JAISTタグ付き自由対話コーパスの復元方法

本データはタグの情報のみを含んでいます。ここではテキストを含めた完全な
コーパスを復元する方法を説明します。

1 名大会話コーパスの入手

以下のウェブサイトから『名大会話コーパス』をダウンロードして下さい。

http://mmsrv.ninjal.ac.jp/nucc/

nucc.zip というファイルをダウンロード後、それを展開して下さい。
以降、ファイルを展開したディレクトリを nucc とします。

2 ツールによるコーパスの復元

コーパスを復元するためには recon_corpus.pl というツールを使います。
このツールは Perl のスクリプトです。復元ツールを使うためには Perl が
実行できる環境が必要です。ツールの動作確認は Perl ver.5.20.2 で行って
いますが、ver.5以降の Perl でしたら正常に動作すると思います。

コマンドライン上で recon_corpus.pl というツールを実行します。
その際、3つのオプションを指定します。

- -n 名大会話コーパスのあるディレクトリ
- -t 本データに含まれる annotation というディレクトリ
  (*.tag と *.idx というファイルを含むディレクトリ)
- -o 出力ディレクトリ
  (このディレクトリの下に復元されたコーパスのファイルが作成されます)

以下は実行例です。

```
perl recon_corpus.pl -n nucc -t annotation -o corpus
```

All files are successfully reconstructed というメッセージが表示されれば、
コーパスは正常に復元されています。
上の例では、復元されたコーパスは corpus というディレクトリに置かれます。
復元後のコーパスのファイルサイズの合計はおよそ8.4MBです。

以上