『日本語話し言葉コーパス』における
文編集データについて
Version 1.0

野畑 周　内元 清貴　井佐原 均
(情報通信研究機構)

目次
1. 本冊子の内容
2. 文編集データの概要
2.1 ファイルの形式・格納場所
2.2 編集操作による文短縮の割合
2.3 編集操作の定義についての考察

1　本冊子の内容

本マニュアルは、『日本語話し言葉コーパス』[1] における要約データの一つである、文編集データの仕様を解説したものである。文編集データは、『日本語話し言葉コーパス』の統合データにおいて、SUW 属性 [2] の一部として提供されている重要文抽出データに対して作業を行ったものである [3]。文編集データ作成の対象講演は、他の要約データ (自由要約データ・重要文抽出データ) と同様に「コア」中の独話 177 講演 (対話・再朗読以外) とコア以外のテストセット 22 講演を含む 199 講演で、そのうち学会講演は 85 講演、模擬講演は 114 講演である。重要文抽出データは、二種類の要約率 (10%,50%) が設定され、各要約率ごとに 3 種類のデータが用意されており、一講演につき合計 6 個の重要文抽出データが存在する。文編集データは、これらの重要文抽出データ各々に対応して作成されたものである。

2　文編集データの概要

文編集作業の目的は、抽出された重要文を、制限された操作の中で可能な限り表現を短縮し、かつ自然な文章に近づけることである。『日本語話し言葉コーパス』においては、単語の品詞などを用いた一定の基準に従って機械的に節単位を認定した後、人手でその結果を修正して最終的な節単位を与えている [4]。ここでは重要文抽出データにおいてと同様に、そのようにして与えられた各節単位を「文」として扱っており、各編集操作の範囲は一節単位に制限している。従って、複数の節にまたがった編集は対象としていない。

作業の入力となる転記テキストでは、括弧付きでラベルが付与されている部分がある。これには、表 1 にあげたようなタイプの違いがあり、形態素情報付与データでの処理に沿って、それぞれ表の右欄のような規則で置き換えて用いた [5]。

文編集データでは編集操作を以下の 5 種類に区分し、それぞれに対応するタグ名を与えている。各編集操作はこのタグで削除操作と挿入操作それぞれの範囲を明示し、編集の種類やその範囲が各々区別で

表 1: ラベルのタイプと整形規則

ラベルのタイプ	例	整形規則
フィラー、感情表出系感動詞	(F あの)	全部削除
言い直し	(D こ) これ、これ (D2 は) が	全部削除
聞きとり、語彙同定、漢字表記に自信なし	(? タオングー)	候補を残す
全く分からない	(?)	削除
複数候補あり	(? あのー, あんのー)	前の候補を残す
音や言葉に関する引用	(M わ) は (M は) と表記	候補を残す
外国語や古語、方言など	(O ザッツファイン)	候補を残す
個人名、差別語、誹謗中傷、など	○○研の (R △△) さんが	候補を残す
基本形で漢字仮名以外の文字を使用する場合	(A イーユー;EU)	前者の候補のみ残す
何らかの原因で漢字表記できなくなった場合	(K い (F んー) ずみ;泉)	後者の候補のみ残す

きるようにしている。

なお、自由要約データと同様に文編集データの作成後に、人名など一部の表現については、転記テキストにおける扱いに沿って「×」記号によって伏字に変換することとしたため、情報が落ちている文があることを御了承されたい (詳細は [6] の (R) タグの項目を参照)。

1. 敬体常体変換 (Honorifics):
   文体を敬体から常体に変更するために、「です・ます」などの文末や節末表現を、「だ・である」などの表現に変換する。(削除タグ<hdel></hdel>、挿入タグ<hins></hins>)

   例:「そんな風に思<hdel>いまし</hdel><hins>っ</hins>た」

2. 重要文抽出による齟齬の修正 (errors by Extraction of sentences):
   直前の文が重要文として抽出されなかったために意味が取れなくなった接続詞など、重要文抽出の影響による齟齬を修正する。(削除タグ<edel></edel>、挿入タグ<eins></eins>)

   例:「最初の一日目は<edel>先程言ったように</edel>夜着いたので…」(対応する先行文が重要文として選択されていない場合)

3. 話し言葉特有の非適格表現の削除 (errors specific to Spontaneous speech):
   繰り返し・言い直し・言い誤りなど、話し言葉特有の構文的に適切でない表現を修正する。(削除タグ<sdel></sdel>、挿入タグ<sins></sins>)

   例:「<sdel>想定は</sdel>南の島を想定してます」

4. 口語文章語変換 (transformation to Written language):
   口語から文章語に変更するため、構文的には問題ないが敬体を常体に変更した結果文体として適切でなくなった表現を修正する。(削除タグ<wdel></wdel>、挿入タグ<wins></wins>)

   例:「多分一人で走<wdel>ん</wdel><wins>る</wins>のが不安で」

5. 意味上の編集 (revision in Meaning):
   その他、意味的に重要でなく削除しても問題ないと思われる表現を削除する。(削除タグ<mdel></mdel>、挿入タグ<mins></mins>)

   例:「上目黒<mdel>というところ</mdel>に住んで」

文編集データからは複数の編集結果を得ることができる。例えば、

「つまり何と言うか性能が向上したっていうことです」

という表現に編集操作を行うと、

「つまり<mdel>何と言うか</mdel>性能が向上した<mdel><wdel>って</wdel><wins>と</wins>いうこと<hdel>です</hdel><hins>だ</hins></mdel>」

といった編集結果が得られる。これによって、例えば、「敬体常体変換」、「口語文章語変換」を行った結果を見たいときには

「つまり何というか性能が向上したということだ」

という結果が得られ、最大限短縮した結果としては

「つまり性能が向上した」

という表現が得られる。

2.1 ファイルの形式・格納場所

文編集データは XML 形式で格納されている。CSJ の本体データとは別ファイルであるが、Talk 属性の TalkID、SUW 属性の ClauseUnitID を共有することによって講演単位、節単位での対応を示している [2]。

ファイル名は、各講演の ID に、ピリオドをはさんで要約率、作業者の区別を示す文字を付し、さらに文編集データであることを示す文字列 sr を付加したものとなる。アルファベットと要約率・作業者の対応を表 2 に示す。文編集ファイルは、要約率ごとにディレクトリを分けて格納されている。例えば、講演 A01F0055 に対する文編集データは、以下のように格納されている:

```
NICT/
     SR/
          A01F0055/
               10PER/
                    A01F0055.1U.sr.xml, A01F0055.1V.sr.xml,
                    A01F0055.1W.sr.xml
               50PER/
                    A01F0055.5U.sr.xml, A01F0055.5V.sr.xml,
                    A01F0055.5W.sr.xml
```

表 2: 要約データのファイル名

要約率を示す文字	要約率
1	10%
5	50%

作業者の区別を示す文字	作業者
U,V,W,X,Y,Z	作業者 (6 名)

表 3: 編集操作による文短縮の割合

	講演中の平均文字数	
編集操作	10%	50%
編集前	548.9	2339.0
(1) 敬体常体変換	529.5(0.035)	2254.9(0.036)
(2) 文抽出による齟齬	544.9(0.007)	2334.8(0.002)
(3) 非適格表現	537.4(0.021)	2292.2(0.020)
(4) 口語文章語変換	530.8(0.033)	2258.2(0.035)
(5) 意味上の編集	537.4(0.021)	2289.4(0.021)
全操作	484.4(0.118)	2073.4(0.114)

2.2 編集操作による文短縮の割合

編集操作によって、講演の平均文字数がどれだけ短縮されたかを表 3 に示す。表内の数字は、編集前と各編集操作を適用したときの文編集データにおける平均文字数である。括弧内の数字は、減少した文字数が編集前の文字数に占める割合を示している。

要約率 10%と 50%のデータでは、ともに全編集操作を適用することで 11%以上文章が短縮されている。また、短縮される文字数の傾向は両者で似通っている。「(2) 重要文抽出による齟齬の修正」における差は、要約率 50%のデータでは抽出される文が多く、齟齬が起きる部分が少なかったためと考えられる。

編集操作の中では、「(4) 口語文章語変換」と「(1) 敬体常体変換」による短縮の度合が大きい。また、繰り返しや言い誤りなど「(3) 話し言葉特有の非適格表現の削除」による短縮の度合も「(5) 意味上の編集」と同程度であり、話し言葉を書き言葉に直すことで 9%程度文章が短縮されることになる。

2.3 編集操作についての検討

文編集データの作成に際しては先に示したように 5 種類の操作を定義したが、各編集操作の対象となる表現は必ずしも自明でない。ここでは、編集操作の対象とする表現について検討した点のいくつかについてふれる。

- 「敬体常体変換」の範囲

  「敬体常体変換」では、基本的に「です」「ます」のみを対象として変換している。ただ、文末・節末から形式的な「です」「ます」を削除するのではなく、現在の書き言葉として適切な表現となるような変換を意図しているため、一部その範囲を越えた変換を行っている。つまり、三分法 (尊敬語、謙譲語、丁寧語) による敬語の分類としては、丁寧語を対象とした変換であるが、一部謙譲語も対象としており、より適切には話題と対話による敬語の種別における「対話の敬語」「聞手に対する敬語」を対象とした変換といえる [7]。例えば、「ございます」「いたします」などの表現は

そのまま「ます」を削除して常体に直すと「ござる」「いたす」となるが、これらの動詞は現在の書き言葉としては違和感があるため、「ある」「する」までの変換を「敬体常体変換」として行っている。また、「御覧ください」には「です」「ます」が含まれないが、同様にそのままでは書き言葉としては不自然であると判断し、「御覧いただきたい」に変換している。

一方、「お話しいただきました」「申し上げます」などの表現は、「敬体常体変換」としては「お話しいただいた」「申し上げる」までの変換に留め、可能な場合は「意味上の編集」として「話があった」「話す」と変換している。

- 文が長くなる編集操作

  文編集作業は要約のための作業であるため、文を短縮することがその主な目的である。「敬体常体変換」も、その変換によって講演全体として表現が短かくなるために導入されている。しかし、「敬体常体変換」を行うことで他の口語的な表現との齟齬が生じるため、それを解決するために「口語文章語変換」を導入した。このため文編集の操作として、例えば「手紙書く」→「手紙を書く」や「してる」→「している」など、自然な文章に近づけるために表現を補い、結果として文がより長くなる操作が含まれている。

- 話者の癖の識別

  文頭の「で」、文末の「けれども」、文中の「ですね」など、話者の癖と考えられる表現が講演の中には多く現れるが、一方でそれらが本来の意味をもって使われている場合もある。今回作成した文編集データでは、話者の癖かどうかは考慮されているが、「口語文章語変換」の一部として処理されている。例えば、文末の「けれども」や文頭の「で」などの表現は、(1) 話者の癖とみなせる場合は、「口語文章語変換」として削除する。(2) 接続の意味をもつ場合は、「で」の場合はそのまま、「けれども」は短縮のために「口語文章語変換」として「が」に変換する。(3) 接続の意味をもつが、重要文抽出の結果前後がつながらなくなった場合には、「重要文抽出による齟齬の修正」として削除する。

参考文献

[1] 前川喜久雄. 『日本語話し言葉コーパス』の概観 (overview.pdf).

[2] 菊池英明, 塚原渉, 小町守, 山田篤, 高梨克也.『日本語話し言葉コーパス』xml 文書について (xml.pdf).

[3] 野畑周, 高梨克也, 内元清貴, 井佐原均. 『日本語話し言葉コーパス』における自由要約・重要文抽出データについて (summarydata.pdf).

[4] 高梨克也, 内元清貴, 丸山岳彦. 『日本語話し言葉コーパス』における節単位認定 (clause.pdf).

[5] 内元清貴, 高岡一馬, 野畑周, 山田篤, 関根聡, 井佐原均. 『日本語話し言葉コーパス』への形態素情報付与. 第 3 回 話し言葉の科学と工学ワークショップ, pp. 39–46, 2 月 2004.

[6] 小磯花絵, 間淵洋子, 西川賢哉, 斉藤美紀, 前川喜久雄. 転記テキストの仕様 (transcription.pdf).

[7] 菊地康人. 敬語. 角川書店, 1994.

付録

[3] で示した、要約データ作成の対象である講演ファイル名一覧を以下に再掲する。全 199 講演のうち、「コア」に含まれるものは 177 講演である。それ以外の 22 講演 (コア以外のテストセット) は表中では斜体で示している。

講演ファイル名 (199 講演)				
A01F0055	A03F0153	S00F0177	S02F0129	S04F1495
A01F0067	A03M0004	S00F0197	S02F0180	S05F0463
A01F0122	A03M0005	S00F0209	S02F0189	S05F1041
A01F0132	A03M0010	S00F0210	S02F0852	S05F1517
A01F0143	A03M0018	S00M0025	S02M0011	S05F1600
A01F0145	A03M0045	S00M0053	S02M0043	S05M0412
A01M0007	A03M0059	S00M0065	S02M0068	S05M0613
A01M0015	A03M0061	S00M0071	S02M0076	S05M1236
A01M0020	A03M0138	S00M0075	S02M0092	S05M1505
A01M0021	A04M0026	S00M0112	S02M0103	S05M1666
A01M0025	A04M0047	S00M0115	S02M0161	S06F0167
A01M0030	A05F0039	S00M0117	S02M0191	S06F1034
A01M0048	A05F0043	S00M0153	S02M0198	S06F1566
A01M0056	A05F0154	S00M0199	S02M0245	S06M0373
A01M0065	A05F0502	S00M0213	S02M1698	S06M0894
A01M0070	A05M0002	S00M0218	S03F0062	S06M0895
A01M0074	A05M0031	S00M0221	S03F0072	S07M0833
A01M0083	A05M0040	S01F0006	S03F0108	
A01M0096	A05M0068	S01F0038	S03F0119	*A01F0001*
A01M0097	A06F0028	S01F0050	S03F0133	*A01F0034*
A01M0099	A06F0049	S01F0074	S03F0184	*A01F0063*
A01M0103	A06F0073	S01F0151	S03F0214	*A01M0141*
A01M0110	A06F0075	S01F0157	S03F0224	*A02M0012*
A01M0115	A06F0120	S01F0166	S03F0232	*A03M0016*
A01M0131	A06F0128	S01F0183	S03F0314	*A03M0106*
A01M0133	A06M0092	S01F1522	S03F0383	*A03M0112*
A01M0137	A07F0844	S01M0005	S03F1477	*A03M0156*
A01M0140	A11M0369	S01M0051	S03F1577	*A04M0051*
A01M0142	A11M0469	S01M0091	S03M0003	*A04M0121*
A01M0147	A11M0846	S01M0101	S03M0046	*A04M0123*
A01M0157	S00F0014	S01M0182	S03M0089	*A05M0011*
A02F0038	S00F0031	S01M0205	S03M0098	*A06F0135*
A02F0082	S00F0041	S01M0225	S03M0106	*A06M0064*
A02F0116	S00F0066	S01M0227	S03M0141	*S00F0019*
A02M0076	S00F0082	S01M0706	S03M0194	*S00F0148*
A02M0098	S00F0083	S02F0012	S03M0201	*S00F0152*
A02M0107	S00F0088	S02F0094	S03M0317	*S00M0008*
A03F0072	S00F0131	S02F0100	S03M0996	*S00M0070*
A03F0108	S00F0134	S02F0113	S03M1133	*S00M0079*
A03F0109	S00F0173	S02F0121	S04F0013	*S01F0105*