

音声対話コーパスのマークアップ

Marking up spoken dialog corpora

土屋俊(千葉大学)

板橋修一(国立情報学研究所、産業技術総合研究所)

大須賀智子(国立情報学研究所)

TUTIYA Syun (Chiba University)

ITAHASHI Shuichi (National Institute of Advanced Industrial
Science and Technology, National Institute of Informatics)

OHSUGA Tomoko (National Institute of Informatics)

May 17, 2006

音声対話コーパスとは？

- 音声対話とは、対話(2人以上の人が参加するコミュニケーション)であって、その参加者による表現活動が主として音声言語の使用を中心に行われるコミュニケーション活動
- 従来は、「転記」テキストによる研究が中心であった
- 1980年代後半から、計算機および周辺機器の性能の向上によって、「音声」、さらに「映像」を保存できるようになってきた。
 - DAT、Compact Disk、DVD、CPU性能等

コーパスの電子化

- したがって、デジタルな録音とデジタルな転記とを関連づけた言語資料を作成し、利用するという可能性が1990年代に生まれた
- 音声認識研究において、統計的手法が活用されるようになり、コーパスへの需要が高まった
- いくつかの試み：
 - Map Task Dialog Corpus(Edinburgh)
 - TRAINS Corpus(Rochester)など

日本語マップタスクコーパス

- エディンバラの設計を模倣
 - 128対話 (Eye contactあるなしで64 x 2)
 - より正確な転記をめざす
 - 時間の表示
 - オーバーラップの記載
 - 発話単位の吟味
- 実物が大事！

TEI的挑戦

- オーバーラップの記述
 - 転記テキストの外にある音声ファイルへの参照
 - <timeline>の活用
 - <link>の活用
- 参加者の記述
 - 同じ人が複数回参加することの記述方法
 - P3だと<copyOf>を使う
 - XML化によって容易な解決
- ついに数ヶ月うちに公開予定(NIIから)

00:04:832-00:05:248 G:おいしょ<1600>

00:06:848-00:07:792 G:(えっと)<3280>;ささやくよう
にしている

00:11:072-00:12:432 G:すたーとちてん*は

00:12:368-00:12:704 F: *はい<272>

00:12:976-00:13:952 G:きゃんぷじょうです<288>

```
<anchor xml:id="u1-1" synch="#time1"/>
<anchor xml:id="u1-2" synch="#time2"/>
```

```
<anchor xml:id="u2-1" synch="#time1"/>
<anchor xml:id="u2-2" synch="#time2"/>
```

```
<timeline>
  <when xml:id="time1" absolute="??:??:???"/>
  <when xml:id="time1"/>
</timeline>
```

```
<linkGrp>
  <link targets="#u1-1 #u2-1 #time1"/>
  <link targets="#u1-2 #u2-2 #time2"/>
</linkGrp>
```