日英ニュース原稿の対訳コーパス化に関する基礎調査

熊野 正　　田中 英輝　　金 淵培　　浦谷 則好

NHK 放送技術研究所

1. はじめに

NHK では，日本語による日々の放送を行うための大量の原稿を作成している．また英語による放送も一部実施しており，これに伴って英語の原稿も作成している．日本語の原稿は電子的に作成しており，データベースとして蓄積している．一方最近になって英語の原稿も電子的に作成・蓄積する体制が整いつつある．

英語の原稿の多くは日本語の原稿を基にして作成しているため，ある放送素材について日英両言語の原稿が存在するものがかなりある．この両原稿間に何らかの対応づけを行っていくことによって，一種の日英対訳コーパスを構築できる可能性がある．我々は日本語のニュース原稿から英語のニュース原稿を作成する現場の翻訳作業を支援することを目的に「類似用例提示型翻訳支援システム」の研究・開発を進めており，対訳コーパスの構築はその中核に位置する．このシステムは IBM Translation Manager/2 [1] や TRADOS Translator's Workbench[1] などに類似するが，対象がマニュアルのような定型的な文章ではなく多様な表現を持つニュース原稿であり，対訳コーパスの構築などの面において従来のシステムとは別種の問題が生じることが予想される．

本稿では，対訳コーパスの基となる日英それぞれの原稿データベースの特徴について述べ，対訳コーパス化のための基礎的な調査を行った結果について報告する．

2. 原稿データベース

2.1. 日本語データベース

NHK では，テレビ・ラジオのニュースでアナウンサが読む「読み原稿」を「汎用原稿」と呼ばれる中間的な記事原稿から作成している．汎用原稿は取材記者が作成・入稿したもので，責任者のチェックを経てさまざまな目的に利用される．

各々の汎用原稿は，見出し情報と記事本文（以後単に「記事」と呼ぶ）からなっており，見出し情報は記事を管理・特定するのに用いられる記事番号を含んでいる．

我々はこの汎用原稿に若干の計算機処理を加え，研究用の日本語原稿データベース（以後単に「日本語データベース」と呼ぶ）として蓄積している．日本語データベースは 1992 年 7 月から蓄積を開始し，現在までに約 7 万記事（約 40 万文）が収集されている．1995 年 4 月にデータベースに登録された日本語の記事の分量を 表 1 に示す．

1 日の記事数	118.4
1 記事の文数	5.2
1 文の文字数	88.9

表 1:　1 日あたりの日本語記事の分量
（1995 年 4 月の平均）

2.2. 英語データベース

総合テレビ・衛星放送などのニュースの副音声やラジオジャパンの英語放送などでは，NHK が取材制作したニュースを英語で放送している．これらの放送の読み原稿の多くは日本語の汎用原稿を翻訳して作成している．

日本語の汎用原稿から英語の読み原稿を作成するには，およそ以下のような手順を経る．

1. 日本人ライタが日本語汎用原稿を基に 1 次原稿を作成
2. 英語を母国語とするリライタが訂正を行い，2 次原稿を作成
3. 責任者による確認・訂正を行い，最終原稿を作成

この一連の作業はワードプロセッサを用いて行っており，全ての段階の原稿は電子的に蓄積されている．

"Primary Investigation on NHK's Japanese and English News Database"
KUMANO Tadashi (kumano@strl.nhk.or.jp),
TANAKA Hideki, KIM Yeun-Bae,
URATANI Noriyoshi
NHK Science and Technical Research Laboratories
1-10-11 Kinuta, Setagaya-ku, Tokyo, JAPAN 157

[1] cf. http://www.trados.com/tww.htm

一般に，1次原稿から最終原稿に近付くにつれて表現や文章構造が日本語原稿から離れてより英語らしくなる．しかし1次原稿のライタも十分に英語の記事を書く訓練を受けた人達であるので，1次原稿の段階でも決して直訳調で不自然な英語というわけではない．

日本語の場合と同様に英語原稿も見出し情報と記事本文からなっており，日本語汎用原稿を基に作成した原稿の多くは見出し情報の中に基となった日本語汎用原稿と同じ記事番号を持つ．

これらの英語原稿に日本語の場合と同様の若干の計算機処理を加えたものを研究用英語原稿データベース（以下単に「英語データベース」）として蓄積している．英語データベースの蓄積は1995年3月より開始し，現在までに約3万文を収集している．1995年4月にデータベースに登録された英語記事（最終原稿）の分量を 表 2 に示す．

1日の記事数	46.6
1記事の文数	7.5
1文の単語数	20.2

表 2: 1日あたりの英語記事の分量
（1995年4月の平均）

3. 対訳コーパス構築のための基礎調査

実際にこの日英のデータベースから日英対訳コーパスを構築するためには，両者の間に何らかの対応づけを行わなければならない．対応づけを行う単位としては

- 記事単位
- 文単位
- より細かい単位（節や句に相当するもの，...）

などが考えられ，より細かな単位での対応づけがとれている対訳コーパスの方が一般的に有用性が高い．

しかしこの日英データベースは対訳コーパス構築を意図して作成しているものではないため，実際に対応づけを行おうとするには不都合な性質が存在する．

3.1. 記事単位での対応関係

英語原稿は基本的には日本語汎用原稿に基づいて作成されており，基となった日本語原稿の記事番号がただ1つ示されていることが期待される．しかし実際には

- 英語原稿作成者の失念により記事番号が欠けている[2]
- 複数の日本語記事を参考に英語原稿を作成したため，基となる日本語記事の記事番号を特定できず，記事番号を記述していなかったり，複数の記事番号を記していたりする

などの理由により，必ずしも全ての英語原稿について対応する日本語原稿が一意に定まるとは限らない（一意に定まる英語原稿は全体の約9割; 表 3 参照）．実際に記事単位での対応づけを行うときには，「複数の日本語記事が1つの英語記事に対応している」という関係を認めた方がよいと思われる．

日本語記事数	英語記事数（最終原稿）	対応がとれた数
3553	714	637 (89%)

表 3: 記事番号による1対1対応がとれた記事数 （1995年4月分）

3.2. 文単位での対応関係

前節 表 3 に挙げた1対1対応がとれた日英の記事（各637記事）について，日英それぞれの1記事が含む文の数の分布を 図 1(a) に示す．また日本語記事を文数が少ないもの（3文以下），中間のもの（4〜6文），多いもの（7文以上）に分類し，それぞれの分類中の日本語記事に対応する英語記事の文数の分布を 図 1(b) に示す．

この結果から以下のようなことが考察できる．

- 図 1(a) より，英語の方が1記事あたりの文数が多い傾向があることが分かる（日本語記事の平均文数: 5.1文，英語記事の平均文数: 7.6文）．このことから，日本語記事中の1文が英語では複数の文で表現されていることが多いと予想できる．
- 英語の原稿は，日本についての知識が少ない人が聞くことを意図して作成している．たとえ元の日本語の記事が多くの文数からなっていても，背景知識の少ない人が関心を持たないような情報（例えば具体的な人名や詳細な地名など）は省略することが多い．逆に背景知識がないと理

[2] また，入力ミスにより誤った記事番号がつけられている原稿も存在する．誤った記事番号の多くは相当する日本語記事が存在しないため，単に対応がつけられない以上の問題は発生しないが，中には関係のない記事を指す記事番号に化けてしまうものもあり，無視できない問題となっている．

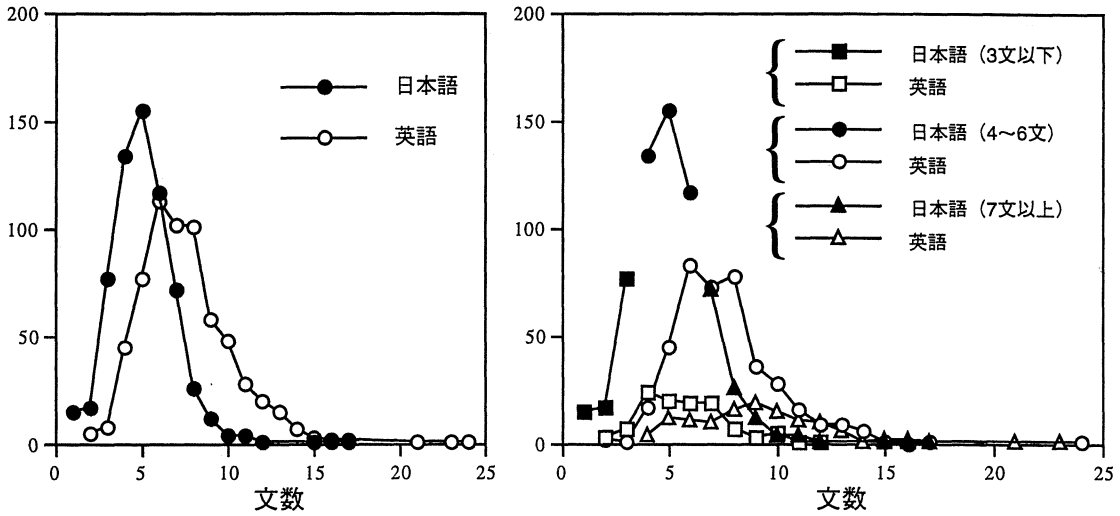

図 1: 日英記事の 1 記事中の文数の分類

解が難しいニュースでは，元の日本語記事には書かれていないような補足説明を加えている．

このため，図 1(b) から読み取れるように，文数の少ない日本語記事に対応する英語記事の文数が必ずしも少ないとは限らない．また逆も同様である．

3.3. 実際の対応関係の調査

前節での考察を確認するため，また実際に対応する日英原稿に対してどの程度表現の対応づけを行うことができるかの可能性を調べるため，人手で日英記事間の表現の対応づけを行ってみた．対象としたのは，前節で用いた 1995 年 4 月の 1 対 1 対応のとれた日英記事のうち，日本語が 5 文のものである．英語側の文数の違いによる対応関係の変化を調べるために，英語の文数が

- 4 文（文数が減少した例）
- 5 文（文数が同じ例）
- 6,7 文（英語側の最も頻度が高い文数の例）
- 9,10 文（文数が増加した例）
- 24 文（文数が非常に増加した例）

となる組合せを各々数例づつ選んで作業を行った．

対応づけは基本的には日本語の「句」もしくは「句の連なり」を単位として行い，日本語中のある句もしくは句の連なりと同じ表現が英語においても同じ文脈中でまとまって表出されている場合に，その組合せに対応関係をマークした．「同じ表現」であるという認定は元の表現の意味を忠実に訳出している場合に行い，例えば

- 「今日」 → 「○○日」,「×曜日」[3]
- 「○○円」 → 「××ドル」

といった言い換えなどに対しては行わなかった．

またこの対応づけ作業とは別に，日本語・英語それぞれの記事中で，相手の記事中には出てこない意味・内容を表す表現を特定する作業も併せて行った．

実際の作業例を 図 2 に示す．

作業を行った組合せのうち 5 文 → 5〜7 文 の 7 組について，対応づけの被覆率（相手側の記事中の表現に対応づけを行うことができた表現の割合）と非訳出率（相手側の記事中に出てこない意味・内容の表現の割合）を算出した結果を 表 4 に示す．

	日本語側（文字数で計算）	英語側（単語数で計算）
被覆率	47.5%	67.9%
非訳出率	11.1%	7.1%

表 4: 対応づけの結果
（5 文 → 5〜7 文 の 7 組の平均）

[3] 英語原稿は日本語原稿が放送に使われる日とは異なる日に使用されることも多く，相対的な日時の表現はあまり用いられない．

s|韓国南部で今朝、列車とバスが衝突して少なくとも十三人が死亡、十五人がけがをしました。
s|事故があったのは韓国南部、チョルラ（全羅）南道のファスン（和順）にある鉄道の踏み切りで列車とバスが衝突しました。
s|この事故でこれまでに少なくとも十三人が死亡した他、十五人がけがをしました。
s|現地の警察当局は、バスの乗客の中に重体の人が何人かいることから死者の数は更に増えるのではないかと見ています。
s|このバスには通学途中の生徒達が大勢乗り込んでいたという情報もあり警察で事故の原因など詳しいことを調べています。

s|At least 14 people died and 20 others were injured when a train and bus collided Tuesday morning in South Korea.
s|The accident occurred at a railway crossing in the town of Hwasun, near central Jeonra Nam Do, in the southern part of the country.
s|Local police said the death toll is likely to rise as some passengers are very seriously injured.
s|Many of about 40 passengers were students on their way to school.

英語側にはない情報

図 2: 対応づけの例

調査結果から，以下のような傾向を読みとることができる．

- 5 文 → 24 文 のように日英の文数が極端に異なっている場合には，記事に含まれている情報の量や情報提示の順序が全く異なることが多い．調査例では，日本語では記事の各所に散在していた情報を英語ではまとめて箇条書的に列挙していたり，日本にいる人には明白なため日本語では省略されているような事柄を英語では補足説明していたりした．
- それ以外（5 文 → 4〜10 文）では記事の情報提示の順序の入れ換わりはあまり見られない．日本語に対する英語の内容の増減は，主として元の日本語の情報提示の順序に対して省略や挿入の形で起こる．

4. まとめと今後の課題

本調査の結果，我々が収集している日英原稿データベースは以下のような性質を持っていることが分かった．

- 英語原稿のほとんどは対応する日本語原稿を特定可能
- 対応する日英記事の文数があまり離れていないときには日英の情報提示の順序が変化していないことが多い．

このことから，我々の持つ日英原稿データベースから対訳コーパスを構築することは困難ながらも不可能ではないという感触を得た．今後は日本語が 5 文以外の記事についても対応関係の調査を行い，より対応づけられたコーパスを自動的に構築するにはどのような日英記事の組合せを選べばよいかについての知見を得る必要がある．

また我々が持つ英語データベースでは，1 つの記事について最終原稿以外に，より元の日本語原稿と類似した 1 次原稿や 2 次原稿を持つものも多い．これを利用して日本語原稿と最終原稿との対応づけの精度を高める手法について検討していきたい．

また我々が研究・開発を進めている「類似用例提示型翻訳支援システム」の用例知識としてこのコーパスを用いる場合，対応づけを行う単位はどの程度の大きさが適切であるかについて考察を行わなければならない．単に入力文に対して類似の日英の文の組合せを提示するだけならば対応づけの単位は文程度でよいのだが，翻訳現場の需要はそのような単純なものではなく，対訳コーパスからはさまざまな粒度の情報を取り出せることが重要である．これらの点をふまえて日英対応関係を実際に付与していきたい．

参考文献

[1] Jean-Marc Langé. MT at IBM France: Research and Products. In *Premières journées franco-japonaises sur la traduction assistée par ordinateur*, pp. 205–207, 1993.