# 統計的手法に基づく品詞解析器の半自動構築

## どんな言語の辞書と文法も半自動で構築することを目指して

山崎　智弘† 　若木　裕美† 　清水　勇詞† 　鈴木　優†

† 東芝研究開発センター知識メディアラボラトリー　〒 212–8582 川崎市幸区小向東芝町 1
E-mail: †{tomohiro2.yamasaki,hiromi.wakaki,yuji.simizu,masaru1.suzuki}@toshiba.co.jp

**あらまし**　市場のグローバル化に伴い、日本語や英語以外の文書を解析したいというニーズが高まっている。我々は統計的手法によってどんな言語にも迅速に対応できる言語解析技術の基盤作りを進めており、第 1 段階として既知言語の情報を元に未知言語の品詞解析器を構築する枠組みを試作した。本論文では単語–品詞ペアからなる品詞辞書、品詞系列の教師データを統計的に生成する手法を説明する。また我々がすでに解析器を持っている英語・スペイン語・エスペラントを既知言語とし、フランス語・ポルトガル語を未知言語として実験した結果について考察する。
**キーワード**　言語解析、辞書構築、文法構築、機械学習、品詞解析

# The Semi-Automatic Construction Of POS Taggers For Specific Languages by Statistical Method

## Aiming To Construct Dictionaries and Grammars of Any Language Semi-Automatically

Tomohiro YAMASAKI†, Hiromi WAKAKI†, Yuji SHIMIZU†, and Masaru SUZUKI†

† Knowledge Media Laboratory, Corporate Research & Development Center, Toshiba Corporation
1. Komukai Toshiba-cho, Saiwai-ku, Kawasaki, 212–8582 Japan
E-mail: †{tomohiro2.yamasaki,hiromi.wakaki,yuji.simizu,masaru1.suzuki}@toshiba.co.jp

**Abstract**　The recent globalization of markets brings us the necessity for analyzing the documents written in the languages other than Japanese and English. We have been making a framework that can quickly develop an analyzer of any language by using statistical processing and machine learning. At present, we confirmed that this framework can build POS taggers of unknown languages with the information of known languages. In this paper, we explain the methods of constructing the POS dictionary and of generating the supervisors of POS sequences. We also describe the result of experiments on the assumption that we know English, Spanish, and Esperanto but we do not know French and Portuguese.
**Key words**　Language Analysis, Dictionary Construction, Machine Learning, POS Tagging

## 1. はじめに

品詞解析をはじめ、構文解析や固有表現抽出といった言語解析技術はテキストデータの解析に欠かせない基盤技術となっており、解析器を構築することで、キーワード抽出、文書分類、機械翻訳といったさまざまな応用が可能となる。しかし従来は日本語や英語といった解析対象となる言語の特性に合わせて辞書や文法の作りこみを人手で行なってきたため、言語依存の部分が多く、対応言語を増やそうとしても新たな解析器の構築には多大なコストがかかるという課題があった。

一方で市場はグローバル化しており、世界各国に対応した製品を開発していくことが求められている。特に、現地の言語で書かれた文書を解析することで

- ユーザの関心や意図を推定してサービスを提供したい、
- 評判や苦情を分析してトラブルを事前に回避したい、

といったニーズが高まりつつある。日本では国際化＝英語化とされがちなこともあってこれまでは主として英語の解析技術に取り組んできたが、これからは英語以外の世界各国の言語に迅

速に対応していかなければならない。

そこで我々は、統計的な手法によって辞書や文法の構築を半自動化することで、どんな言語の解析器でも迅速に構築できる仕組みを確立すべく研究に取り組んでいる。具体的には品詞解析のほか、より高度な言語解析を行なう解析器までをも、表層情報や統計情報などの組み合わせだけを用いて実現することを目指している。ただしまったく何も手がかりがない状態から解析器を構築することは困難なので、既知言語の情報はある程度人手で与えるものとする。

本研究における技術的なポイントは、辞書の構築と文法の構築に大きくわけることができる。そこで本論分では品詞解析器を例に取り、既知言語との対訳が存在する未知言語に対して

- 単語–品詞ペアからなる品詞辞書、
- 品詞系列の教師データ、

を統計的に生成する手法について説明する。そして生成した教師データを用いて未知言語の文法を機械学習し、半自動で得られた未知言語の品詞解析器の精度について検証を行なう。

### 1.1 関連研究

近年、大量のタグつきコーパスが用意できば機械学習によって自然言語処理タスクにおけるさまざまな課題が容易に解決できることがわかってきた。しかし教師データとなるタグつきコーパスをタスクごとに用意することは大きな問題となる。一方タグなしコーパスはインターネットなどから簡単に入手できるようになってきたため、生のテキストコーパスから教師データなしで解析器を生成する手法についても研究されている。例えば [3] は最小記述長原理 (MDL) に基づいて語形変化を推定するものである。ただしこの手法はあらかじめ人手で確率文法を生成しておき、採用すべきものとそうでないものを MDL を用いて判別しているため、ある程度対象となる言語を知っている必要がある。そのため未知言語に対して半自動で解析器を生成することは難しい。

またここ数年、大量のタグつきコーパスを用意できない問題を解消する手法として半教師あり学習が注目されている。これは大量のタグなしコーパスに少量のタグつきデータを与えることで、大量のタグつきコーパスを与えたときと同じ効果を狙ったものである。例えば [2] は最初に与えたシード語からルールを学習することで固有表現認識を行なうためのコーパスを自動生成し、最終的に固有表現抽出器を生成している。しかし [2] に限らず、少量のタグつきデータに誤りが含まれていると自動生成した教師データに急激に誤りが増えてしまい、十分な精度が得られないということが知られている。未知言語に対しては少量であっても正確なタグつきデータを与えることは難しいため、誤りの伝播 (error propagation) を起こさないための工夫が必要となる。

### 1.2 方針

既知言語との対訳を扱うためにはコーパスの入手しやすさが大きな問題となる。大手のニュースサイトであれば、現地の言語のほか英語でも同じ記事を配信していることが多い。そのため対英であれば対日よりもさまざまな言語の対訳を入手しやすいが、それでも言語によっては難しい場合がある。またたとえ X 語–Y 語対訳が入手できたとしても、文同士が対応付けられているとは限らない。一般に単語同士の対応関係がある程度わかっていないと文同士を精度よく対応付けることは難しいため、未知言語の場合は困難が伴う。

そこで我々は実験用コーパスとして聖書を用いるものとした。聖書は世界中で最も読まれている文書であり、非常に多くの言語訳がインターネット上で公開されている [1]。またそれぞれの聖典は章節が細かく区切られているが、どの言語でも基本的には同じ章節となっている。各章節には多くても数文しか含まれないため、文同士がほぼ正しく対応付けられていることになる。

一方、我々は表層情報や統計情報などの組み合わせだけを用いて高度な言語解析を行なう解析器までをも実現することを目指している。そこで最初の取り組みとしては

- UNICODE 上に文字が定義されていない、
- 分かち書きされていない、
- 普通名詞と固有名詞を書き分けない、

といったテキスト処理が難しい言語は対象としないものとした。分かち書きされていない言語としてはタイ語、カンボジア語、ラオス語などがあるが、日本語と同じく単語分割が非常に難しいためである。また大文字小文字の区別がないアラビア語、ヘブライ語、ヒンディー語などや、名詞をすべて大文字で始めるドイツ語は固有名詞の判別が簡単ではないためである。すなわち、ドイツ語を除くヨーロッパ系の言語、中でもなじみやすさの観点からラテン文字を用いる言語を主な対象とする。ただしキリル文字もギリシャ文字も扱いやすさという観点ではラテン文字と大差ないので、それらの文字を用いる言語についても今後は対象としていく予定である。

## 2. 単語同士の対応関係抽出

本論文における品詞辞書を構築する手法には、言語内の統計情報を用いて単語を抽出する処理と言語間の統計情報を用いて単語同士の対応関係を抽出する処理が存在する。以下ではそれぞれの処理について説明する。

### 2.1 言語内の統計情報を用いた単語抽出

本節では言語内の統計情報を用いて固有名詞、連語、表記類似語を抽出する手法について説明する。表記類似語を抽出するのは、後段の処理で同一語の語形変化を推定するためである。

1.2 節で述べたように、普通名詞も大文字で始めるドイツ語を除いたヨーロッパ系の言語を対象とすることにしたため、常に大文字で始まる単語は固有名詞である可能性が高いと考えられる。ただしこれらの言語では文頭も大文字で始めるため、それらは除去する必要がある。これらをふまえ、コーパスをスペースと記号で分割して得られたそれぞれの単語 $w$ に対し

- すべて小文字である出現の回数 $\mathrm{lower}(w)$ が 0 かつ
- 大文字始まりまたはすべて大文字である出現の回数 $\mathrm{upper}(w)$ が 5 以上、

であるものを固有名詞として抽出する。固有名詞でない単語に対して $\mathrm{upper} \geqq 5$ となる確率は、文頭に来る確率を 1/2 とかなり大きく見積もっても $(1/2)^5 = 1/32$ 以下なので 5% 以下の有意水準で固有名詞であると検定することができる。

続いて、文書から連語を抽出するためには C-value [4] と呼ばれる手法が古くから知られている。この手法はコーパスにおける単語間の結合度を計算するものである、連語 $\mathbf{w} = w_1 \dots w_l$ に対して C-value $= (l-1)(n - t/c)$ で定義される。ここで $n$ は $w$ の出現回数であり、$t$ は $w$ を含む $w$ より長い連語の出現回数、$c$ は $w$ を含む $w$ より長い連語の異なり数である。

連語 $\mathbf{w}$ に含まれる単語間の結合度が高ければ $\mathbf{w}$ がまとまって出現しやすいため、$n$ と比べて $t$ が小さくなり、C-value が大きくなりやすい。しかしながら短い連語の場合は $n$ と比べて $c$ が大きくなり、C-value が不当に大きくなる傾向が見られる。そこで我々は C'-value [5] も連語を抽出するために併用するものとした。具体的には C-value と C'-value が閾値 (今回の実験では 50) 以上の連語を抽出する。

ところで、今回対象とすることにしたヨーロッパ系の言語は文法的にはほとんどが屈折語と呼ばれる分類に属している。屈折語においては文法的機能を表す要素が語の内部に埋め込まれ、動詞だけでなく名詞や形容詞も格・性・数などによって語形変化するという特徴がある。そのため、表記が異なっていても同一語の語形変化であると判定できる必要がある。

世界の言語の中にはスワヒリ語のように語頭が変化するものや、アラビア語やヘブライ語のように語中が変化するものもあるが、今回対象とすることにした言語はほとんど語尾が変化するものであるため、共通接頭辞の長さが一定以上の単語同士を表記類似語とし、同一語の語形変化の候補として抽出するものとした。具体的にはコーパスをスペースと記号で分割して得られた単語のすべての組み合わせ $(w_1, w_2)$ に対し、以下のような処理を行なう。

- $l = \min(|w_1|, |w_2|), L = \max(|w_1|, |w_2|)$ とおく。
- $l \geqq L/2$ かつ共通接頭辞の長さ $\mathrm{pre}(w_1, w_2) \geqq L/2$ iff $w_1 \sim w_2$ と定義する。
- $\sim$ の反射推移閉包 $\sim^*$ の同値類によって表記類似語の集合を抽出する。

語の組み合わせにおける類似性さえ判定できれば、反射推移閉包は類似性の判定方法によらず抽出することができる。そのため、今回は実験していないが、共通接頭辞の長さの代わりに共通接尾辞の長さで判定すれば語頭が変化する言語に、共通部分文字列の長さで判定すれば語中が変化する言語にも同じ手法が適用可能であると考えられる。

### 2.2 言語間の統計情報を用いた単語同士の対応関係抽出

本節では言語間の統計情報を用いて単語や連語同士の対応関係、表記類似語に含まれる同一語の語形変化を抽出する手法について説明する。

X 語の単語 $w^x$ と Y 語の単語 $w^y$ が対応関係にある場合、対訳における $w^x$ の出現位置と $w^y$ の出現位置には関連があると考えられる。一般には文同士が対応付けられているとは限らないため出現位置に関連があるかどうかを判定することは簡単ではないが、対応付けられている対訳であれば以下のようにして簡単に判定することができる。例えば $n$ 文からなる X 語–Y 語コーパスにおいて、X 語の単語 $w^x$ と Y 語の単語 $w^y$ の出現を集計したところ、表 2 のようであったとする。対応する文においてどちらも出現したのが $a$ 文、$w^x$ のみ ($w^y$ のみ) 出現したのが $b$ ($c$) 文、どちらも出現しなかったのが $d$ 文あったことを表している。

表 2 $w^x, w^y$ の出現の集計

| | $w^y$ が出現 | $w^y$ が非出現 | 計 |
|---|---|---|---|
| $w^x$ が出現 | $a$ | $b$ | $e = a + b$ |
| $w^x$ が非出現 | $c$ | $d$ | $f = c + d$ |
| 計 | $g = a + c$ | $h = b + d$ | $n = a + b + c + d$ |

このような集計表において期待値からのズレの 2 乗を期待値で割り、すべてのマスについて合計した値は $\chi^2$ 値と呼ばれ、$\chi^2$ 分布と呼ばれる分布に従うことが知られている。表 2 の場合、$\chi^2 = n(ad-bc)^2/efgh$ と求められるので、この値が 0.384 以上であれば 5% 以下の有意水準で $w^x$ と $w^y$ に関連があると検定することができる。なおこの検定は文において出現しているかどうかだけを用いているので、言語はもちろん文の長さにも依存しない。また単語同士だけではなく 2-gram や連語同士でも同様に対応関係を判定することができる。

しかしながら文において出現しているかどうかだけでは文中の出現位置の情報が失われるため、対応関係にあると判定される語が複数出てきてしまう場合がある。複数出てきてしまっても、最終的に単語–品詞ペアからなる品詞辞書を構築することができれば特に問題ないともいえるが、今回はなるべく正しく対応させることを目的とし、単語同士の表記の類似度を計算して最も近いものを対応関係にある単語として選択するものとした。これは今回対象とすることにしたヨーロッパ系の言語はほとんどが印欧語族と呼ばれる言語グループに属しており、対応関係にある単語同士は似たような表記ないし発音になりやすいためである。今回は実験していないが、固有名詞はどのような言語でも似たような発音になると考えられることを考慮すると、表記から発音に変換した上で類似度を計算するほうがどんな言語にも適用可能となり、より汎用的である。

さて、ここまでは言語に依存した情報は積極的には使わず統計情報のみで対応関係を求める手法について説明してきたが、ここからは X 語については既知である (= 品詞解析器がある) と仮定した上で未知の Y 語の語形変化を抽出する手法について説明する。

前節で述べたように、表記類似語は同一語の語形変化の候補を含んでいる。X 語については仮定により品詞解析器があるため、標準形を求めることで同一語かどうかを判定することができる。すなわち Y 語のある表記類似語 $\mathrm{sim}^y = \{w_1^y, w_2^y, \dots\}$ に含まれるすべての部分集合と X 語の標準形との関連性を判定し、最も関連性が高いと判定された部分集合を選択すれば、X 語の標準形に対応する Y 語の語形変化を抽出することができると考えられる。具体的には以下のような処理を行なう。

- X 語の標準形 $\overline{w}^x$ と Y 語の表記類似語 $\mathrm{sim}^y = \{w_1^y, w_2^y, \dots\}$ を選択する。
- $\mathrm{sim}^y$ に含まれるすべての部分集合 $\overline{\mathrm{sim}}^y \in 2^{\mathrm{sim}^y}$ に対して関連性判定を行ない $\chi^2$ 値を計算する。

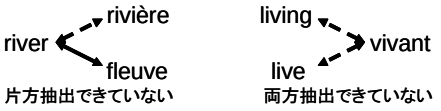
表 1　フランス語聖書から抽出された単語の例

| 固有名詞 | 連語 | 表記類似語 |
|---|---|---|
| Jubal, Assyrie, Jébusien, Guérar, Nimrod, Calakh, Gaza, Dikla, | en paix, le livre, car vous, nos pères, l'autel, de guerre, sa famme, d'Égypte, | {répara,réparent,réparé,réparât,réparer,réparèrent}, {sanctifient,sanctifie-la,sanctifier,sanctifié,sanctifieras,sanctifiée,sanctifie, sanctifierai,sanctifierez,sanctifiés,sanctifiez-vous, sanctifièrent,sanctifiez, sanctification,sanctifiaient,sanctifiait,sanctifiez-le,sanctifieront,sanctifiât}, |

- $\chi^2$ 値が最大となる $\overline{\mathrm{sim}}^y(\overline{w}^x)$ を求め、これを $\overline{w}^x$ と対応する単語の語形変化であるとみなす。

表 1 からわかるように、表記類似語の個数は 15 を越えることも多く、部分集合のサイズは非常に大きくなってしまうことがある。このような場合にまともに検定を行なうと計算量が非常に多くなってしまう。しかし前述したように検定は文の長さには依存しないことに着目すると、適当にマージした文を新たな文とみなして検定することも可能である。すなわち $M$ 個の文をマージして計算すれば $\chi^2$ 値の計算量は $1/M$ になるため、高速化が見込まれる。実際、今回の実験でまともに計算していたときは 24 時間以上かかっていたが、表記類似語の個数が 15 以上のときは 32 文をマージして計算するようにしたことで、語形変化の抽出処理が 2 時間以内に完了するようになった。

以上 $\chi^2$ 検定による手法を説明してきたが、基本的に $w^x$ と $w^y$ について対称であることからもわかるように、1 対 1 対応にある単語同士以外は抽出することが難しいという欠点がある。例えば英語聖書では「川」を表す単語として river しか使われていないが、フランス語聖書では rivière と fleuve が使い分けられている。逆にフランス語聖書では「生きている」を表す単語として vivant しか使われていないが、英語聖書では形容詞の living と動詞の live が使い分けられている。このような状況では「対応関係にある単語同士の出現・非出現は一致する」という前提が成り立っていないため、そのままでは片方ないし両方がうまく抽出できない。




そこで連語同士の対応関係を利用し、1 対 1 対応にないものの抽出も行なうものとした。対応関係にある連語同士について、連語に含まれる単語同士の対応関係があればそれらを取り除いていくということを繰り返し、最終的に単語同士が残ったとすればそれらは対応関係にあると推定することができる。新たに対応関係にあると推定された単語同士をさらに他の連語から取り除いていくことで、さらなる単語同士の対応関係を推定していくことができる。具体的には固有名詞、単語、2-gram、3-gram の対応関係を元に以下を繰り返す。

- 対応関係にある X 語の 2-gram $w_1^x w_2^x$ と Y 語の 2-gram $w_1^y w_2^y$ に対し、$w_1^x$ と $w_1^y$ が対応する固有名詞のときは $(w_2^x, w_2^y)$ を抽出する。同様に $w_2^x$ と $w_2^y$ が対応する固有名詞のときは $(w_1^x, w_1^y)$ を抽出する。
- 対応する 2-gram、3-gram から対応関係にある単語同士を取り除き、最終的に単語同士が残ったときはそれらを抽出する。

表 3　連語同士の対応関係から抽出された単語の例

| 英語 | living | only | praising | sharp | then | wise |
|---|---|---|---|---|---|---|
| フランス語 | vivant | unique | louant | tranchante | alors | sage |

表 3 は連語同士の対応関係から抽出された単語の例である。先ほどはフランス語の vivant に対応すべき英語の living と live が両方とも抽出できていなかったが、連語同士の対応関係を利用することによって living の方は抽出できたことがわかる。しかし live の方は抽出できていないため、さらなる手法の改善が必要である。

## 3. 辞書構築と教師データ生成

前節では X 語と Y 語の対訳、および X 語の品詞解析器を用いて、対応関係にある X 語の単語 $w^x$ と Y 語の単語 $w^y$ を抽出する手法について述べた。最終的に単語–品詞ペアからなる Y 語の品詞辞書を構築するためには、$w^y$ の品詞を推定する必要がある。仮定により X 語の品詞解析器があるため $w^x$ の品詞は決定できるので、その品詞を $w^y$ の品詞とみなせばよい。

表 4　用いた品詞一覧

| A | C | D | I | M | N |
|---|---|---|---|---|---|
| 形容詞 | 接続詞 | 限定詞 | 間投詞 | 数詞 | 名詞 |
| P | R | S | V | 0 | _ |
| 代名詞 | 副詞 | 前置詞 | 動詞 | 数字 | 記号 |

ただし $w^x$ の品詞は一意に決まらないことがあることに注意が必要である。例えば英語は名詞にも動詞にもなる多品詞語が非常に多いことが知られている。英語の name は「名前」という名詞にも「名づける」という動詞にもなるが、フランス語の nom は名詞にしかならない。そのため適合率の観点からは文脈に応じて品詞を推定するほうがよいと考えられるが、今回は手法を簡単にするため $w^x$ の品詞のうち出現回数が全体の 1/4 以上を占めるものはすべて $w^y$ の品詞とみなすものとした。

一方今回対象とすることにしたヨーロッパ系の言語はほとんどが印欧語族と呼ばれる言語グループに属しており、基本的な文法には大きな差がないとされる。逆に言うと対応付けられている文同士の品詞系列には言語間で大きな差がないことを意味している。すなわち既知の X 語の品詞系列を元に以下のような条件で最小マッチング問題を解くことで、未知の Y 語の品詞系列の教師データを生成することができる。

- 限定詞、代名詞、前置詞、数字、記号は異なる品詞とはマッチさせない。これらの品詞は他の品詞とはマッチしないと考えられるためである。

表 5　実験に使用した言語と聖書の版

| 言語 | 版 | 章節数 | 単語数 (延べ) | 単語数 (異なり) |
|---|---|---|---|---|
| 英語 | American Standard | 31103 | 918287 | 13256 |
| スペイン語 | Reina-Valera | 31103 | 824760 | 28874 |
| エスペラント | 英国聖書教会 | 31103 | 796700 | 30760 |
| フランス語 | Darby | 31103 | 935222 | 24924 |
| ポルトガル語 | Almeida Atualizada | 31103 | 828352 | 29306 |

- スキップするコストは $c_{\mathrm{skip}}$。
- X 語の品詞が Y 語の品詞候補に含まれるときのコストは 0、含まれないときは $c_{\mathrm{diff}}$。ただし Y 語の品詞候補が $\emptyset$ のときは 0。

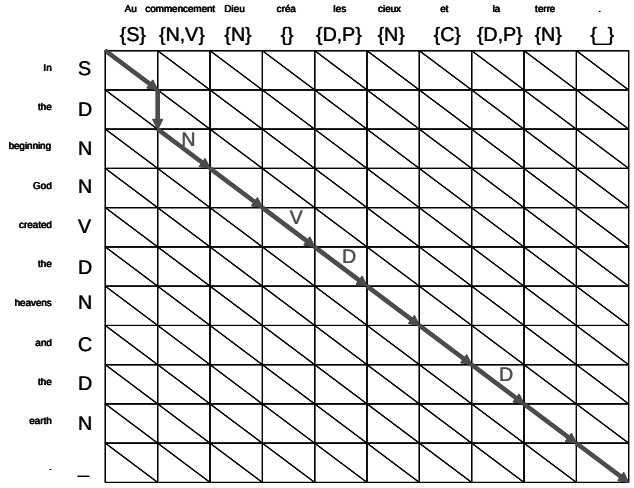


図 1　動的計画法を用いた最小マッチング問題の解

なお今回は手法を簡単にするために単語同士の対応付けを直接は用いていないが、対応する単語同士を優先してマッチさせるようなコスト設定にしてもよい。

## 4. 実験結果

本章では、我々が品詞解析器を持っている英語・スペイン語・エスペラントを既知言語とし、フランス語・ポルトガル語を未知言語として実験した結果について説明する。なお言語によっては訳者や時代で聖書にいくつも版がある場合があるが、今回は表 5 に示すものを利用した。

まず本手法によって獲得できた単語によるカバー率を図 2 に示す。多少のばらつきはあるものの、獲得できた単語によるカバー率は既知言語が英語・スペイン語・エスペラントのどれでも同程度であり、言語に依存しない安定したアルゴリズムになっていると考えられる。また延べではカバー率が 0.98 を越えていることから、統計的手法によってほぼすべての単語に対して何らかの品詞が推定できていることがわかる。ただし異なりではポルトガル語のカバー率が 0.75 程度とそれほど高くないため、まだ改善の余地がある。

続いて、上記の品詞辞書を元に生成した教師データから CRF [6] を用いて文法を学習し、半自動で得られた品詞解析器の精度について評価した結果を図 3 に示す。聖書にはそれぞれの単語の品詞情報がついているわけではないので、フランス語・ポルトガル語ともに教師データを除く 60 文 900 語程度を抽出して人手で品詞の正解を作成し、評価を行なった。図 3 を見ると、半自動で構築した品詞解析器でありながら 0.9 程度の高い正解率が出ていることがわかる。

なお不正解となっていた箇所を調査したところ、既知言語の文法的特徴を引きずっていることが要因である場合が多いことが確かめられた。例えばフランス語の voici という単語は、副詞でありながら主語＋動詞を兼ねることができるという特徴がある。そのため品詞解析器はある意味では正しく動詞と推定しており、不正解となっていた。あるいはフランス語やポルトガル語では形容詞が名詞の後ろに来るにも関わらず、英語では形容詞が名詞の前に来るため、名詞にも形容詞にもなる可能性がある単語が並んでいる箇所で前を形容詞と推定してしまっていた。

また容易に想像できるように、現代の文書には聖書に出現しない単語が数多く出現する。構築した品詞辞書にない単語については周辺の単語から品詞を推定しなければならないため、カバー率が低くなるととその分精度も低くなってしまう。すなわち既知言語の情報だけを使うのではなく、未知言語に対しても文法書に記載されているような最低限の文法知識はあらかじめ人手で与え、Wikipedia のような大規模コーパスから現代的な単語のうち名詞や動詞や形容詞になりうるものを抽出する手法を取り入れる必要があると考えられる。

## 5. おわりに

本論文では、既知言語との対訳テキストデータを用いて未知言語の単語–品詞ペアからなる品詞辞書を構築する手法と、品詞解析のための文法を学習するための教師データを自動生成する手法について説明した。また対訳テキストデータとして聖書、既知言語として英語・スペイン語・エスペラントを用いることで、フランス語・ポルトガル語の品詞辞書、ならびに品詞解析器を半自動で構築することができた。しかも 0.9 程度の高い正

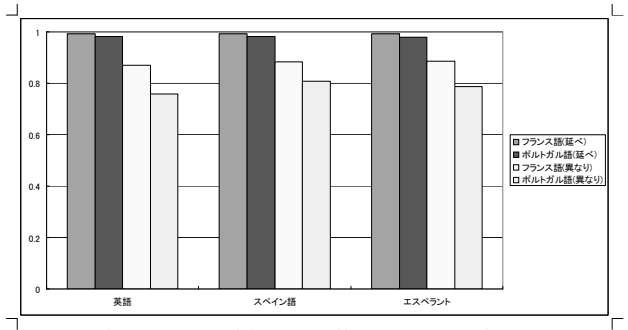


図 2　既知言語ごとの未知言語の単語カバー率 (延べ・異なり)

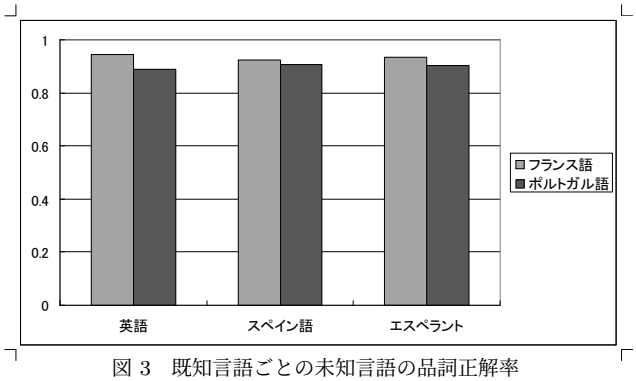


図 3　既知言語ごとの未知言語の品詞正解率

解率が出ることを確認することができた。

今回はキリル文字やギリシャ文字の言語は対象としなかったが、今後はこれらの文字を用いるロシア語やウクライナ語、ギリシャ語などにも範囲を広げていく予定である。また品詞解析だけではなく固有表現抽出やより高度な言語解析を、多言語展開できる言語処理基盤として確立していくため、それぞれの解析のための辞書・文法の自動構築にも取り組んでいく。

## 文　　献

[1] Unbounb Bible. http://unbound.biola.edu/
[2] C. Niu, W. Li, J. Ding, R. K. Srihari, A bootstrapping approach to named entity classification using successive learners, Proc. ACL 2003, pp.335–342
[3] J. Goldsmith, Unsupervised Learning of the Morphology of a Natural Language, Journal of Computational Linguistics (2001), Vol.27, No.2, pp.153–198
[4] K. Frantzi, S. Ananiadou, Extracting nested collocations, Proc. COLING-96, pp.41–46.
[5] 山崎智弘, 強連結成分分解を利用した電子番組表からの話題抽出, Journal of DBSJ, vol.7, No.1, pp.1–6.
[6] J. Laffert et al., Conditional random fields, Proc. Machine Learning (2001), pp. 282–289.