自然言語処理　138－4
（2000. 7. 18）

構文情報の定量化とそれを用いた言語比較

高村 大也
奈良先端科学技術大学院大学
情報科学研究科自然言語処理学講座
〒 630-0101 奈良県生駒市高山町 8916-5
0743-72-5246,5248
hiroya-t@is.aist-nara.ac.jp

杉原 厚吉
東京大学
工学系研究科計数工学専攻
〒 113-8656 東京都文京区本郷 7 - 3 - 1
03-3812-2111 内線 6905
sugihara@simplex.t.u-tokyo.ac.jp

あらまし

語順や格変化など，文の構造を決定する情報（構文情報）を，情報理論の概念を用いて定量化する方法を提案する．さらに，その方法により独語と英語について具体的数値を算出し，両言語を比較した結果，独語では名詞句の格変化や動詞の活用が大きな構文情報を担い，一方英語では語順が支配的であることを定量的に示すことに成功した．

キーワード　エントロピー，語順，格変化，定量化，文構造

Quantification of Syntactic Information and its Application to Language Comparison

Hiroya Takamura
Nara Institute of Science and Technology
Graduate School of Information Science
8916-5 Takamaya, Ikoma, Nara 630-0101, JAPAN
+81-743-72-5246, 5248
hiroya-t@is.aist-nara.ac.jp

Kokichi Sugihara
Tokyo University
Dept. of Mathematical Engineering
7-3-1, Hongo, Bunkyo-ku, Tokyo, 113, Japan
+81-3-3812-2111 ext. 6905
sugihara@simplex.t.u-tokyo.ac.jp

Abstract

We propose a method to quantify syntactic information (word order, conjugation, etc) using information theory. We compute the actual values of various syntactic features of German and English. The result quantitatively shows that, for German, the declination of noun phrases and the conjugation of verbs contain a larger amount of syntactic information. However, for English, the word order contains larger information.

key words: entropy, word order, conjugation, declination, sentence structure

1 序論

人間の言語活動を情報伝達として捉えると, 伝達内容はアルファベット列あるいは音韻列としてコード化されているとみなすことができる. また, 一段階上の視点から, 単語列としてみなすことも可能である. このような考え方は Shannon による情報理論と共に誕生し, 発話をアルファベットのマルコフ過程とみなすなどして, その情報理論的考察がなされてきた [1, 2, 3, 4].

ここで視点を変えて, 文構造のコード化という考え方に着目してみよう. 例えば, "He has a book." という英語文がある. この文に含まれる文構造に関する情報（構文情報）には, 次のようなものがある.

- "He" の形態（格変化）は,「he」という個体がこの文の主語であることを示す.
- "has" の形態（活用）は, この文の主語は三人称単数であることを示す.
- "he" と "a book" の語順は,「he」という個体が主語,「a book」という個体が目的語であることを示す.

このような捉え方は, 構文情報が素性の集合としてコード化されていると考えることに相当する.

構文を知るための情報には, 語順が与える情報, 名詞句の形態（格変化や格助詞）が与える情報, 動詞の形態が与える情報, さらに意味論や文脈が与える情報などがある. これらの情報が実際に文構造に関与する度合は言語により異なるであろう. 例えば, 英語では, "The student has a book." では "The student" が主語だが, 語順を入れ換えた "A book has the student." では "A book" が主語となる. これは語順が文の構造を決定していることを示している. ところがドイツ語では, "Der Student hat ein Buch." という文の語順を入れ換えて, "Ein Buch hat der Student." としても依然として "der Student" が主語である. これは名詞句の形態が文構造を決定することを示している.

コーパスを利用してこれらを情報量という形で算出し, その言語が含む曖昧性や情報の冗長性を具体的数値として求めることにより, 種々の言語の特性を論じることができる. そのための方法を提案し, さらに英語とドイツ語に対して実際に数値計算をしてみようというのが本論文の主旨である.

2 文構造決定

2.1 文構造決定のモデル化

言語における表層格として, 主格, 直接目的格（対格）, 間接目的格（与格）, それにいくつか の前置詞格を考える（印欧言語を念頭においている. 日本語などを扱う場合はまた別の定式化が必要であろう）.

ここで, 文構造を決定するということを定義する.

定義 2.1 文構造を決定するとは, 文の動詞以外の各要素の表層格を同定することである.

動詞以外の要素の表層格が同定されれば, 動詞によって, 要素の深層格が同定される [5]. つまり, 文の意味が決定されるのである.

さて, 以下の解析のために文構造を単純化する. まず, 前置詞格についてはその形態から格が明らかなので, 全て無視する（存在しないものと考える）. また, 動詞は同定可能なものとし, それぞれの要素の境界も与えられるとしておく. 要するに, 動詞句と名詞句だけが与えられると考えるのである.

ある文が与えられ, その文を構成する n 個の名詞句を出現順序を保ったまま取り出したものを,

$$\mathbf{NP}_1, \mathbf{NP}_2, \cdots, \mathbf{NP}_n$$

とする. また, 可能な格が m 種類あり, それらを,

$$\mathbf{CASE}_1, \mathbf{CASE}_2, \cdots, \mathbf{CASE}_m$$

とする. ここでは, 主格 (**S**), 直接目的格 (**DO**), 間接目的格 (**IO**) の三つのみを扱うので,

$$\mathbf{CASE}_1 = \mathbf{S}, \mathbf{CASE}_2 = \mathbf{DO}, \mathbf{CASE}_3 = \mathbf{IO}$$

としておく.

本論文では, 次の二つが成り立っているものとする.

- 一名詞句一格の原則：一つの名詞句にはただ一つの格が対応する.
- 一格一名詞句の原則：一つの格にはただ一つの名詞句が対応する.

一格一名詞句の原則については [5] に言及されている. 実はこれには "Ich frage es ihn." のような例外が存在するが, ごく少数なので無視する.

ここで文中の名詞句と格との対応付けを格割当と呼んで, a_i $(1 \leq i \leq k)$ で表すことにする. k は格割当の種類の数である.

ここでは主格が必ず出現するような言語のみを対象にするので, 次の 11 種類の格割当が考えられる：

$$a_1 = \left[\begin{array}{ccc} \mathbf{NP_1} & : & \mathbf{S} \end{array} \right], a_2 = \left[\begin{array}{ccc} \mathbf{NP_1} & : & \mathbf{S} \\ \mathbf{NP_2} & : & \mathbf{DO} \end{array} \right],$$

$$a_3 = \left[\begin{array}{ccc} \mathbf{NP_1} & : & \mathbf{DO} \\ \mathbf{NP_2} & : & \mathbf{S} \end{array} \right], a_4 = \left[\begin{array}{ccc} \mathbf{NP_1} & : & \mathbf{S} \\ \mathbf{NP_2} & : & \mathbf{IO} \end{array} \right],$$

$$a_5 = \left[\begin{array}{ccc} \mathbf{NP_1} & : & \mathbf{IO} \\ \mathbf{NP_2} & : & \mathbf{S} \end{array} \right], a_6 = \left[\begin{array}{ccc} \mathbf{NP_1} & : & \mathbf{S} \\ \mathbf{NP_2} & : & \mathbf{DO} \\ \mathbf{NP_3} & : & \mathbf{IO} \end{array} \right],$$

$$a_7 = \left[\begin{array}{ccc} \mathbf{NP_1} & : & \mathbf{S} \\ \mathbf{NP_2} & : & \mathbf{IO} \\ \mathbf{NP_3} & : & \mathbf{DO} \end{array} \right], a_8 = \left[\begin{array}{ccc} \mathbf{NP_1} & : & \mathbf{DO} \\ \mathbf{NP_2} & : & \mathbf{S} \\ \mathbf{NP_3} & : & \mathbf{IO} \end{array} \right],$$

$$a_9 = \left[\begin{array}{ccc} \mathbf{NP_1} & : & \mathbf{DO} \\ \mathbf{NP_2} & : & \mathbf{IO} \\ \mathbf{NP_3} & : & \mathbf{S} \end{array} \right], a_{10} = \left[\begin{array}{ccc} \mathbf{NP_1} & : & \mathbf{IO} \\ \mathbf{NP_2} & : & \mathbf{S} \\ \mathbf{NP_3} & : & \mathbf{DO} \end{array} \right],$$

$$a_{11} = \left[\begin{array}{ccc} \mathbf{NP_1} & : & \mathbf{IO} \\ \mathbf{NP_2} & : & \mathbf{DO} \\ \mathbf{NP_3} & : & \mathbf{S} \end{array} \right].$$

2.2 文構造を決定するための手がかり

次に, 文の構造を決定するための手がかり (trace) を次の 4 つに分類する.

- T_{num}:名詞句の数による手がかり
- T_{ord}:語順による手がかり
- T_{np}:各名詞句の形態による手がかり
- T_v:動詞の形態（名詞句との属性の一致）による手がかり

コーパスの用例集合を E と表し, その要素, すなわち一つの文 $e \in E$ から, その要素が表す手がかりへの写像を t_i $(i \in \{num, ord, np, v\})$ と書く.

$$t_i : E \to T_i$$

$$t_i : e \mapsto t_i(e)$$

定義 2.2 用例 e のある手がかり $t_i(e)$ $(i \in \{num, np, v\})$ を与えたときに, $t_i(e)$ と矛盾しない格割当を可能格割当と呼び, そのような格割当の集合を $J(t_i(e))$ で表す.

ここで語順による手がかり t_{ord}は特別扱いするので, 上の定義には含まれていない.

不可能格割当は可能格割当の補集合として定義する. また, 二項演算子∩は二つの手がかりの統合を表す.

例 2.1 用例として e ="Die Frau hat schwarze Haare." という独語文を与える. 手がかりとして, $t_{np}(e) \in T_{np}$（名詞句の形態が与える手がかり）, すなわち「$\mathbf{NP_1}$（Die Frau）は $\mathbf{S}$（主格）か $\mathbf{DO}$（直接目的格）であり, $\mathbf{NP_2}$（schwarze Haare）も $\mathbf{S}$（主格）か $\mathbf{DO}$（直接目的格）である」という情報のみを与えてみる. 可能格フレーム $J(t_{np}(e))$ は,

$$J(t_{np}(e)) = \left\{ \left[\begin{array}{ccc} \mathbf{NP_1}(\text{Die Frau}) & : & \mathbf{S} \\ \mathbf{NP_2}(\text{schwarze Haare}) & : & \mathbf{DO} \end{array} \right], \left[\begin{array}{ccc} \mathbf{NP_1}(\text{Die Frau}) & : & \mathbf{DO} \\ \mathbf{NP_2}(\text{schwarze Haare}) & : & \mathbf{S} \end{array} \right] \right\}$$

だが, ここで, hat という形態から導かれる「$\mathbf{S}$ となる名詞句は単数である」なる $t_v(e) \in T_v$を加えると,

$$J(t_{np}(e) \cap t_v(e)) = \left\{ \left[\begin{array}{ccc} \mathbf{NP_1}(\text{Die Frau}) & : & \mathbf{S} \\ \mathbf{NP_2}(\text{schwarze Haare}) & : & \mathbf{DO} \end{array} \right] \right\}$$

と格割当は一意に決まる.

3 格割当の生起確率

ここで次のような確率変数 X を導入する:

$$X = \left(\begin{array}{ccc} a_1, & \cdots & ,a_k \\ p(a_1), & \cdots & ,p(a_k) \end{array} \right).$$

ここで $p(a_i)$ は, 用例 e の正しい格割当が a_iである確率である. 先に定義した手がかりを与えた場合にこの確率がどのように計算されるかを考えていく.

3.1 語順による手がかりが与えられていない場合

まず, t_{ord}が与えられていないときは, 各可能格割当の生起確率は等しいと仮定する. すると, 手がかり t (t_{ord}を含まない) に対し,

$$p(a_i|t) \equiv \begin{cases} 0 & a_i \notin J(t) \\ \frac{1}{|J(t)|} & a_i \in J(t) \end{cases} \quad (1 \leq i \leq k) \qquad (1)$$

となる. ここでは $p(a_i|t)$ は手がかり t が与えられた場合の格割当 a_iの生起確率を表す.

例 3.1 手がかりがいっさい与えられていない場合は，全ての格割当が可能でありかつ等しい生起率を持つので，生起率は次のように計算される．

$$p(a_i|\phi) = \frac{1}{|J(t)|} = \frac{1}{11}. \quad (1 \leq i \leq k)$$

例 3.2 名詞句の数が 2 であるという手がかり $t_{num}(e)$ が与えられているときは，可能格割当は 4 つ存在し，生起率は次のように計算される．

$$p(a_i|t_{num}(e)) = \frac{1}{|J(t_{num}(e))|} = \frac{1}{4}. \quad (i \in J(t_{num}(e)))$$

3.2 順序付格フレーム

次に T_{ord}が与えられている場合を考えていく．ここでは可能格割当の概念を用いて，以下のようにその値を計算することにする．

まず，順序付格フレームを $f_i(1 \leq i \leq k)$ で表し，その集合を F と表すことにする．ここでは，

f_1 =**S**,
f_2 =**S-DO**,
f_3 =**DO-S**,
f_4 =**S-IO**,
f_5 =**IO-S**,
f_6 =**S-IO-DO**,
f_7 =**S-DO-IO**,
f_8 =**DO-S-IO**,
f_9 =**DO-IO-S**,
f_{10} =**IO-S-DO**,
f_{11} =**IO-DO-S**

$F = \{f_1, f_2, f_3, f_4, f_5, f_6, f_7, f_8, f_9, f_{10}, f_{11}\}$
とする．

$f_i(1 \leq i \leq k)$ の出現確率 $q(f_i)$ は，次のようにコーパスから推定する．

$$q(f_i) \simeq \frac{freq(f_i)}{|E|}. \quad (2)$$

ただし $freq(f_i)$ は，順序付格フレーム f_iを持つ用例の頻度を表す．

格割当の集合から順序付格フレームの集合 F への写像 f を次のように定義する:

$$f(a_i) = f_i.$$

例 3.3 $\mathbf{NP}_1$, $\mathbf{NP}_2$を持つ文に対して，

$$f\left(\begin{bmatrix} \mathbf{NP}_1 & : & \mathbf{S} \\ \mathbf{NP}_2 & : & \mathbf{DO} \end{bmatrix}\right)=\mathbf{S\text{-}DO},$$
$$f\left(\begin{bmatrix} \mathbf{NP}_1 & : & \mathbf{DO} \\ \mathbf{NP}_2 & : & \mathbf{S} \end{bmatrix}\right)=\mathbf{DO\text{-}S}.$$

3.3 語順による手がかりが与えられている場合

さて，t_{ord}が与えられるということは，値 $q(f_i)(1 \leq i \leq k)$ を格割当の生起確率の計算に使用できるということであると考える．よってこれを用いて，t_{ord}と t (t_{ord}を含まない) が与えられた場合，

$$p(a_i|t_{ord}(e) \cap t) \simeq \begin{cases} 0 & a_i \notin J(t) \\ q(f(a_i)) \cdot \frac{1}{\sum_{a_j \in J(t)} q(f(a_j))} & a_i \in J(t) \end{cases} \quad (3)$$

と値を推定する．

以上のように，格割当の生起確率は，式 (1), (3) のようにモデル化された．これは，人手で与えた文法規則と，コーパスによる文法規則（順序付格フレームの頻度）を組み合わせたモデルである．コーパスによる学習部分を少なくすることにより，学習すべきパラメータ数が大きく減少している．

4 他のモデルとの関係

4.1 N クラスモデルとの関係

ここで挙げたような格割当の確率は，N クラスモデル [6] の特殊な形と考えることができる．N クラスモデルにおいて $N = 1$ とし，さらに言語単位として格割当を考え，クラスとして順序付格フレームを考えると，

$$Pr(a_i) = Pr(a_i|f)Pr(f) \quad (4)$$

となる．この式において，

$$Pr(a_i|f) = \frac{1}{\sum_{a_j \in J(t)} p(a_j)} \quad (5)$$

$$Pr(f) = q(f(a_i)) \quad (6)$$

とみなすことにより，(3) が得られる．

つまり，格フレームの生起は無記憶であると仮定していることになる．

4.2 全文最大エントロピーモデルとの関係

本論文のモデルと全文最大エントロピーモデル [7, 8, 9] とは, 文を単語の条件付き確率の積でなく, 素性の集まりと見なしている点で共通点がある. 実際, 本モデルでは, 「人称や数の一致」という素性を取り入れている. しかし, 全文最大エントロピーモデルでは文の生起そのものをモデル化しているが, 本モデルでは格割当をモデル化しており, ここに根本的な違いがある. もちろん確率の計算方法も異なっている. また, 本論文のモデル化の目的が構文情報の定量化にあることも独自の特長である.

5 エントロピー関数の導入

5.1 エントロピー関数

前章で導入したモデルを用いて言語エントロピーを算出するために, エントロピー関数を導入する.

$H(X|T_{i_1}, T_{i_2}, \ldots)$ $(i_1, i_2, \ldots \in num, np, v, ord)$ は, $T_{i_1}, T_{i_2}, \ldots$を統合した手がかりが与えられたときのエントロピーである. コーパス中の各用例について平均をとることによりエントロピーが計算される:

$$H(X|T_j) =$$

$$\frac{1}{|E|} \cdot \sum_{e \in E} \sum_{a_i} -p(a_i|t_j(e)) \log p(a_i|t_{num}(e)). \quad (7)$$

5.2 情報量

エントロピーを組み合わせると, 以下のような情報量が考えられる. 以下のそれぞれにおいて名詞句の数が与えられた時のエントロピー $H(X|T_{num})$ を基準にしたのは, 語順や名詞句の形態を与える時は自動的に名詞句の数も与えられることになるので, それぞれの情報量を比較するためにはこのようにするのが良いと思われるからである.

- 語順が与える情報量.:
  $I(T_{ord}) = H(X|T_{num}) - H(X|T_{num}, T_{ord})$
- 各名詞句の形態が与える情報量.:
  $I(T_{np}) = H(X|T_{num}) - H(X|T_{num}, T_{np})$
- 動詞の形態が与える情報量.:
  $I(T_v) = H(X|T_{num}) - H(X|T_{num}, T_v)$
- 各名詞句の形態と動詞の形態が与える情報量 (形態が与える情報をまとめたもの).:
  $I(T_{np}, T_v) = H(X|T_{num}) - H(X|T_{num}, T_{np}, T_v)$
- 全ての手がかりが与える情報量.:
  $I(T_{ord}, T_{np}, T_v) = H(X|T_{num}) - H(X|T_{num}, T_{ord}, T_{np}, T_v)$

6 計算

これまでに述べた考え方に沿って, コーパスを用いて実際に計算してみる. ここでは独語と英語に関して計算を行ない, その値を比較してみる.

6.1 データについて

以下の数種類の異なったコーパスに対して計算を行なった. 各コーパスはそれぞれ同じ内容が独語及び英語を用いて書かれている.

以下の数種類のコーパスに対して計算を行なった. 各コーパスは同内容が独語及び英語で書かれている.

- ニュース (News): http://www.mathematik.uni-ulm.de/germnews/, http://www.mathematik.uni-ulm.de/de-news/ より各語に関して文数 100 のデータが二つずつ.
- 文学作品 (Literature): Franz Kafka の "Die Verwandlung" から, 独語はは文数 102 と 88, 英語は文数 101 と 106 のデータ. 出典は WWW 上の http://gutenberg.aol.de/gutenb.htm, http://www.vr.net/ herzogbr/index.html.
- 自然科学の論文 (Science): Sigmund Freud の "Psychoanalyse" より, 独語は文数 105 と 95, 英語は文数 103 と 129 のデータ. 出典は WWW 上の http://freud.t0.or.at/.
- 童話 (Fairy Tales): Jakob Grimm 及び Wilhem Grimm による, "Die Bremer Stadtmusikanten" の全文 (独語は文数 103, 英語は文数 92) と, "Der Wolf und die sieben jungen Geißlein" の全文 (独語は文数 88, 英語は文数 85). 出典は http://www.vcu.edu/hasweb/for/menu.html.

同内容でありながら独語と英語で文数が違うのは, 両言語で文章構成が異なっており, 各文が両言語間で一対一対応しているわけではないからである.

また, コーパスは前もって解析されている必要がある. この解析は以下の基準に従って行なわれた.

- 疑問文や感嘆文においては語順に関して特別な規則が存在すると予想されるので, これはデータとして取り入れない. その他, 会話文で出現するような動詞を持たない発話も無視した.
- 名詞句が and などで並列接続している場合は, 両名詞句の情報を統合した上で全体を複数の名詞句とした. 例えば独語において, "die Frau und der Mann" は, "der Mann" が主格であることを示しているので, 複数の主格としてデータ化する.
- to-不定詞句 (独語では zu-不定詞句) や that 節 (独語では dass 節) などは, 主格及び直接目的格になりうる単数の句として扱った (間接目的語にはならないものとしている). 例えば, "I know that he has a book." はそのような場合である. ただし, "I went there to see her." などのように, 副詞句として用いられている場合は, 一般の副詞句と同様にこれを無視した.
- and (und) や but (aber) などで複数の文が結ばれている, いわゆる並列接続の場合は, 全ての文をデータとして取り入れた. 一方, 従属節などにおいては語順に関して他と同じ規則が働いているという保証がないので, これを無視し, 主文のみをデータとして取り入れた. 例えば, "I know that he has a book." において, "he has a book." という文はデータとして取り入れられていない.

6.2 計算結果

数値は小数点以下四桁目を四捨五入して小数点以下三桁まで記した.

表 1 は, 独語と英語それぞれの全てのデータから得られた順序付格フレームの生起率を示す. この数値を, 各コーパスのエントロピー値の計算に用いた. 表 2 は, ニュースと文学作品に関する種々の情報量を示す. 一つの欄に二つ数字があるが, それぞれ同種類の異なるコーパスに対する計算結果である. 表 3 は表 2 と同様で, 科学論文と童話に関する数値である. 表 4 は, 独語と英語について, それぞれのデータを全て合わせたデータに対して情報量を計算したものである.

図 1 は独語と英語に関する構文情報のうち, $I(T_{ord}), I(T_{np}), I(T_v)$ の値を図に表したものである. 図 2 は, 横軸には各コーパスデータから得られた $I(T_{ord})$ の値を, 縦軸には同じく $I(T_{np}, T_v)$ の値を, 独語と英語の区別だけを示して図に表したものである. 図 3 は, 図 2 と同じデータに対して, コーパスの種類を区別してプロットしたものである.

6.3 考察

独語と英語の両言語に対して同じ内容の文をデータとして採用したのは, どちらの言語においても代名詞は格変化が著しいので, 代名詞の頻度が計算結果に影響を与えると予想されたからである.

図 1, 図 2 を見ると, 英語では独語に比べて語順 (元となる格確率分布) が構造決定に大きな役割を果たしていること, 独語では英語に比べて名詞句の形態が大きな役割を果たしていることがわかる. 個々のコーパスから算出された値で比べてみると, 前述の性質をほとんどのデータが満たしている (例外は, 自然科学論文の一つ目のコーパスから算出された値である). これは, 序論で例示した独語と英語の差異と同じ結果であり, 計算結果は我々の感覚と一致する. さらに, 図 2 において特徴的なのは, 二つのデータ群 (独語と英語) が完全に線形分離している点である. ただ, 独語の値は比較的互いに近いのだが, 英語は散らばっている. 実際, 全体の値を基準値とした分散は, 独語は 0.145, 英語は 0.255 であり, 英語の方が分散が大きい.

図 3 は, 図 2 と同じものをコーパスの区別をして表示したものである. 童話や文学において, 他のコーパスと比べ, 形態による情報が大きくなっているが, 結論を導くのは危険だろう.

7 結論

本論文において, 構文情報を, 名詞句の形態が与える手がかり, 動詞句の形態が与える手がかり, 語順が与える手がかりなどに分類し, エントロピー関数を用いてそれぞれを定量化する方法を提案した.

本モデルでは, 名詞句の形態や動詞句の形態が与える手がかりによって格割当を可能格割当と不可能格割当に分割し, コーパスから得られた語順の頻度により各可能格割当に確率を比例分配するという方法で格割当の確率分布が計算されている. そして, 各手がかりを与えたときに格割当の曖昧性がどれだけ減少する

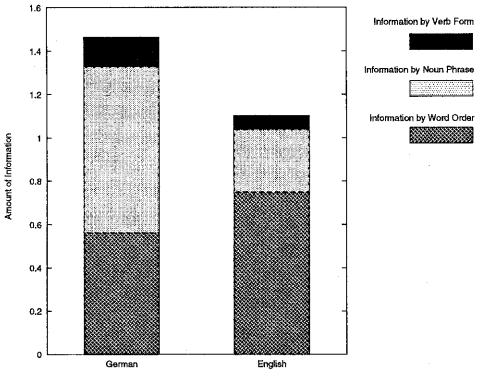

図 1: 独語と英語の各構文情報

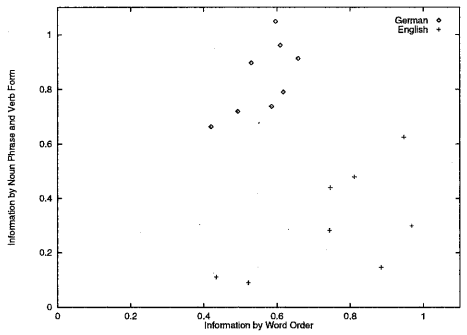

図 2: 語順と形態が与える情報量（言語による比較）

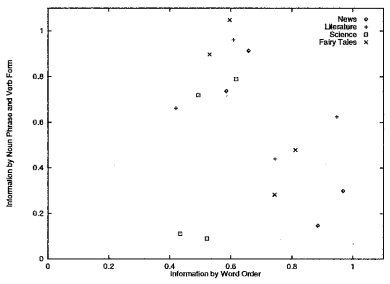

図 3: 語順と形態が与える情報量（コーパスの種類による比較）

かによって，その手がかりが持つ情報量を測っている．さらにその方法を用いて実際に数値計算を行なった．計算結果からは，英語では語順が，独語では逆に形態（名詞句の形態と動詞の形態）が，構造決定に大きな役割を果たしていることがわかった．

しかし，本方法はまだ多くの問題点を持つ．まずはデータ量の問題がある．現時点では，独語と英語を合わせて約 1600 文という小量のデータしか用いていないので，計算結果の信頼度が高いとは言えない．データを増やすには解析済みの大規模コーパスの使用が不可欠だが，直接利用することはできない．なぜならば，曖昧性などを算出するためには，ある句がコーパス中でどのような機能を持っているかだけでなく，他にどのような解釈の可能性があるかという情報が必要だからである（一般のコーパスはそのような情報を持っていない）．例えば，独語の “der Student” という名詞句は必ず主格であるが，“die Studentin” は主格にも対格にもなりうる．これを自動的に認識するプログラムの作成が必要である．

また，文の要素として，主語，直接目的語，間接目的語しか考えてなく，構造は極度に単純化されている．前置詞句や副詞句など，もっと多くの要素を扱えるようにすれば，より実際の言語現象に近い結果が得られるであろう．しかし，それにはモデルの大幅な改良が必要である．それは，格割当の種類の数が m の増加に伴い急激に増加するからである．

以上に挙げたような方法が確立されると，非常に広範囲の発展が期待される．まずは種々の言語に対して計算を行い，それらを比較することである．これにより新しい視点からの言語分類が期待される．その際に現われる最も本質的な問題は，何を名詞句とみなすかということである．格の数が非常に多い言語もたくさん存在し（例えばクロアチア語では 7 つ），格パターンの種類が非常に多くなってしまい，計算が困難になる．また，日本語のような膠着言語，あるいは印欧言語でもイタリア語のように主語の省略が許されている言語はいかに扱うべきかという問題も未解決である．

他言語への応用の他には，いろいろなコーパスに対して計算を行い，それらを比較するというのも興味深い．日常会話と新聞の文章では違った結果が出ることも予想される．ここから，文章の「らしさ」（新聞記事らしさ，日常会話らしさ等）に関する定量的な指標が導かれる可能性もあるだろう．

表 2. 英独両語の構文構造における種々の情報量

	News		Literature		Science		Fairy Tales	
Language	German	English	German	English	German	English	German	English
$I(T_{ord})$	0.658	0.968	0.420	0.745	0.617	0.434	0.596	0.743
	0.585	0.885	0.609	0.947	0.493	0.522	0.530	0.811
$I(T_{np})$	0.800	0.211	0.663	0.440	0.726	0.111	0.970	0.282
	0.596	0.127	0.961	0.625	0.635	0.075	0.788	0.479
$I(T_v)$	0.240	0.246	0.059	0.000	0.101	0.058	0.180	0.050
	0.120	0.050	0.102	0.081	0.107	0.039	0.184	0.000
$I(T_{np}, T_v)$	0.913	0.299	0.663	0.440	0.790	0.111	1.048	0.282
	0.737	0.147	0.961	0.625	0.719	0.090	0.897	0.479
$I(T_{ord}, T_{np}, T_v)$	1.108	1.023	0.756	0.870	1.029	0.466	1.128	0.812
	0.934	0.906	1.105	1.082	0.846	0.548	0.984	0.959
$I(T_{ord}) + I(T_{np}) + I(T_v)$	1.698	1.425	1.141	1.186	1.444	0.603	1.747	1.075
	1.301	1.062	1.673	1.653	1.235	0.636	1.502	1.290

表 3. 英独両語の順序付格フレームの生起確率

Case Frame	German	English
S	0.458	0.536
S-DO	0.388	0.408
DO-S	0.090	0.038
S-IO	0.015	0.000
IO-S	0.005	0.000
S-IO-DO	0.026	0.016
S-DO-IO	0.004	0.000
DO-S-IO	0.006	0.000
DO-IO-S	0.003	0.000
IO-S-DO	0.003	0.000
IO-DO-S	0.000	0.000

参考文献

[1] Shannon, C. E., Weaver, W., *The Mathematical Theory of Communication*, University of Illinois (1949).

[2] Shannon, C. E., Prediction and Entropy of Printed English. *Bell Systems Technical Journal*, Vol. 30, pp. 50–64 (1951).

[3] Cover, T. M. and King, R. C., A Convergent Gambling Estimate of the Entropy of English. *IEEE Transactions on Information Theory*, Vol. IT-24, No. 4, pp. 413–412 (1978).

[4] Brown, P. T., Pietra, V. J. D., Mercer, R. L., Pietra, S. A. D. and Lai, J. C., An Estimate of an Upper Bound for the Entropy of English. *Computational Linguistics* Vol. 18, No. 1, pp. 31–40 (1992).

[5] Fillmore, C. J., The Case for Cases, *Universals in Linguistic Theory*, Holt Rinehart and Winston, New York, pp. 1–88. (1968).

[6] Niesler, T. R. and Woodland, P. C., *Variable-Length Category-based N-Grams for Language Modeling* (Tech.Rep.No. CUED/F-INFENG/TR.215). Cambrigde University Engineering Department.

[7] Rosenfeld, R., A Whole Sentence Maximum Entropy Language Model. *Proceedings of the IEEE Workshop on Speech Recognition and Understanding.*, pp. 230–237 (1997).

[8] Chen, S. and Rosenfeld R., Efficient Sampling and Feature Selection in Whole Sentence Maximum Entropy Language Models, *Proceedings of Acoustics, Speech and Signal Processing*, pp. 549–552 (1999).

[9] Zhu X., Chen S. and Rosenfeld R., Linguistic Features for Whole Sentence Maximum Entropy Language Models, *Proceedings of Eurospeech'99*, Vol. 4, pp. 1807–1810 (1999).