

日本語の数量表現とその英語への機械翻訳に関する一考察

吉武 淳

三菱電機(株)コンピュータ製作所

高山 泰博

鈴木 克志

太細 孝

三菱電機(株)情報電子研究所

日英機械翻訳システムMELTRAN-J/E において数量表現処理を実装した。本報告では、そこでの考察にもとづき、下記項目について議論する。

- (1) 日本語の数量表現について、文節レベル、構文レベルそれぞれでの新しい考察
- (2) その英語への機械翻訳の際の注意点
- (3) 通常と少し異なる観点から整理されたシソーラスが必要ではないかということ
- (4) 数量数量詞、属性数量詞についての考察

本報告の方法を拡張していくことで、数量表現の翻訳能力の向上が期待できる。

A View on Japanese Quantifiers and their Machine Translation into English

Jun Yoshitake

Computer Works, Mitsubishi Electric Corporation
325, Kamimachiya, Kamakura, Kanagawa, 247 Japan

Yasuhiro Takayama

Katsushi Suzuki

Takashi Dasai

Information Systems and Electronics Development Laboratory,
Mitsubishi Electric Corporation

5-1-1, Ofuna, Kamakura, Kanagawa, 247 Japan

A quantifiers translation program is implemented in our Japanese-English machine translation system, MELTRAN-J/E. Through its implementation, some new ideas have been acquired:

- (1) some new viewpoints on Japanese quantifiers both in the 'bunsetsu' segment level and in the syntactic level
- (2) points to pay attention to in their translation into English
- (3) the thesaurus which is organized in the viewpoint different from a usual one may be needed for quantifiers processing.
- (4) views on quantitative quantifiers and attributive quantifiers

These are discussed in this paper.

1 はじめに

数量表現の翻訳は、技術文書、マニュアル 類の翻訳の 1 つのポイント である。その機械処理については、今までも 1 文中で局所的に処理する研究^{1), 3)} などが行われてきたが、まだ十分とはいえないのではないだろうか。

言語学において数量詞の研究はさかんになされてきた^{4), 7), 8)} が、それらの多くは数量詞そのものよりも数量詞の 1 文中における移動についてのものであり、言語の機械処理という点からは十分役に立つといえるものではなかった。

このような中で、日英機械翻訳システム MELTRAN-J/E²⁾ (以下本報告では MELTRAN と呼ぶ) での数量表現の実装を通して、日本語の数量表現とその英語への機械翻訳にあたって新しい考察が得られたので報告する。

以下本報告では第 2 章で日本語の数量表現についての新しい考察について述べ、第 3 章ではそれらの英語への機械翻訳の際の注意点について説明する。第 4 章ではシソーラス や数量数値詞、属性数量詞に関して述べ、第 5 章でまとめと今後の課題について述べる。

なお MELTRAN はトランスファ方式の機械翻訳システムであり、人工知能用ワークステーション MELCOM PSI II 上で開発されている。本報告の数量表現処理ももちろんそこで動作している。

2 日本語の数量表現

2.1 文節レベルでの日本語の数量表現

田中¹⁰⁾は、文節レベルの日本語の数量表現は図 2.1-1 の形式で記述できるとしている。^{*}

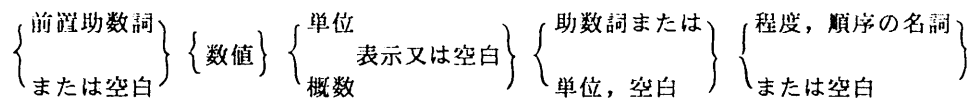


図 2.1-1 田中¹⁰⁾による数量表現

これを BNF 風の表現になおすと図 2.1-2 のようになる。

```
(数量表現) ::= [前置助数詞] (数値) [単位概数表示] [助数詞単位]
               [程度, 順序の名詞]
(単位概数表示) ::= (単位表示) | (概数表示)
(助数詞単位) ::= (助数詞) | (単位)
```

図 2.1-2 図 2.1-1 の BNF 風表現

ここで [] は省略可能であることを示す。

また、天野ら¹⁾は本節でいう文節レベルの数量表現に相当するものを「数字文節」と呼んでその構造を示しているが、それを BNF 風の表現になおすと図 2.1-3 のようになる。

図 2.1-3 の「数字」、「後置助数詞」はそれぞれ図 2.1-2 の「数値」、「助数詞単位」にほぼ相当し、「数詞接辞」は、図 2.1-2 の「程度、順序の名詞」に含まれるものと考えられる。

図 2.1-4 に本報告での文節レベルの日本語の数量表現の定義の概略を示す。

田中による定式化や天野らの構造と本報告での定義との大きな違い (本報告での新しい考察) は次

*)もちろん文節レベルの日本語の数量表現の意味である。以下、2.1 節で単に「数量表現」と記述した場合はすべてこの意である。

(数字文節) ::= [前置助数詞] (数字) [後置助数詞列] [数詞接辞]
 (後置助数詞列) ::= (後置助数詞) [後置助数詞列]

図2.1-3 天野らによる数字文節の構造

(数量表現) ::= [数量接頭語] (数) [助数詞名詞列] [数量限定語]
 (助数詞名詞列) ::= (助数詞名詞) [助数詞名詞列]
 (助数詞名詞) ::= (助数詞) | (NumN型の名詞) | (区切り記号)

図2.1-4 本報告での数量表現の定義

の通りである。

- ① 田中による形式では助数詞、単位の繰返しが考慮されていないが、本報告での定義では十分考慮されている。

2人月、10Mbyte/sec

のように、助数詞や単位は容易に結びついて連なっていく性質を持っているのでこのような繰返しに対応できるようにしておくことは実用上重要である。

また、天野らによる構造では繰返しが考慮されているが、あとの③で述べるような区切り記号を明瞭には意識していない。

- ② 田中による形式でも天野らによる構造でも数量表現が助数詞なしで名詞を直接修飾する表現をあまり明瞭に意識していないが、本報告での定義でははっきりと意識している。助数詞なしで名詞を直接修飾する表現（以後NumN型の表現と呼び、直接修飾される名詞をNumN型の名詞と呼ぶ）とは、たとえば、

私は3文字書いた。

のような表現である。この時、「3」は直接「文字」を修飾している。

私は5本の[^]ンを持っている。

という表現で、「5」が「本」という助数詞を伴って「[^]ン」を修飾しているのとは違うわけである。表2.1-1にNumN型の名詞の例を示す（「3文字」という表現があっても「3[^]ン」という表現はないように、すべての名詞がNumN型というわけではない）。

表2.1-1 NumN型の名詞の例

球、票、文字、字、行、株、国、人、歩、店、点、面、辺

- ③ 田中による形式でも天野らによる構造でも区切り記号の存在をあまり意識していないが、本報告での定義でははっきりと意識している。区切り記号は、助数詞や単位と違ってそれ自体では意味をもたないので区別しておいた方がよい。

また、区切り記号も英語への翻訳の観点から2種類に区別される。たとえば"/"は、
 10000語/時 → 10000 words/hour

のように、英語への翻訳時、"/"の前の語は複数形になることがあるが、たとえば"・"（中黒）の前の語は複数形になることはない。

また、本報告の定義では、m やkgのような単位を示す英文字類が数に続いた場合、助数詞として切り出すこととし、それらを辞書には登録していない。これは、単位を示す英文字には実にいろいろな

もの・いろいろな組合せがあり、それらをすべて助数詞として辞書に登録しておくことは実質的に不可能だからである。

なお、本報告の定義で「数量接頭語」、「数」、「数量限定語」は、それぞれ田中による形式の「前置助数詞」、「数値」、「程度、順序の名詞」にほぼ対応する。

2.2 構文レベルでの日本語の数量表現

ここでは、名詞修飾の日本語の数量表現について述べる。それには、以下のような多様性がある。奥津⁷⁾は、名詞修飾の日本語の数量表現には次の3つがあるとしている。

- 舞台に主役3人が現われた。 (NQC型) (i)
舞台に主役が3人現われた。 (NCQ型) (ii)
舞台に3人の主役が現われた。 (Q/N型) (iii)

ここでNは名詞、Qは数量詞(助数詞付きの数量表現)、Cは格助詞を表す。

また、柴谷⁸⁾はこの他に次のような表現をあげている。

- 舞台に3人主役が現われた。 (QNC型) (iv)

しかし、名詞修飾の日本語の数量表現にはこれら以外に次のようなものもある。

- 舞台に主役の3人が現われた。 (N/Q型) (v)

さらに、2.1 節で述べたNumN型の表現がある。

(i)~(v)の文が同じことを意味することを、奥津⁷⁾の議論の一部を借りて、念のために少し厳密に確かめておく。

まず、いずれの文においてもNは不定である。すなわち、たとえば英語に翻訳した時に定冠詞やそれに相当する語句はつかない。

また、いずれの文においてもNとQとは同格である。すなわち、主役 = 3人 という関係が成り立つ。

このように(i)~(v)の文は同じことを意味しているわけである。

2.3 数量表現と名詞との修飾関係の解析

2.2 節で述べたように構文レベルでの日本語の名詞修飾の表現には少なくとも6つのものがある。

MELTRAN ではこれらについて以下のように修飾関係を解析している。

Nの直後にQがあればNQC型とする。

N, C, Qという語の並びがあった場合、NまたはNの意味コトが持つ、それがとり得る数量数量詞(4.2 節に詳述)のリストにQの助数詞が含まれるならばNCQ型とする。こうしているのは、

- 学生が3人来た。 (NCQ型)
私が3文字書いた。 (NCQ型ではなく、NumN型)

などの区別が難かしいからである。

Q, 「の」, Nという語の並びがあった場合、Q/N型とする。

QNC型の解析はNCQ型と同様である。

N/Q型の解析は未実装であり、今後よく検討する必要がある。

NumN型の場合は、一度Qとして解析した後、その前後に修飾すべき名詞がなく、Q中の助数詞がNumN型の名詞になり得るものならば、NumN型と解析される。

3 内部構造を英語へ変換する際の注意点

2章で述べた解析の結果、日本語の内部構造(MELTRANでの内部表現)が生成されるがそれを英語へ変換する時の注意点は、

- ・ 英語で訳出される助数詞、訳出されない助数詞の区別
- ・ 助数詞の複数変化処理
- ・ 数量表現によって修飾される名詞の複数変化処理
- ・ 数量限定語の訳出
- ・ 冠詞の問題
- ・ ブラック挿入の問題

- ・ 助数詞にも、数量表現によって修飾される名詞にもなるものの処理
- ・ a piece ofの類の訳出
- ・ その他

である。以下順番に説明する。

3.1 英語で訳出される助数詞、訳出されない助数詞の区別

「300 円の本」、「20m の橋」の「円」、「m」ように英語で訳出される助数詞は、そのまま“300 yen book”、“20m bridge”というように訳出している。

「5本のペン」、「10個のボール」のように訳出されない助数詞は、“5 pens”、“10 balls”というように訳出しないようにしている。

これらの区別は、助数詞の辞書項目に英語への訳出の可、不可の情報を持たせ、日本語の内部構造を英語の内部構造に変換する時にその情報をもとに助数詞を削除する、しないの処理を行っている。

3.2 助数詞の複数変化処理

英語に訳出される助数詞でも複数変化するものとししないものがある。

私はその本を500 円で買った。

→ I bought that book with 500 yen.

その計算機は2バイト でデータを転送する。

→ That computer transmits data at 2 bytes.

また、複数変化するものでも名詞修飾の場合は複数形にならない。

その計算機は2バイト のデータを転送する。

→ That computer transmits 2 byte data.

さらに、英語に訳出される助数詞が複数個ある場合は、

- ・ 区切り記号の“/” がある場合は、数に一番近い“/” の直前の助数詞を複数変化させる。

MELTRAN J/E は10k 語/時で文書を翻訳します。

→ MELTRAN J/E translates document at 10k words/hour.

- ・ “/” がない場合は、一番最後の助数詞を複数変化させる。

翻訳速度は10k 語だ。

→ The translation velocity is 10k words.

もちろんこの場合も名詞修飾の場合は、“/” がないにかかわらず複数変化させない。

MELTRAN J/E は10k 語/時の機械翻訳システムです。

→ MELTRAN J/E is 10k word/hour machine translation system.

3.3 数量表現によって修飾される名詞の複数変化処理

複数形になるものはそのように変化させる。複合名詞の場合は一番最後のものを複数変化させる。

私は普通名詞を2個書いた。

→ I wrote 2 common nouns.

複数変化するものでも、その前に助数詞が訳出されている場合は複数変化させない。

私は500 円の本を持っている。

→ I have 500 yen book.

3.4 数量限定語の訳出

助数詞の“not” が訳出される場合、一番最後にくる助数詞の後に訳出する。

MELTRAN J/E は10k 語/時以上の機械翻訳システムです。

→ MELTRAN J/E is 10k word/hour or above machine translation system.

助数詞の“not” が訳出されない場合で、名詞修飾の場合はその名詞の後に訳出する。

私はペンを5本以上持っている。

→ I have 5 pens or above.

助数詞の“not” が訳出されない場合で、名詞以外のものを修飾する場合は数の後に訳出する。

数は5以上だ。

→ The number is 5 or above.

なお、「以上」「以下」等の訳を“or above”, “or less” などとしているが、これは“more than”のような比較級にするよりも簡便な方法である（「10以上」は“more than 9”のように数を1ひかなければならないことにも注意）。

3.5 冠詞の問題

内部構造の英語への変換にあたり、名詞に複数であるという情報が設定されない場合は、通常単数形として出力の英文で冠詞が付けられる。しかし、助数詞が訳出される場合で名詞修飾の場合は、名詞に複数であるという情報が設定されないにもかかわらず、数のノットがあるので冠詞が訳出されないという弱点がいま実装してあるものにはある。

私は500 円の本を持っている。

→ I have 500 yen book.

3.6 フランク 挿入の問題

2.1 節で述べたm やkgのような単位を示す英文字類と、区切り記号の前後のみフランクを挿入しないようにしている（“5 kg / m”のように間延びさせず、“5kg/m”のようにするという）。このような注意も翻訳処理の際には必要である。

3.7 助数詞にも、数量表現によって修飾される名詞にもなるものの処理

たとえば、「本」は「500 円の本」（“500 yen book”）、「5本のペン」（“5 pens”）のように助数詞にも数量表現によって修飾される名詞にもなり得る。このようなものに対する注意も必要である。

いま実装しているものでは、文節内解析が終了した時点で「本」が助数詞として用いられる場合は内部構造にその情報を持たせることによって区別している。

また、単独で助数詞として用いられている時は削除される助数詞であっても、他の助数詞または数量表現によって修飾される名詞とともに用いられると、削除されない助数詞または数量表現によって修飾される名詞となるものがある。

たとえば「人」は「3人の学生」の時は“3 students”となって削除される助数詞である。しかし、「2人月」「5kg/人の米」の場合は、それぞれ“2 man month”、“5kg rice per person”のように「人」は訳出されるべきである（現在実装されているものはこれに対処できていない）。このような例は「80人/台」等実際に数多くあり、今後本格的に対処していく必要がある。

3.9 a piece ofの類の訳出

「4枚の紙」→“4 sheets of paper”のように英語の方で“a piece of”の類の表現を訳出しなければならないものがある。名詞の辞書項目に“a piece of”の類をとるかとらないか、とるとすればどのようなものをとるかの情報を設定し、出力の英文生成時にその情報を参照することで対処している。

3.10 その他

いま実装しているものでは、たとえば、

それは10kgだ。

→ “It is 10kg.”

になってしまう。より英語らしくは、“It weighs 10kg.であらう。また、

それは長さ10m だ。

→ It is 10 meters long.

など、本当に英語らしい英文を訳出するためにはまだまだ研究すべき点が多々ある。

4 シソーラス、数量数量詞、属性数量詞

4.1 シソーラス

計算機上のシソーラスについても実にさまざまな研究が行われてきている。⁹⁾ 田中¹⁰⁾も助数詞と対象物

の関係について簡単な考察を行っているが、ここでは数量表現の処理には通常のシリーズとは少し異なる視点で整理されたものが必要であることを指摘しておきたい。

いま、

5本の青色のペン

青色の5本のペン

という2つの文でどちらの文でも「5本の」という文節が「ペン」にかかるようにするには、「本」という助数詞が「ペン」という細長いものを数えるときに使われるという知識が必要である。

また、

3本の質の高い映画

質の高い3本の映画

のような文でも「本」という助数詞が「映画」というその物理的媒体を巻くことができるものを数えるときに使われるという知識が必要である。

このように考えていくと、たとえば図4.1-1 に示すようなシリーズが役立つ。

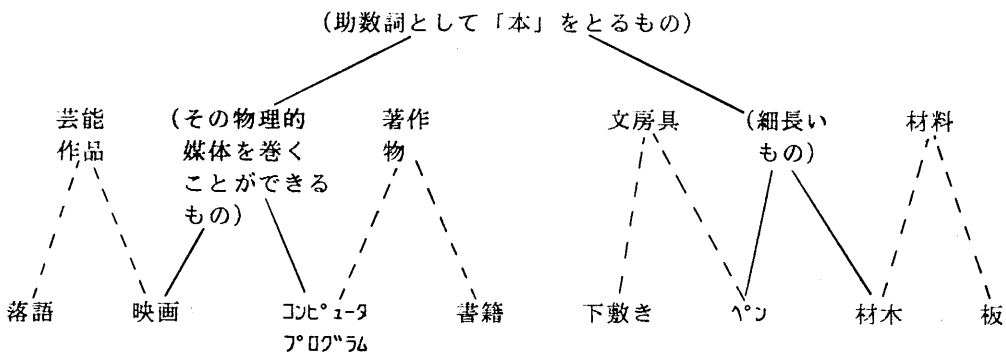


図4.1-1 数量表現に役立つように整理されたシリーズの例(参考文献5)を少し参考している)

図4.1-1 では実線が数量表現用の上下位関係、点線が通常のシリーズでの上下位関係である。「(助数詞として「本」をとるもの)」、「(その物理的媒体を巻くことができるもの)」、「(細長いもの)」という言い方は変かもしれないが、通常のシリーズとは少し異なる視点で体系的に整理されたものが将来必要ではないかと考えられる。

4.2 数量数量詞、属性数量詞

数量数量詞とは、

5本のペン、3冊の本

の「本」、「本」のように修飾する名詞の数量そのものを表すものである。

属性数量詞とは、

1000ccの車、400ページの書

の「cc」、「ページ」のように修飾する名詞の属性を表し、一般の修飾語句にすぎないものである。⁷⁾

これらの区別も機械翻訳での翻訳精度の向上に役立つ。たとえば、

- ① 上述の例からもわかるように、数量数量詞には英語に訳出されない助数詞が多い一方、属性数量詞には訳出されるものが多い。そして、属性数量詞が英語で訳出された場合、それ自身も修飾される名詞も単数形のままであることが多いと思われる。これは属性数量詞の性質から考えても当然のことと考えられる。
- ② 属性数量詞はQノN型しかとらない。たとえば、
1000ccの車を買った。

という言い方はしても、

車1000ccを買った。

という言い方はしない。⁷⁾ この知識は日本語の数量表現の解析の精度や速度の向上に役立つと考えられる。

- ③ 数量数量詞、属性数量詞両方が同時に使われた場合の翻訳精度の向上にも役立つ。

200cc のびん3本 → three 200cc bottles

などが例として挙げられる。

- ④ 1文中で1つの名詞に対して、それを修飾する数量数量詞は1つだが、属性数量詞は2つ以上となることがある。

図4.1-1 の例は、数量数量詞を対象に考えていたが、このように考えてくると属性数量詞についても同じようなものが必要になる可能性もある。

ただ、数量数量詞、属性数量詞を体系的に整理しようとするならば次のようなことに気をつけておく必要がある。

- ① ひとつの助数詞が場合によって数量数量詞、属性数量詞のどちらにもなり得る。たとえば、「2000ccの車」という時の「cc」は属性数量詞であるが、「牛乳150cc」という時の「cc」は数量数量詞である。
- ② 「1つ」「2つ」の「つ」のような、修飾する名詞との関係があまり深くないようなものをどうするか。
- ③ 属性数量詞はすべて整理しきれぬのか（たとえば、あらゆる具体物は重さを持つがその単位はまちまちである）。

5 おわりに

日英機械翻訳システムMELTRAN-J/E に実装した数量表現処理の特長、日本語の数量表現についての新しい考察、英語へ機械翻訳する際の注意点、ソラリス、数量数量詞、属性数量詞についての考察について述べてきた。

本報告ではふれなかったが、度数、回数を表す表現や部分数量の表現、外国語へ翻訳する時の単位の変換等、数量表現にはまだまだ研究、検討すべきことがたくさんある。また、製品番号等数を使用した表現に対象を広げるとさらに研究、検討すべきことがふえる。今後このような方面でも大規模な言語データの蓄積、研究が望まれる。

謝辞 本研究を行うにあたって、貴重な資料を気持ちよく提供して下さった姫路短期大学経営情報学
科田中康仁教授と、有益な御意見を下さった三菱電機情報電子研究所自然言語処理グループの皆さんに感謝します。

[参考文献]

- 1) 天野, 河田: かな漢字変換における局所意味分析, 昭和55年度電子通信学会総合全国大会, 1190 (1980).
- 2) 太細, 吉武他: 日英機械翻訳システム, 三菱電機技報, Vol.62, No.5, pp.427-430(1988).
- 3) 飯野, 亀井: 英語数量表現の解析, 第38回情報処理学会全国大会論文集, pp.394-395(1989).
- 4) 井上和子: 日英対照 日本語の文法規則, 大修館書店(1978).
- 5) 宮地, 辻, 金枝上: 知識処理応用のためのテラ-3A1-3利用環境について, 電子情報通信学会, DE88-31, pp.9-16(1989).
- 6) 日本工業規格: 単位記号の情報交換用表記方法 JIS X0124 (1981).
- 7) 奥津敬一郎: 数量詞移動再論, 人文学報160号, pp.1-24, 東京都立大学人文学部(1983).
- 8) 柴谷 方良: 日本語の分析, 大修館書店(1978).
- 9) 田中, 仁科: 上位/下位関係ソラリスISAMAP1の作成, 情報処理学会自然言語処理研究会, 64-4 (1987).
- 10) 田中康仁: 日本語の数表現の解析, 情報処理学会計算言語学研究会, 24-3(1980).
- 11) 富井 篤: 英語の決めて数量表現, インタープレス(1986).