ]の前にカンマ(,)を打つときにはカンマと[ ]との間にスペースを入れ、[ ]の後にカンマ(,)やピリオド(.)を打つときには 間は空けない。

[ ] は通常の単語と同様に扱い、前後の単語とはスペースなどのデリミタで適切に区切る。

- yes, [eh] we can. /

[ ]の内部に聞き取れないことを示す記号を埋め込むことは禁止する。すなわち、その内部では@_uk_@は使用してはならない。聞き取れない場合は、すべて{・・・|@_uk_@}のように記述する。

母音の単音/長音は、明らかに言い誤りとわかるもの以外は基本的に間投詞扱いとし、以下のように表記を統一する。

(ア/アー)	[ah]
(イ)	[i]
(イー)	[ii]

（ウ）　　[u]

（ウー）　　[uu]

（エ）　　[e]

（エー/エイ）　　[eh]

（オ/オー）　　[oh]

感動詞的な ah や oh も区別せずに［ah］［oh］と間投詞扱いにする

1 間投詞に 1 つの[ ]を使用し、間投詞が連続する場合にはそれぞれを[ ]でくくる。（1 間投詞は 1 単語とは限らない。）

× ［ah ah］
○ ［ah］［ah］
　［oh dear］

不正確な発音と文法的な間違い

不正確な発音のために実在しない語に聞こえるものや言い誤りなどでのうち、明らかに正しい語が推測できる場合は、正しい語に書き改める。

- *analysist {analyst||@_uk_@}*
- *kay {okay||@_uk_@}*
- *childrens {children||@_uk_@}*

直接話法

直接話法の場合、ダブル引用符(″)を使用する。

- my friend told me, ″are you still taking care of father?″ /

that 節が使用されているにも関わらず、代名詞の人称が直接話法と同じ場合は直接話法と見なし、ダブル引用符(″)を使用する。

- the professional home helpers told me that, ″your father would pass away in two months if you put him in the hospital.″ /

1 つの引用は途中で改行することなく最後まで 1 行で処理する。

引用の途中でのクエスチョンマーク(?)、ピリオド(.)の後は 1 スペース空ける。

引用の途中には文境界(/)を含まない。

・ Mr. Okura said to me that, ``there is a hole in my wall in my hotel and that wall is reserved for you. I would like to have your orange crystals on the wall.'' /

引用箇所の後にさらに文が続く場合、引用の終りはカンマ(,)、クエスチョンマーク(?)を使用する。

・ ″it's all right,″ said he. /

直接話法の中にさらに直接話法などの引用がある場合は、シングル引用符(')を 使用する。 シングル引用符(')中の文の終わりは、カンマ(,)もしくはクエスチョンマーク(?)を使用する。

・ my acquaintance said that, ″one student came into the university, who wants to study the German literature and I asked that, 'have you ever read Hermann Hesse?' and he says, 'I've never heard of Hermann Hesse.'″ /

※ この場合、文末はピリオド(.)、シングル引用符(')、ダブル引用符 (″)の順となる。

タイトル

芝居のタイトルや書籍、雑誌、新聞名などは、引用符でくくらず、各々の単語を通常と同じように (普通名詞は普通名詞として、固有名詞は固有名詞として)表記する。

・ I'd like to see the phantom of the opera. /
・ New York times says that ...

綴り

名前などの綴りを言う場合は、各文字の前にを付与した上で、大文字とし、文字間は1スペース空ける。ハイフンは使用しない。
× M-A-R-Y
○ M A R Y

普通名詞+数字

「普通名詞+数字」で特定のものを表す場合、すべて小文字とする。

・ room eight o three
・ national route number one
・ exit three

5.5 英語テキストファイル表記の基準

個々の表記の正規形については、原則として研究社のリーダーズ英和辞典 （以下 READERS）の見出しに準じる。

略号

READERS の見出しに準じる。

- six P.M. six p.m.
- five AM five a.m.

省略形

省略形は、発音されている通りに用いてよい。

- I can't go there./

数字の表記

数字間では 2 桁表記する場合のみハイフンで結ぶ。

- fifty-three (53)
- three thousand five hundred and fifty-three (3553)
- nineteen twenty-nine (1929 年)
- nineteen tens (1910 年代)
- August twenty-second (8 月 22 日)

数詞

アラビア数字/ローマ数字は使用せずに、原則すべてスペルアウトする。（略語の一部の数詞については以下の 「略語」節を参照）

0（ゼロ）を「オー」と発音している場合は、o と表記する （oh としない）。

- price: eighty-five dollars
- address: fifty-third street
- phone number: o seven five one two three one seven o o
- time: nine a.m.

固有名詞・地名

固有名詞と地名は、各単語の先頭にを付与した上で、それぞれ大文字で開始する。

- I live in Japan./
- but Jacques Cousteau the French oceanographer spent ...

{}の中においても単語で書き表せる該当箇所にはを付与する。例えば、

- {||when I was in Japan}

のように書く。

日本語のローマ字表記

固有名詞など、公式の綴りがわかっている場合にはそれを利用する。 それ以外の場合は、**表 3：日本語のローマ字表記法**と**表 4：外来音節の表記法**に従ってローマ字に書き下す。

表 3: 日本語のローマ字表記法

	ア	イ	ウ	エ	オ
ア行	a	i	u	e	o
カ行	ka	ki	ku	ke	ko
サ行	sa	shi	su	se	so
タ行	ta	chi	tsu	te	to
ナ行	na	ni	nu	ne	no
ハ行	ha	hi	fu	he	ho
マ行	ma	mi	mu	me	mo
ヤ行	ya	--	yu	--	yo
ラ行	ra	ri	ru	re	ro
ワ行	wa	--	--	--	--
ン	n				
ガ行	ga	gi	gu	ge	go
ザ行	za	ji	zu	ze	zo
ダ行	da	ji	zu	de	do
バ行	ba	bi	bu	be	bo
パ行	pa	pi	pu	pe	po

キャ行	kya	--	kyu	--	kyo
シャ行	sha	--	shu	--	sho
チャ行	cha	--	chu	--	cho
ニャ行	nya	--	nyu	--	nyo
ヒャ行	hya	--	hyu	--	hyo
ミャ行	mya	--	myu	--	myo
リャ行	rya	--	ryu	--	ryo
ギャ行	gya	--	gyu	--	gyo
ジ（ヂ）ャ行	ja	--	ju	--	jo
ビャ行	bya	--	byu	--	byo
ピャ行	pya	--	pyu	--	pyo

表 4: 外来音節の表記法

音	綴り
ファ	fa
ティ	ti
フィ	fi
ディ	di
シェ	she
チェ	che
フェ	fe
ジェ	je
フォ	fo

「オー/オ」などの音は表記上区別しない。

× Taroo, Taroh, Tarou
○ Taro

原則としてハイフンは使用しない。

× Osaka-jo-koen
○ Osakajokoen

「歌舞伎」などは固有名詞とせず、小文字で始める。

× Kabuki
○ kabuki

略語

略語（複数の単語の頭文字をつなげた語の類）は、その読み方によって以下の 2 通りの記号を各単語の先頭に付与した上で、それぞれすべて大文字で記述する。

1）各文字のアルファベットの読みの列として発音するものには、先頭にを付与する。

- FBI
- FEMA （読みが「エフイーエムエー」の場合）

2）以外のものには、先頭に@を付与する。

- @NATO
- @FEMA （読みが「フィーマ」などの場合）

略語の中に含まれる数字には、例外的にアラビア数字を用いる。 このとき、これらの数字も含めて、すべての文字が各々の読みの列として発音されているときのみ先頭にを付与し、それ以外のものには先頭に@を付与する。

- A320 （読みが「エイスリーツーオー」として）
- @COP3 （読みが「コップスリー」として）
- G7 （読みが「ジーセブン」として）

本来略語の一部である数字が離れている場合もアラビア数字を使用する。

- A320 and 330 are ... （330 の読みが「スリースリーオー」 として）
- @COP4 as well as @3 ...

略語が間投詞やいい直しで分断されている場合も通常の語の分断と同様の処理とする。

- A3 [ah] 20
- A3 {|@_uk_@} 20
-

6　音素転記ファイル(.trs)の表記

音素の表記には、各国語での音素記号を用いる。 日本語の場合、現在は以下の 27 個の記号である。

```
a  i  k  j  o   zh  z  u  d  m
g  ch ng r  sh  ts  s  e  b  q
t  w  n  p  h   f   -
```

境界設定区間の時刻情報の単位としては''秒(sec)''を用いる。 0.25 秒以上の無声区間が存在する場合に、境界設定区間の無音(@_ns_@)を設ける。 trs の１行は書き起こし（txt）の１行に対応する。

7　形態素情報ファイル(.mor)

形態素情報ファイルの詳細は、

http://lab.slt.atr.co.jp/dept3/TDMT/component.html#DB

に置かれる。各言語の形態素仕様の詳細が必要な場合には、上の URL の情報を参照すること。

8　データベース作成作業仕様

作業手順 ：書き起しと形態素データの作成まで

書き起しから形態素データ作成までのステップを示す。

ステップの冒頭にある'*'の記号は、そのステップが人手による作業を伴うことを示している。 記号が冒頭についていないステップはプログラムで行なわれる。

書き起しの表記についての詳細については 5 章の記載に従うものとする。

形態素データの各フィールドへの記載事項は 2.3 節および 7 章に従うものとする。

1) 音声波形の自動切りだしと認識を行う。
◎各音声区間の始終端時刻と認識結果が記録される。

2) *音声区間の確認を行う。
音声区間±３秒の波形を表示し、1 で得られた区間をマークする。
表示された波形、および音から音声区間切りだしにミスがないかを確認する。ミスがあれば正しい区間（波形表示された区間内で）をクリップする。
◎確認後の各音声区間の始終端時刻が記録される。

3) *2 の区間の音を聞いて書き起こしを行う。
一区間、一行とし、行末と文の区切り目には必ず半角の「/」を入れる。
/ は機械的に挿入されるが、専門作業者は、明らかな誤り（継続中の音声の中に置か

れた / など）がないかどうかを検査し、誤りは修正する。
書き起し時の表記の詳細は章に示す。

[あ]ありがとうご{ざ||な}いました。/よろしくお願いします。/

◎各音声区間に対する書き起こし。

4）3)の書き起こしを一文、一行とするとともに、音声区間の始終端時刻を埋め込む。

<S350.4>[あ]ありがとうご{ざ||な}いました。
よろしくお願いします。<E353.6>

◎始終端時刻付きテキスト（作業の中間結果のファイル）。

5）4)からクリーンなテキストを抽出する。

ありがとうございました。
よろしくお願いします。

6）*形態素解析を行う。
１行に１つの形態素の情報を書く。

ありがとうございました|ありがとうございました||/感動詞||
。|。||/記号||
よろしくお願いします|よろしくお願いします||/感動詞||
。|。||/記号||

◎読み、言い誤り、および時刻情報のない mor ファイル

7）*未知語に読みを付け、マスター辞書登録する。
未知語に音素列を付け、認識辞書登録する。
◎未知語追加されたマスター辞書、認識辞書。

8）{||}の情報を埋め込む。
上記 4)の結果と上記 6)の結果との間で DP マッチングを行なうことにより、4)の情報（{}内の情報）を埋め込む。{||}の|の最左のフィールドの情報は 形態素の行の左から 3 番目のフィールドに L が入っている行のように書く。|の最右の情報は同じフィールドに A が入っている行のように書く。

あ|あ|あ|あ/間投詞||//*A*/|||/350.4|/
ありがとうございました|ありがとうございました|ありがとうございました|
¥

ありがとうございました/感動詞||//*L*/|||/|/
ありがとうごないました|ありがとうごないました|ありがとうございました|
¥

ありがとうございました/その他|//A/|||/|/

。|。|。|。/記号|//|||/|/
よろしくお願いします|よろしくお願いします|よろしくお願いします|￥

よろしくお願いします/感動詞|//|||/|/

。|。|。|。/記号|//|||/|353.6/

◎音声区間始終端時刻付き mor

9) *未知の間投詞を認識辞書登録する。
◎未知間投詞の追加された認識辞書。

10) 発声に従った始終端時刻付き単語列を抽出する。

350.4
あ|あ|あ|あ/間投詞|//A/|||/350.4|/
ありがとうごないました|ありがとうごないました|ありがとうございました|
￥

ありがとうございました/その他|//A/|||/|/

よろしくお願いします|よろしくお願いします|よろしくお願いします|￥

よろしくお願いします/感動詞|//|||/|/

353.6

11) 10)の単語列から最尤の読み付き単語列と音素系列を得る。
「L」付きの単語は直後の「A」付き単語の表記（なければ空）
と DP コスト最小のものを読みとする。

あ|あ|あ|あ/間投詞|/ア/A/|||/350.4|350.5/

ありがとうございました|ありがとうございました|ありがとうございました|
￥

ありがとうございました/感動詞|/アリガトウゴザイマシタ/L/|||/|/

ありがとうごないました|ありがとうごないました|ありがとうございました|
￥

ありがとうございました/その他|/アリガトウゴナイマシタ
/A/|||/350.6|351.1/

。|。|。|。/記号|||///|||/|/
よろしくお願いします|よろしくお願いします|よろしくお願いします|￥

よろしくお願いします/感動詞||/ヨロシクオネガイシマス//|||/351.3|353.4/

。|。|。|。/記号|||///|||/|/

◎読み、単語境界時刻付き mor
◎trs ファイル

12）*.wif ファイルと.trs ファイルの不整合をチェックする。音素転記ファイルは基本的に自動作成するが、自動作成されたファイルのチェックは専門作業者が行なう。なお、その音素転記ファイルに誤りがあった場合には、その箇所をリストアップする。誤りはすぐに手で直すのではなく、

(a) 辞書を確認する
(b) 書き起こしテキストを確認する

の順に確認を行い、誤りの原因を探し、 各ファイル間の整合性を保ちつつ、誤りの訂正を行っていく。

ここで得られる最終結果（読み、単語境界時刻付き mor ファイル）から、本データベースを構成する各ファイル（例えば、書き起しテキストファイル（txt）や音素転記ファイル（trs））を生成できる。 また、本データベースとしては格納されないが、クリーンなテキストファイルも書き起こしテキストファイル（txt）から生成できる。

9 おわりに

本稿ではデータベースの仕様（第 1 版）を記した。特に、本仕様書は現時点で収集の予定がわかっている種類のデータの格納が行なえることを優先して作成された。従って、 今後、新たな種類のデータの格納が必要な場合には、改訂が必要である。

本テクニカルレポートは仕様の第 1 版である。同じものを

http://lab.slt.atr.co.jp/database/slt_db.html

にも置く。また、以後の改訂は WEB のホームページ上にも行なう。また、適宜テクニカルレポートとしても改訂版を所内公開する予定である。

今後の検討事項は以下の通りである。

- 書き起こしの表記法を LDC 表記に合わせるかどうか

    - 現時点ではATRが独自に英語データを集めるかどうかが不明である。そのため、英語コーパスのためのLDC表記への対応については今後必要に応じて行なうこととした。
    - なお、LDC表記のうちの、``Acronyms I''、``Acronyms II''、``Proper names''、および、``Individual letters''は本仕様で採用した。
    - また、 LDC表記のうちのいくつかはATRの表記からの変換が可能である。 LDC表記とそれに対応する本仕様によるATR表記とを本仕様書の付録に示した。
- 画像データのための波形情報ファイル（.wif）の記述法

また仕様以外の案件は以下の通りである。

- ユニコード関連のツール整備
  コード関連のサブグループを中心に整備をすすめる。
- 既存のATRデータの新仕様への変換作業
  ツールの統一と維持にかかる負担の軽減などのために変換を行なうこととした。 変換の対象は SDB、SLDB、LDB。 作業の進め方は各研究室の代表で決めて行なわれる。
- ファイルの識別番号の発行方法
  各研究室の代表で決めて別途指定される予定である。

10 謝辞

1999年末からの本仕様の検討には、塚田元氏、中村篤氏、内藤正樹氏も参加されました。また、本仕様の原案は、内藤正樹氏、中村篤氏、伊藤いずみ氏によって作成されました。そして、ラベリング等の専門作業者の皆様からは本仕様へのコメントをいただきました。以上の皆さんと、各研究室の担当部分の詳細を検討下さった各研究室の担当者をはじめとするATR音声言語通信研究所の皆さんに感謝します

付録：書き起こしにおける LDC 表記と ATR 表記との対応

```
================================================================================
CONDITION
        DESCRIPTION (LDC による表記の説明)
LDC     LDC の表記法
ATR     上の LDC の表記に対応する ATR の表記法
================================================================================
* Numerals
        all numerals are written out in full
LDC     twenty-two
ATR     twenty-two
--------------------------------------------------------------------------------
* Acronyms I
        acroynyms pronounced as a single word
LDC     @NATO
ATR     @NATO (LDC の表記法を採用する)
--------------------------------------------------------------------------------
* Acronyms II
        acronyms pronounced as series of letters
LDC     ~FBI
ATR     ~FBI (LDC の表記法を採用する)
--------------------------------------------------------------------------------
* Individual letters
        Pronounced individual letters
LDC     ~A ~B ~C
ATR     ~A ~B ~C (LDC の表記法を採用する)
--------------------------------------------------------------------------------
* Proper names
        both proper names and place names should be
LDC     ^Homer
ATR     ^Homer (LDC の表記法を採用する)
--------------------------------------------------------------------------------


* Partial words
        partial words are indicated with a dash
LDC     absolu-
ATR     {absolute||@_uk_@}
--------------------------------------------------------------------------------
* Mispronounced words
        mispronounced word
LDC     +probably
ATR     {probably||@_uk_@}
--------------------------------------------------------------------------------
```

```
* Idiosyncratic words
        speaker uses a ``made-up'' word
LDC     *schlump
ATR     {@_uk_@||@_uk_@}
-------------------------------------------------------------------------
* Speaker noise
        sound made by the talker. Use only those sounds described
LDC     {breath}
ATR     {||@_x_@} で x = ng,uk,br,sz,cg,lg
-------------------------------------------------------------------------
* Background
        sound not made by the talker(usually background)
LDC     [text]
ATR     別チャネルとして格納
-------------------------------------------------------------------------
* Backgroung noise
        start / end of non speaker noise
LDC     [text/] (extended) [/text]
ATR     別チャネルまたは@_x_@記号 (x = ng,uk,br,sz,cg,lg) で記述
-------------------------------------------------------------------------
* Semi-intelligible speech
        unintelligible speech. This is the transcriber's best attempt
LDC     ((text))
ATR     {推測できた単語||@_uk_@}
-------------------------------------------------------------------------
* Unintelligible speech(token)
        completely unintelligible speech
LDC     (( ))
ATR     {||@_uk_@}
-------------------------------------------------------------------------
* Unintelligible speech
        long period of unintelligible (long span) speech - skipped
LDC     [[skip]]
ATR     {||@_uk_@}    (注)上との区別は時間に閾値を設けた場合に可能
-------------------------------------------------------------------------
* Repeated section of speech
        Period of repeated speech(broadcast)
LDC     [[repeat]]
ATR     {||repeated words}    (注) 挿入誤りも同じ表記をとる点が問題
-------------------------------------------------------------------------
* Foreign language
        this is used to indicate foreign speech
LDC     <language text>
ATR     保留中 ・・・mor file に新領域を設ける案が有る
-------------------------------------------------------------------------
* Speaker aside
```

aside made from main speaker to background individ.
LDC　〈as〉 text 〈/as〉
ATR　{||@_ng_@}で書く。
---
* Overlapping speech (same channel)
used to indicate overlapping speech on
LDC　〈ov〉 text 〈/ov〉
ATR　別チャネルに格納予定
==========
* Non-lexemes　辞書にない間投詞類のリスト
* Interjections　間投詞リスト
それぞれ別途定義する。