

議会会議録における整文過程の分析

-- 衆議院と欧州議会の比較及び過去十年の変化 --

河原達也（京都大学）

1. はじめに

会議録作成は音声テキストに変換するもので、2つの要件を満たす必要がある。一つは正確性、つまり音声に忠実であることで、もう一つは可読性、つまり読んで理解しやすいことである。この2つはしばしば相反する関係にあり、議会では一定の基準（明文化されていないものも含む）に則って会議録が作成される。ただし、この基準は言語や国によって異なるし、時代による変遷もある。本稿では、会議録作成の過程で行われる整文について、衆議院と欧州議会の比較及び衆議院における過去十年の変化に関する分析を報告する。

2. 会議録作成における整文

会議録作成の過程において様々な要因により整文が行われる。まず、言い淀みは除く必要があるし、他の冗長な部分も除去される。次に、文法的な誤りが修正される。いくつかの口語的表現が改まった表現に言い換えられることもある。さらに音声では文の区切りが明確でないので、句読点を挿入する必要がある。これらに加えて、可読性をよくするために倒置を入れ替えるなどの構造的な整文が行われることもある。さらに、明らかな言い誤りを訂正する場合もあるが、これは内容にも関わることなので注意を要する。これらについて以下に説明する。

2.1. 冗長箇所の削除

「えーと」や「あの一」などのフィラーは例外なく削除される。速記者はこれらをそもそも書き取らないし、音声認識システムでもこれらは自動的に削除される。繰り返しや言い直しなども削除されるが、これらを自動的に削除するのは容易でない。

「まあ」や「～ですね」などの談話標識はそのままにしておいてもよいが、あまり多いと可読性を損なう。「議長」などの呼びかけもそのままにしておいてもよいが、削除した方が可読性がよい。

2.2. 文法的誤りや口語的表現の修正

いくつかの文法的誤り、特に助詞や助動詞の脱落は修正する必要がある。例えば、「これする」は「これをする」に、「してる」は「している」に各々修正される。また、「じゃ」を「では」に言い換えるなど、口語的表現を改まった表現に修正する場合もある。ただしこのような修正は時代によって変遷している。方言の扱いも問題となる。方言は多くの人に理解されない場合もある反面、話者のアイデンティティを示すために使われる場合もある。

2.3. 構造的・意味的な修正

倒置表現を入れ替えたり、長い文を区切ったり、文の構造的な修正が行われる場合もある。これらは文章の推敲が必要となる。

一方、意味的な修正は注意が必要である。「億」と「兆」の違いなど明らかな言い間違いは一定の手順を経て修正されるが、固有名詞や事実関係の違いを修正してよいかは議論が分かれる。特に言い間違いが後々のやりとりに影響を与えた場合には修正すべきでない。

3. 衆議院と欧州議会におけるコーパス分析

3.1. コーパス

衆議院と欧州議会の会議録のコーパスを用いて、整文に関する分析を行った。欧州議会の方は、2007年の本会議のいくつかのセッションから英語で発言された部分を用いた。衆議院の方は、2005～2007年に行われた各委員会から複数のセッションを用いた。公式の会議録に加えて、元の発言音声からフィラーを含めて忠実に書き起こしたものも用意した。この忠実な書き起こしは音声認識システムの研究開発のために作成されたものである。これらのコーパスの仕様を表1に示す。

3.2. 整文カテゴリの分析

前節で述べた整文の各カテゴリの統計量を表2に示す。衆議院会議録における整文の大半はフィラーと談話標識の削除であるのに対して、英語の会議録作成では文法的誤りの訂正や文の構造の入れ替えなどが多数行われていることがわかる。このように言語による差は顕著といえる。

表 1 コーパスの仕様

	欧州議会	衆議院
忠実な書き起こしの単語数	30.9K	418K
公式の会議録の単語数	27.1K	379K
整文された単語の割合	20.5%	12.9%

表 2 整文カテゴリの統計量

		欧州議会	衆議院
削除	フィラー	11.6%	46.7%
	繰り返し／言い直し	11.0%	9.4%
	談話標識	1.8%	18.4%
	その他の冗長表現	16.8%	3.0%
修正	文法的誤り	20.1%	7.5%
	口語的表現	18.0%	8.4%
構造的整文		19.6%	5.9%

衆議院の会議録でフィラーと談話標識以外で顕著に削除されていたのは、「委員長」という呼びかけであり、挿入されていたのは「い」「を」「は」「と」「が」などの助詞・助動詞である。修正の頻度が多かったパターンは、「けども→けれども」「やっぱり→やはり」「じゃ→では」「いろんな→いろいろな」「ていう→という」である。

3.3. 会議のカテゴリ及び時間的変遷の分析

衆議院における委員会毎の整文の頻度割合を図1に示す。本会議以外では、憲法調査会において整文が顕著に少ない。憲法調査会は質問者と答弁者の1対1のやりとりでなく自由討論形式であるので、逐語的に記録される傾向が強いと考えられる。それ以外では、テレビ中継がされる予算委員会は少ない傾向にあるなど若干の差はあるが、顕著な違いは認められない。

一方、2007年と2016年を比較すると大きな差がある。すなわち、この10年間で整文の割合が40%も減少している。

4. 考察

整文が減少している原因はいくつかある。最も顕著なのは、置換などの文構造の入れ替えがほとんど行われなくなったことである。談話標識も削除されなくなる傾向にあり、繰り返しも強調などで用いられる場合はそのままにされるようになった。さらに、口語的表現の多くが許されるようになった。これらから、会議録がより逐語的になっていることが示唆される。

この傾向の理由として以下が考えられる。まず、音声認識システムの導入である。2011年度からすべての会議で音声認識システムが導入され、原稿作成者は認識誤りを含む忠実な書き起こしを整文するようになった。それ以前は速記者がテキストを入力していたので、その段階で自ずと整文が行われていたと考えられる。次に、審議映像のインターネット配信の影響も考えられる。すべての

会議が動画配信され、またアーカイブも常にアクセスできるようになり、ソーシャルメディアで参照できるようになった。これにより、校閲者の基準も変化したのではないかと推察される。ただし衆議院では、明文化された整文の基準はないとのことである。

5. 音声認識精度の変遷

衆議院では2011年度から音声認識システムを用いた会議録作成を運用している。議員などの発言者の音声を直接認識するシステムの導入は国会レベルでは世界に先駆けるもので、現在も他に例はないと思われる。導入後も新しい語彙や話題に対応するために単語辞書と言語モデルを毎年更新している。音響モデルは総選挙の後の審議音声を用いて更新している。

音声認識精度は継続的に計測している。2011年度の導入時は文字単位の正解率で89.7%であったが、その後の更新によって改善し、2012年以降は91%前後で安定している。ただし最近話題の深層学習（ディープラーニング）を用いることにより、3-4%程度改善することを著者の研究室で確認している。

6. まとめ

コーパス分析により以下のことが明らかになった。

- (1) 議会会議録では10-20%の単語で整文が行われている。
- (2) 日本語と英語では整文の傾向がかなり異なる。
- (3) 過去10年において整文が大幅に減少した。

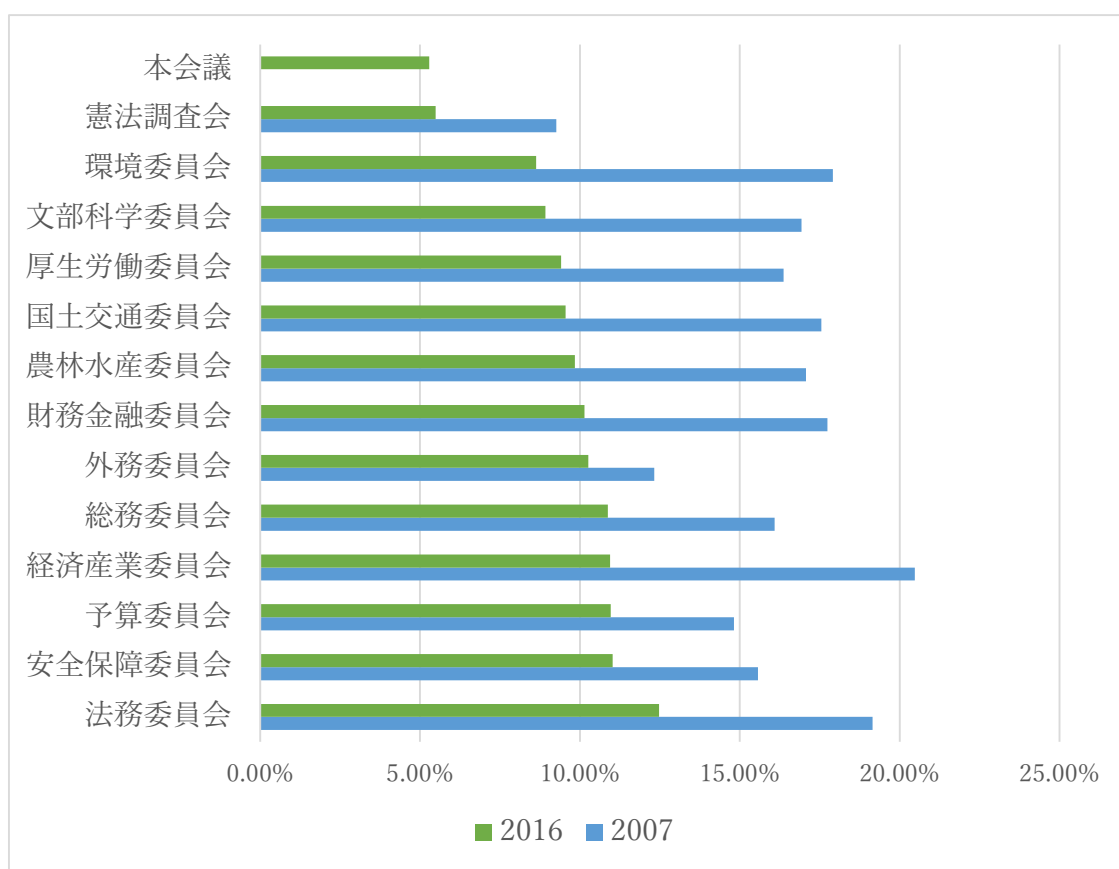


図1 2007年と2016年の委員会毎の整文の割合