5H－9

発話コーパスを作る前に - その問題点と選択肢 - *

土井 晃一
株式会社 富士通研究所
doy@iias.flab.fujitsu.co.jp

1 はじめに

自然な発話を書き起こしたもの(以後コーパスと呼ぶ)は、言語心理学・言語社会学など文系の分野で使われてきた[1]。特に会話分析[2]の分野では、自然な発話を詳細にかつ正確に起こすための技術が工夫された。

近年、工学でも自然な発話を扱うようになり、特に自然言語処理やソフトウエア工学の分野で扱われるようになった[3, 4, 5, 6, 7, 8]。そこでは、計算機に入力し、何らかの処理をすることが前提となっている。計算機可読になったがための懸念点(明白に問題になるかどうかがまだわからないが、なんとなく気になる点)が生じている。特に書き起こす時の表記の問題が大きい。また、出来上がったコーパスをどのように利用するかには関わりなく、まずは発話の場をなるべく忠実に、また一様に(作業者による個人差がなく)再現することが重要となる。

そこで本稿では、懸念点に対して考えておくべき観点を挙げる。つまり、書き起こし規則の、習得容易性、面倒臭さ、弁別性、種々の選択肢。さらに出来上がったコーパスの人間の可読性である。

これらの事前に把握しづらいが、なおざりにすると後々大きな工数を要求することになる懸念点を観点を軸にして整理した。

2 何を書き起こすか?

本節では、コーパスを作成するに当たって、書き起こせる範囲について言及する。

自然な発話に対して書き起こせるものは、音声表現・音声付随表現・その他の三種に大別できる。

ここで言う音声表現とは、自然な発話の主として現象的な側面を指す。以下のようなものが挙げられる。

1. いわゆる言語表現
2. 文法的には感動詞として分類されるフィラー
3. 言い間違い
4. 言い直し

次に音声付随表現とは、自然な発話の主として認知的な側面を指す。以下のようなものが挙げられる。

1. イントネーション
2. 速さ
3. 挿入発話[9]
4. 声の大きさ
5. 重なり
6. 感情
7. 笑い
8. 間
9. 表情
10. 身振り
11. 動き

また、その他のものとして、

1. 聞きとれない
2. 起こした人のコメント
3. 時刻

が挙げられる。

*Before Making Corpus of Speech - Its Problems and Choice Points
Kouichi DOI (Fujitsu Laboratories)

3 あらかじめ考えておくべき懸念点

本節では、コーパスを作成するに当たって、あらかじめ考えておくべき懸念点について詳説する。

まず、懸念点を列挙してみる。

1. 数字・英語・記号・カタカナについて全半角のどちらをとるか
2. 数字は算用数字を使うか、それとも漢数字を使うか
3. 送りがなをどうするか
4. ひらがなと漢字の両方の表記の可能なものをどうするか
5. 外来語はアルファベットで書くか、それともカタカナで書くか
6. いわゆるカタカナ語の表記のゆれ、発話とのずれはどうするか
7. イントネーションはどうするか
8. 発言の終了をどうとらえるか
9. 同じ漢字で読みが違うものをどうするか
10. 発話の速さによる強調はどうするか
11. 同時発話をどうあつかうか
12. 言い間違いをどうあつかうか
13. 聞き取りにくいところをどうあつかうか

これらを順に説明していく。

1については、全半角は計算機で自動的に変換する方法が存在するから実はあまり問題ではない。しかし、半角カナだけは読めなくなってしまうことがあるので要注意。

2については、可能な限り算用数字を使うことにすれば容易に解決できる。

3はどういう懸念点かというと、送りがなのふりかたが一般に難しいという問題と、同じ漢字が二通り以上の読みを持つことがあり、作業者がそれに気がつかないことがあるという問題である。後者に関しては読みを付記することで解決できる。例えば、「行ってしまう」という例は「行(い)ってしまう」、「行(おこな)ってしまう」と付記することによって解決できる。ただしこのような例をあらかじめ枚挙しておく必要があり、それは一般には難しい。

4はどういう例が挙げられるかというと、「わかる」・「分かる」のような例である。これは可能な限り漢字で書けば解決で

	習得容易性	面倒臭さ	認知的	弁別性	選択肢	人間の可読性
1 全半角						×
2 漢数字						×
3 送りがな	×	×		×		×
4 ひらがなと漢字	×	×				×
5 外来語		×		×	×	×
6 カタカナ語の表記		×	×	×		×
7 イントネーション	×	×	×		×	×
8 発言の終了	×	×	×			×
9 同じ漢字で読みが違うもの	×	×				×
10 発話の速さによる強調	×	×	×			×
11 同時発話	×	×	×	×	×	×
12 言い間違い		×	×	×		×
13 聞き取りにくいところ		×	×	×		

表 1: 観点と懸念点のマトリックス

きる。ただし、わかりにくい例では、あらかじめ練習が必要であるという問題も持っている。

5はまさに慣用によるとしか言いようのないところで、こなれていない場合はアルファベットで書くようにすればかなり解決できる。さらに必ずなかぐろ「・」を打つようにすればよい。

6も慣用によるしかない。発話が慣用の表記と異なる時は発話の読みを併記することでかなり解決できる。

7は現象面を優先させるか、認知面を優先させるかで異なった表記になる。認知面を優先させれば、「?」(主として疑問)・「…」(主として発話維持調)・「!」(主として断定)を発話の最後にふることで解決できる。また、現象面を優先させれば、上昇調(おおむね疑問)・維持調(おおむね発話維持の意図が見受けられる)・下降調(おおむね断定・ひとりごと)に分けられる。それぞれ発話の最後に、例えば(↑)(→)(↓)を付記することで対応できる。

8はどういう気分で発言が終了したかを記述する方法があるか、という懸念点である。発話の終りに、例えば「(怒って)」と注書きすることで部分的には解決する。

9は必ず読みをふることで対応できる。例えば、「間(かん)」「間(あいだ)」「何(なに)」「何(なん)」などである。しかし、これも例をあらかじめ枚挙しておく必要がある。また、あらかじめ練習が必要となる。

10は完全に認知的現象であり、作業者の主観がでてしまうのが大きな問題と言える。

11は純粋に我々の直観であるが、いわゆる同時発話には認知的にいろいろな場合がありそうである。

12は言い間違えた方を括弧に入れることで解決できる。

13も12と同様に聞き取りにくいところを括弧に入れることで解決できる。

基本的には日本語の正書法を守り、発話と一致しない時は発話を小括弧に入れて表記し、小括弧は特殊記号とし、全半角は関知しない、なかぐろは打つようにし、後段の機械処理の前に括弧内は括弧を含めて取り除くようにすればかなりの問題は解決できる。

これらの懸念点を観点を軸にまとめてみる。観点は以下のように整理できる。

1. 習得容易性
2. 面倒臭さ
3. 認知的
4. 弁別性
5. 選択肢
6. 人間の可読性

「習得容易性」とは、ルールを作業者が習得するさいの容易性である。「面倒臭さ」とは、ルールを作業者が遂行する時の困難さである。「認知的」とは作業者の主観が入ってしまい、結果が一様にならないことを指す。「弁別性」とは、他のものとの弁別が容易にできるかどうかを指す。「選択肢」とは、解決策の選択肢がどれくらいあるかを指す。「人間の可読性」とは、できあがったコーパスの人間にとっての読みやすさを表す。

懸念点と観点を整理してみると表1のようになる。表中、問題になりそうなところを我々の主観で×で示した。これからコーパスを作ろうとする人は、×印で示された点で問題が生じる可能性があるので、なんらかの解決策を準備しておいたほうがよいと思われる。例えば、本節1,2,3,4,5,6,9の懸念点に関しては、ワープロの機能をうまく使うなどすれば、かなり改善できるであろう。

4 おわりに

コーパスを起こす前に考えておくべき懸念点/観点について整理した。これからコーパスを作ろうとしている方々の参考になれば幸いである。

参考文献

[1] 海保博之, 原田悦子. プロトコル分析入門 発話データから何を読むか. 新曜社, 1995.

[2] Stephen C. Levinson. *Pragmatics*. Cambridge University Press, 1983.

[3] 大森晃, 土井晃一. オフライン要求獲得法の提案. 情報処理学会第48回全国大会, Vol. 5, No. 4K-6, pp. 373–374, 3月 1994.

[4] 片山佳則, 蓬莱尚幸, 渡部勇, 土井晃一, 園部正幸. ユーザ指向ソフトウェア開発のための要求獲得/分析方法. 日本ソフトウェア科学会第12回大会, 1995.

[5] 片山佳則, 蓬莱尚幸, 渡部勇, 土井晃一, 園部正幸. ユーザ指向ソフトウェア開発のための要求分析法の実践に ついて. 日本ソフトウエア科学会 ソフトウエアプロセス研究会, 3月 1996.

[6] 土井晃一. 自然な発話における言い間違いに関する考察. 自然言語処理研究会, 7月 1995.

[7] 土井晃一. 要求獲得オフライン法での非機能/未分化要求の抽出. ソフトウエア工学研究会, 1月 1996.

[8] 土井晃一. オフライン要求獲得法における「笑い」の利用による「本音」情報の抽出. 自然言語処理研究会, 7月 1996.

[9] 島津明, 川森雅仁. 対話データの表記法. 情報処理学会第44回全国大会, Vol. 3, pp. 99–100, 1992.