東京大学
情報理工学系研究科 創造情報学専攻
修士論文

データ圧縮に基づく
多言語混合文書の言語に関する分割

Segmentation of multilingual mixed text by language
using data compression

山口 洋
Hiroshi Yamaguchi

指導教員　千葉 滋 教授

2013 年 1 月

概要

本研究では多言語混合文書を扱う。多言語混合文書とは、言語の異なる複数の文書片を繋ぎあわせた構造を持った文書のことである。このような文書はウェブ上に数多く存在しており、特に少数言語で顕著に見られる。また、その処理には、書かれた言語ごとに文書を分割し、単一言語で書かれた文書に分解する前処理が必要とされる。しかしながら、多言語混合文書の存在は十分認知されていないため、その手法は未だ確立されていない。本研究では、データ圧縮を用いた言語判別の研究を元に、この分割問題をデータ圧縮の符号長とモデルの記述長を用いた最小化問題として定式化した。さらに動的計画法を用いて、厳密解を求める線形時間アルゴリズムを示した。また、Wikipedia と世界人権宣言から作成した人工データを用いて、複数の文字種を含む合計 200 以上の言語について実験し、提案手法が有効に機能することを確認した。

Abstract

This paper deals with multilingual mixed texts, which mean concatenation of text segments written in different languages. There are a lot of such texts in the WWW, especially in sites for minor language speakers. In automatic processing of such texts, they must first be segmented by language, and the language of each segment must be identified. The problem is formulated as a minimizing problem using data-compression and solved with linear time algorithm. Empirical results are presented for experiments using artificial texts generated from the Universal Declaration of Human Rights and Wikipedia, covering more than 200 languages including multiple character systems.

目次

第 1 章

序論

世界には, 諸般の事情から書かれた言語の異なる文を含んでいるような文書が存在する. それに対する一般的な名称はないが, ここではそれを多言語混合文書と呼ぶことにする. Wikipedia など, World Wide Web 上に無視できない数が存在するが, 割合としてはかなり少ないことや, 少数言語ページに多いなど分布の偏りもあって, その存在はまだ, 十分に認知されておらず, 処理方法は未だ確立されていない.

1.1 多言語混合文書

本論文では, 文書を単一言語からなる文書片に分割する方法を扱う. その目的は主に,

1. 多言語混合文書を言語情報資源として利用しやすくするため,
2. 多言語混合文書を自然言語処理の手法で解析するため,

の 2 つに大別することができる. 前者は, 多言語混合文書は, 単一言語の言語情報資源として見た場合, 他の言語で書かれたノイズを含んでしまっているので, それを除去して質を改善することが必要ということである. 実際, 少数言語版の Wikipedia においては頻度がかなり高いので大きな意味がある. また, 後者は, 自然言語処理のための解析器は基本的に単一言語のみで書かれた文書用しかないため, 文書を分割することによって, 単一言語からなる文書の解析に帰着しようということである.

このように, 前処理として使うことを想定しているため, 教師ありで分割し, 分割と同時に各文書片に言語タグを付与することも行う.

1.2 データ圧縮に基づく手法

分割の具体的な方法として, 本論文ではデータ圧縮を用いた符号長の最小化問題として分割問題を解くことを提案する. これは, 関連する, データ圧縮の文書が似ているほど符号長が短くなるという性質を利用した, 言語判別の研究 [24] [15] [6] を応用したものであり, 学習にあまり

大きなデータを必要としないため, 少数言語を多数扱う上でも都合がよい. データ圧縮手法としては, 数あるものの中から, 辞書型・統計型の代表として MMS [15], PPM [9] を採用した. また, 分割には個数に応じた損失を付加しており, 過剰な分割を制限している. これは, モデルの記述長として説明されるものであり, 最小記述長 (MDL, Minimal Description Length) 原理 [20] を参考にしているが, いくらか異なる点もある.

また, 同時に動的計画法を用いた最適解の計算法も提案する. これにより, 厳密解を言語数と入力文書の線形時間で処理することができ, 実用的にも問題ない速度で処理することができる.

1.3　少数言語への対応

多言語混合文書は少数言語によく見られることもあり, これらの言語を網羅するためできるだけ多くの言語に対応することにする.

比較的入手しやすい多言語コーパス[*1]として, 世界人権宣言と Wikipedia があり, 入手できるほぼすべてを対象とする. 文字種は約 30, 言語数はのべ約 300 である.

現代では, Unicode が存在するため, これを使用することで文字コードの差異を気にする必要はなくなったが, それでも, ごく限られた主要な言語のみに対応する場合に比べて, 以下のような困難がある.

作業量　数が多いため各言語ごとに特化した作りこみを必要とするアプローチは困難である,

入手可能性　一般的な自然言語処理においては, コーパスを大量に用意して学習を行うことが多いが, 特定の少数言語とわかるコーパスを十分に揃えるのは手間がかかる. また, 多言語混合文書のが存在するという問題もあり, 依存関係が循環する.

言語の類似性　言語数が多くなると, 言語集合の中に紛らわしいものが増えてくる. 系統的に近いもの, 借用によって共通する単語が非常に多くなってしまったものなどがある. これにより, 素朴すぎるアプローチでは誤分割・誤分類が多発して非常に使いにくくなる.

表記の多様性　句読点などの約物はある程度統一されているように思われるかもしれないが, 実際は言語により差異があり, 言語によっては全く違う使い方がなされる場合がある.

計算資源の問題　オンメモリで処理したい場合, 言語数が多いためメモリも制約となりうる. 例えば, メモリが 1GB 使えるとして, 300 言語では 1 言語あたり 3MB, 接尾辞木を使うとすると 20 倍程度は必要なので, 学習に使えるのは 150kB となる. このサイズだと, 単語ベースのアプローチはつらくなる.

前述のデータ圧縮を用いた提案手法は, それほど大きな学習データを必要とせず, 識別能力も高いため, これらの問題点をほぼ解消することができる.

[*1] これらを構成するそれぞれの文書自体は単一言語のみからなると想定されていて, 基本的に多言語混合文書ではない

1.4 本論文の構成

本論文は次のように構成される.

- 2 章では, 多言語混合文書の定義とその実態に関する調査について述べる.
- 3 章では, 多言語混合文書の処理, 言語判別, 文書分割の手法等に関する関連研究について紹介する.
- 4 章では, 分割問題の符号長最小化問題としての定式化について述べる.
- 5 章では, データ圧縮とそれを用いたクロスエントロピーの推定法について述べる.
- 6 章では, 4 章で定義した最小化問題に対する動的計画法による線形時間解法について述べる.
- 7 章では, 人工データを用いた実験の結果を載せる.
- 8 章では, 実データを用いた実験の結果を載せる.
- 9 章では, まとめと今後の課題を述べる.

第 2 章

多言語混合文書の定義と実態

この章では, 多言語混合文書の定義を述べるとともに, 多言語混合文書を目にしたことのない人たちのために, 主に Wikipedia を例に, その存在実態について説明する.

2.1 多言語混合文書の定義

本論文では多言語混合文書を次のように定義する.

定義 2.1 文書 X に対して, 文字列の列 $X_1, X_2, \ldots, X_n$ $(n \geq 2)$ への分割を考える. このとき,

1. 各 X_i は節より大きな文書片を含み,
2. 各隣り合う文字列 X_i, X_{i+1} の主とする言語が異なる,

ような分割が存在するなら, 文書 X を多言語混合文書と呼ぶ.

Alex らの英文中のドイツ語由来の単語を指摘する研究 [3] [4] のように, 他の言語由来の単語, あるいは句を指摘するような研究も考えられるが, 本論文ではあくまで節より大きな単位のみを扱い, 言語の帰属を単語単位では行わない. これは,

- 借用語・外来語が元々のどの言語に属していたかという議論を避けるため,
- 関係のない言語間での綴りの似た単語の判別にはある程度の文脈が必要となるため,

といった理由がある.

しかしながら, 外来語・借用語等を指摘する問題も重要な自然言語処理の問題であることから, これをサポートする手法は今後の課題の 1 つとしたい.

なお, 細かい単位での分割を考慮するために, 分割の候補として, すべての単語境界, あるいは文字間を考慮する必要がある. 一般に文字列が短いほど判別が困難になるため, 本当に文より細かい単位を考慮する必要があるのか, と疑問に思われるかもしれない. しかしながら, http://lb.wikipedia.org/wiki/V%C3%ABlkermord_an_der_Vend%C3%A9e において

は, ルクセンブルク語文中に副文としてフランス語の節が含まれている箇所があり, 実際に文中に埋め込まれていると判断すべきものは存在するため, 考慮すべきであると考えている.

2.2　多言語混合文書の存在理由と分類

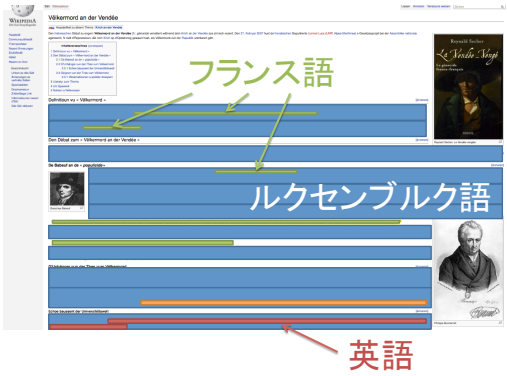

図 2.1. 引用の例

http://lb.wikipedia.org/wiki/V%C3%ABlkermord_an_der_Vend%C3%A9e

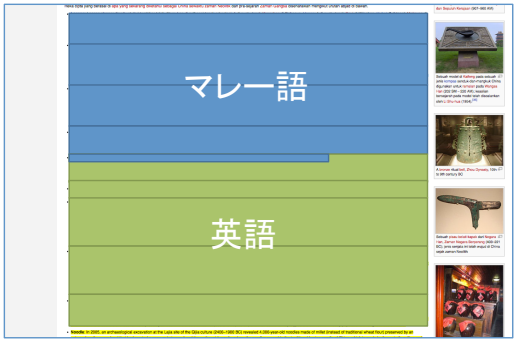

図 2.2. 不完全な翻訳の例

http://ms.wikipedia.org/wiki/Senarai_reka_cipta_China

多言語混合文書を意識して見たことがある人はそれほど多くないため, そのようなものが存在する理由がわからないかもしれない. そこで, 多言語混合文書の存在理由あるいは成り立ちについて説明する.

多言語混合文書の代表的なものは, 大きく分けると次の 4 つになる.

1. 意図的なもの
   (a) 引用

   非常によく見かける場合としては引用がある. 引用では, 原文をそのまま示すことが重要であるため, 元が他言語である場合, 文書内に他の言語で書かれた部分が存在する原因となりうる.

   形態としては, 段落中に挿入されているもの, 字下げされた別のブロックに分けられているもの, など様々である. 正書法に則った整った文書であれば, 引用符, 斜体などでわかりやすく示されていることが多く, これらの情報が利用できる場合には, 問題になりにくい. しかしながら, 必ずしも示されているわけではなく, ラテン文字の規則の影響の少ない言語の文書, 口語的な文書, 正書法の整備されていない言語, すなわち少数言語においては, どこに挿入されているかはそれほど自明ではない.

   (b) 多言語併記

   他のよくある場合としては, ロゼッタストーンに代表される, 多言語併記がある. 実用上も並列コーパス (並行コーパス / 平行コーパス) として利用できるため非常に

利用価値が高い. 基本的には文書の構造により示されている場合が多いため, それらの情報を使えば十分であり, あまり問題にならない.

(c) 他言語話者に対する案内

見過ごされがちではあるが, 他の重要な場合として, 他言語の話者に対する案内文がある. 少数言語の場合, 要約や, 本文に関する注記事項などを伝える対象を, その言語の話者に限ることができない場合が非常に多くなる.

基本的には本文と区別して書かれるべきであるが, 本文中に含まれてしまっている場合もあり, 多言語混合文書が存在する原因の 1 つとなる.

2. あまり意図的でないもの

(a) 不完全な翻訳

出版物にはあまり見られないものの, 多人数で協力して文書を作る過程が見られるウェブに特徴的な多言語混合文書の場合として, 不完全な翻訳というものがある. これは, 翻訳者の力量不足か, 多言語併用環境下で翻訳の必要性を感じなかったのか, 翻訳途中でそのままにされたものである. 途中の段落まで訳されたもの, 1 段落目だけ訳したもの, 各段落の 1 文目だけ訳したもの, モザイク状に訳したものなどが見られる.

このように, 様々な理由から多言語混合文書は生まれており, ウェブ上には数多く存在しうる.

2.3　Wikipedia における多言語混合文書

言語ラベルが付いている多言語コーパスとして重要なものとして, Wikipedia がある. もちろん, Wikipedia も多くの多言語混合文書を多く含んでいる. 前節の分類に当てはめて, 詳しい状況について述べることにする.

(a) 意図的なもの

i. 引用

まず, 引用は Wikipedia においても多く見られる. 例としては, `http://lb.wikipedia.org/wiki/V%C3%ABlkermord_an_der_Vend%C3%A9e` があり, ゲルマン系のルクセンブルク語の文書中にロマンス系のフランス語などの文が多く引用されている.

ii. 多言語併記

次に多言語併記は, Wikipedia ではあまり見られない. なぜなら, 言語が分けられているため, それぞれ別の記事に書かれることが好ましいためである. 歌詞など韻文の紹介などにわずかに見られる.

もう少し広げて, 正書法の違い, ラテン文字転写などを含めると, もう少し数がある. 例えば, 中央アジアでは, ロシアの進出により, 正書法に用いる文字が, アラビア文字・キリル文字・ラテン文字で 2 転 3 転した言語があり, そういっ

た言語では複数の表記を併記している場合がある．あるいは, 過去に独自文字を用いていた言語でラテン文字を

iii. 他言語話者に対する案内

他言語話者に対する案内もいくらか見られる．基本的にテンプレートで示されるものであるが, 本文中に埋め込めれているものもある．例としては, http://xh.wikipedia.org/wiki/Iphepha_Elingundoqo があり, これは, コサ語による歓迎の文書と混在して, ここはコサ語版の Wikipedia であることが書かれている．

(b) あまり意図的でないもの

i. 不完全な翻訳

不完全な翻訳は Wikipedia では非常に多く見られる．例としては, http://ms.wikipedia.org/wiki/Senarai_reka_cipta_China があり, 途中までマレー語に訳したものの, そこより後ろは英語のままとなっている．

多くの場合, 翻訳途中であることを示すテンプレートが貼られているが, 必ずしも存在するわけではない．さらに悪いことには, 分量が大きい記事で多く見られるため, 次の節で説明するような問題の大きな原因ともなっている．

特に Wikipedia を処理しようとする上で問題となるのは, (2b) である．日本語版 Wikipedia など主要な言語版においても存在していて, 例えば,

- http://ja.wikipedia.org/wiki/%E7%84%A1%E6%94%BF%E5%BA%9C%E5%85%B1%E7%94%A3%E4%B8%BB%E7%BE%A9
- http://ja.wikipedia.org/wiki/%E3%82%A6%E3%82%A3%E3%83%A9%E3%83%BC%E3%83%89%E3%83%BB%E3%83%B4%E3%82%A1%E3%83%B3%E3%83%BB%E3%82%AA%E3%83%BC%E3%83%9E%E3%83%B3%E3%83%BB%E3%82%AF%E3%83%AF%E3%82%A4%E3%83%B3
- http://ja.wikipedia.org/wiki/%E3%83%9E%E3%83%8D%E3%83%BC%E3%82%B5%E3%83%97%E3%83%A9%E3%82%A4
- http://ja.wikipedia.org/wiki/%E3%82%B8%E3%83%A7%E3%83%B3%E3%83%BB%E3%82%B9%E3%83%9F%E3%82%B9_%28%E5%8A%B4%E5%83%8D%E5%85%9A%E5%85%9A%E9%A6%96%29
- http://ja.wikipedia.org/wiki/%E3%82%B9%E3%82%AD%E3%83%83%E3%83%97%E3%83%AA%E3%82%B9%E3%83%88
- http://ja.wikipedia.org/wiki/%E3%82%A2%E3%83%B3%E3%83%89%E3%83%AC%E3%83%BB%E3%82%B7%E3%82%A7%E3%83%8B%E3%82%A8
- http://ja.wikipedia.org/wiki/%E4%BA%8C%E6%AC%A1%E7%94%9F%E6%88%90%E7%89%A9
- http://ja.wikipedia.org/wiki/%E3%82%A2%E3%82%A4%E3%83%A4%E3%83%BC%

E3%83%B4%E3%82%A1%E3%83%AA

- http://ja.wikipedia.org/wiki/%E3%82%AB%E3%83%BC%E3%83%80%E3%83%BC%E3%83%AB%E3%83%BB%E3%83%A4%E3%83%BC%E3%83%8E%E3%82%B7%E3%83%A5
- http://ja.wikipedia.org/wiki/%E3%82%B5%E3%83%B3%E3%83%BB%E3%82%AB%E3%83%AB%E3%83%AD%E3%82%B9%E3%83%BB%E3%83%87%E3%83%BB%E3%83%AA%E3%82%AA%E3%83%BB%E3%83%8D%E3%82%B0%E3%83%AD

などは, 英語を含んでおり,

- http://ja.wikipedia.org/wiki/%E3%83%9E%E3%83%83%E3%82%AF%E3%82%B9%E3%83%BB%E3%83%AA%E3%83%83%E3%83%88%E3%83%9E%E3%83%B3

のドイツ語のようにそれ以外の言語を含んでいる場合もある. おおよそどこの言語版にも混入しているのは英語であり, 地域によってフランス語, スペイン語, ロシア語, 中国語, 分野によってドイツ語, 日本語などが含まれているものが見受けられる.

また, 主要言語との組み合わせだけを考慮していればよいかというとそうでもない. トゥイ語 (ガーナ) 版 Wikipedia の記事にケチュア語 (ボリビア) で書かれた記事[*1] http://tw.wikipedia.org/wiki/Kristiyanu_i%C3%B1iy が存在するなど, 状況は複雑である.

ただし, 主要な言語版では他の言語を含む, あるいは他の言語のみで書かれている記事は割合が低く, ほとんど問題がない. また, 編集も活発なため, これらの記事の寿命も短い傾向にあり, おおよそ短期間で解消される. [*2]一方で主要でない言語版においては深刻な問題となりうる.

2.4　チワン語版 Wikipdia における統計

主要でない言語版の実態を示すため, このチワン語版 Wikipedia を[*3]例に実際にカウントを行った結果を示す. チワン語は中華人民共和国に住むチワン族によって用いられる言語であり, タイ・カダイ語族に属する.

なお, テンプレート・見出し・表・リスト・表題・図題・参考文献などは, 短文が多かったり, 他の言語で書かれていたり, URL・固有名詞等のみで成り立っていたりすることが多く, 扱いが面倒なことや, はっきり分類することが難しいことから, あらかじめ除去し, 本文部分のみについてカウントした.

なお, 使用したのは 2012 年 12 月上旬頃のデータである.

[*1] ただし, ケチュア語だけで書かれているのでこの記事そのものは多言語混合文書ではない. しかしながら, 記事ごとの言語判別によるフィルタリングを省略してコーパスを作成した場合, 多言語混合文書が発生する原因となりうるため, 潜在的には多言語混合文書とも言えなくもない.

[*2] そのため, 上のリストは 2014 年には全く役に立たなくなっている可能性が高い

[*3] http://za.wikipedia.org/wiki/

表 2.1. チワン語版 Wikipedia における多言語混合記事数

チワン語のみ	英語を含む	中国語を含む	ドイツ語を含む
580	15	14	1
95.1%	2.4%	2.3%	0.2%

表 2.2. チワン語版 Wikipedia における多言語混合記事のデータ量

チワン語	多言語文書	
	チワン語部分	他の言語部分
116kB	13kB	299kB
27.1%	3.0%	69.9%

まず, 記事数である. 表 2.1 の通り, カウントされた 610 [*4]の記事のうち, 他の言語を含むものは 30 で, およそ 5% であり, 少数派ではあるものの無視できるほど少ないとはいえない. なお, 境界はいずれも段落間にあり, ほとんどが不完全な翻訳に分類され, 2 つの記事を除き, 境界の数はいずれも 1 箇所のみであった. Wikipedia の少数言語における多言語混合文書は, 同様の傾向を示す.

しかしながら, 問題は多言語混合文書のデータ量である. 他の言語を含む記事はどれもサイズが大きいため, 表 2.2 のように, 実に全体の 73% がそのような記事に含まれているという結果になった. 質の低下というより, どの言語のデータを集めているかわからないレベルであり, 多言語混合文書を処理しなければ使用が不可能な状態にある.

なお, 多言語混合文書を捨てずに, チワン語の部分を抽出するとチワン語のデータ量が 11% 増えることから, 言語情報資源を有効活用する意味でも重要であるといえる.

このように, チワン語にかぎらず, 多言語混合文書を処理することは, Wikipedia の少数言語と向き合う上で非常に重要なことであるといえる. [*5]

[*4] Wikipedia の公式統計 http://ja.wikipedia.org/wiki/Wikipedia:%E5%85%A8%E8%A8%80%E8%AA%9E%E7%89%88%E3%81%AE%E7%B5%B1%E8%A8%88 によると, 純記事数は 619 となっている

[*5] もっとも, さらにマイナーな言語になると部分的な翻訳すらされていない場合も多い.

第 3 章

関連研究

この章では, 主に多言語混合文書の分割と言語判別の問題について, 関連研究を紹介する.

3.1 言語に関する文書分割

多言語混合文書に対する認知度はそれほど高くないため, その分割に関する研究はあまり行われていない. 少ないながらも, 言語に関する文書分割の方法を挙げると, おおよそ,

1. 段落, 文, 単語など適当な構文単位で分割して, 個別に言語を判定する方法,
2. 内容の大きく変化する点を検出して分割して, 個別に言語を判定する方法,
3. 文書全体について目的関数を最小化する分割と言語の組を同時に求める方法,

の 3 種類に分けられる. それぞれの分類ごとに紹介する.

3.1.1 適当な構文単位で分割し個別に言語判別する方法

言語に関する文書分割の方法として, 最もよく見られるものとしては, (1) である. 例としては, KDE4 への搭載を目指して開発されていたスペルチェッカー Sonnet があった. [*1] [21] [22] これは, 段落単位 (オプションで文単位も選択できる) であらかじめ分割し, 各言語ごとに用意された 3-gram の確率モデルと辞書を補助的に用いて言語を判別するものであり, 40 言語に対応していたとされる.

今回の問題設定に照らし合わせると,

- 分割の粒度を小さくしようとすると, 判定にかける文字列が短くなってしまい, その分だけ精度が下がってしまう. そのため, 文字, 単語レベルの粒度での分割はほとんど不可能である.

[*1] 記事では KSpell 2 を置き換えること目標に, となっていたが, その後の続報も見つからず, KSpell 2 が現役で使われているので, 開発中止になったのかもしれない. 入手できなかったため, 実際にどのように動作するかを使って確認してみることはできなかった.

- 単語と段落の中間に位置する粒度の文字列分割を用意するのは難しい. Sonnet では文も選択できるようになっているが, 句読点などの約物の使い方は言語ごとに細かな差異があるため, 言語の判別ができていない段階で, 多数の言語に対応する検出は難しいためである. 適合率を犠牲にして, ありとあらゆる句読点らしい箇所で切ってしまう方法もあるが, 落とし穴が存在して, コイサン語族では, 多彩な吸着音を表現するため “!” など文の切れ目に使われるような記号が単語の綴りの一部として使われる. [*2]また, 日本語・タイ語など言語によっては分かち書きがなされないので, 単語境界すらわからないこともある.

といった難点があり, 利用は難しい.

3.1.2 内容の変化点で分割し個別に言語判別する方法

次に, 言語に関するものではないが, 一般に (2) は文書分割によく使われていて, 代表的なものに *text tiling* [14] がある. これは, 文書を適当な単位で分割し, 適当な関数で隣り合うブロック間で類似度を計算し, その 2 階差分が閾値を超えた点で分割を行うものである. [14] ではこれを用いて, 文書のトピックによるパラグラフ単位での分割を行った. 今回のように, ブロックが非常に小さくなるときには, 類似度の計算に前後に連続するブロックも含める, スムージング処理によって精度の低下を抑えるのが普通である. なお, 教師なしで分割するため, 言語判別は分割したあとでやはり個別に行うことになる. (1) と比べると, 分割の最小単位ごとに言語を判別する必要がなくなったので, 性能の向上が見込まれるが,

- スムージング区間が短いと精度が下がり, 長いと短めの別言語区間を見逃してしまうというトレードオフが存在するため, 多言語混合文書に汎用的なパラメータ設定が難しい,
- 2 階差分を使うためノイズに負けやすく, 分割してほしい位置ちょうどになることはあまり期待できない,
- 似ている言語対などで境界が目立たないときには検出できない,

といった問題点が残る.

3.1.3 文書全体について目的関数を最小化する分割と言語の組を同時に求める方法

最後に (3) があり, 提案手法もここに含まれる.

ここに含まれる非常に近い研究として, PPM と呼ばれるデータ圧縮を拡張することによって実現した Teahan の方法 [24] がある. データ圧縮を利用している点も近い. Teahan は, 直前数文字の文脈から次の文字を予測するという PPM の原理を利用した脱字補完を行い, その

[*2] ただし, Unicode では本来見た目がよく似たの記号 U+01C3 (*LATIN LETTER RETROFLEX CLICK*) を使うのが正しいようであるが “!” で代用されることもよくある

応用として, 空白文字を補う単語分割や, 特殊な言語切替文字を補う言語に関する文書分割を提案した. 具体的には, 脱字補完の原理は,

1. 脱字のない文書群から PPM を用いた圧縮器を学習し,
2. 脱字を補った場合と補わなかった場合で出力される符号長を比較し,
3. 最終的に短い方を出力とする

というものであり, これに,

- 各言語ごとに別々に用意した複数圧縮器を使ったモデルにする,
- 特殊な言語切替文字によって, 使用する圧縮器を切り替えるようにする,

といった拡張を行うことで文書分割に対応するようになっている.

この方法は,

- 可変長の区間の言語判別を行うことや, 特殊な言語切替文字の挿入にコストがかかることで短かすぎる区間の考慮を防いでいるため, (1) (2) と違って, 文字列が短いときに言語判別の精度が下がる問題を回避しやすくなっている.
- また, 文書全体について最適化を行うため, (2) のように, 局所的なノイズに影響されにくく, アドホックなスムージングを必要としない.

といった利点があり, 非常に優れているように見える. さらに, モデルに学習以外で決める外部パラメータを持たないため, 学習データさえ用意すれば, 使用できるという点も利点と見える.

しかしながら, この方法を実用化することを考えた場合, 大きな問題点が存在している. それは, 特殊な言語切替文字の挿入コストは, 学習データ量を増やすと発散するパラメータであるということである. すなわち, 一定以上の量の学習データを与えると量を大きくすればするほど性能が低下することを意味している. この原因は, 言語切替文字は通常, 学習データ中に存在しないという点に起因している. PPM は, 統計を用いてゼロ頻度の文字の出現頻度を上手に予測する点で非常に優れたモデルであるが, それに依存した Teahan の手法では, 学習データ中に存在しない文字に適切な符号長を与える手段がないのである.

ここで, Teahan の方法で, 言語切替文字に割り当てられる符号長を見積もってみることにしよう. エスケープ方式は C あるいは D とする. 割り当てられる符号長は, 学習データに存在しない文字なので,

1. エスケープ記号の符号長,
2. -1 次のモデルにおける符号長,

の 2 つの成分からなる. 前者は, エスケープ方式によって細かな違いはあるが, いずれも学習量を大きくすると $\mathrm{O}\left(\log\left|\text{学習データ}\right|\right)$ に漸近する. また後者は, いずれも学習量を大きくすると, 学習データ中に存在しないがアルファベット集合には入っている文字の数を u' として,

$\log u'$ に収束する．すなわち, 実は言語切替文字のコストには隠れたパラメータとして, 学習データ量と学習データ中に存在しないがアルファベット集合に入っている文字の数の 2 つが存在していて, これらは学習できない．一方で, 学習量を大きくすれば, 出力される符号長は理論上は真のクロスエントロピーに収束する[*3]ので, 分割に対するコストはこれと同程度に設定されなければならないはずである．すなわち, いわゆる過学習とは無関係のところで学習データを増やし過ぎたことによる性能低下が存在し, 無関係な文字の数の増減させるというナンセンスな方法によるパラメータ調節が必要な手法ということになる．[*4]むしろ, 使う側からするとアルファベット数や, 学習データ量に関してロバストであってほしいと思われる．よって, このことから, 適切に利用するためには, 多言語混合文書でも, 単一言語のみからなる文書でも構わないが, 言語切替文字を適切に挿入した文書で学習して用いるのが望ましい手法であると考えている．

しかしながら, 深く考えることなく使用する手法としてはそれほど悪くはなく, Teahan は, 聖書の創世記[*5]を利用して, 英語・フランス語・ドイツ語・スペイン語・イタリア語・ラテン語について実験し, かなりよい結果が得られたとしている．ただし, 聖書は内容・文体が揃っているため, かなり綺麗なデータセットであるとみなせることや, ごく限られた言語でしか実験しておらず, 少数言語版の Wikipedia を含むようなセットでうまくいくかは別に検証しなければならない課題だと考えている．

なお, Teahan の方法は非常に多くの組み合わせを考慮するため, 一見非常に遅いように見えるかもしれないが, 動的計画法によって線形時間で処理する方法が, 論文中に具体的には書かれていないため, 詳細は不明であるが, 示唆されている情報から推定すると, 計算量に起因する問題はあまりない．

3.2　データ圧縮に基づかない言語判別

多言語混合文書の分割は, 単一言語からなる文書の言語判別に基礎を置くことになるため, 言語判別問題に関する研究についても述べる．そのような研究は数多く行われており, 中でも重要なものとしてデータ圧縮を利用しないものから 4 つの研究を紹介する．

1 つ目は Grefenstette による研究 [12] である．Grefenstette は文章を構成する単語に注目した,

- 文字 3-gram に分解し, その頻度を数える方法,
- 短い単語のみを抽出して, その頻度を数える方法,

の 2 種類の手法を比較し, 3-gram の方が性能が良いという結果を得たとしている．データ圧

[*3] 文字セットが一致している場合．英語とロシア語のように異なる文字種を使う場合には収束しない

[*4] PPM の圧縮器の実装上も, このようなアルファベット数を増減させる機構を入れるのはかなりトリッキーになる

[*5] 英語版でおよそ 75kB 程度

縮の中でも PPM を用いる方法は, より 3-gram を洗練したものと捉えることもできる.

2 つ目は, Kruengkrai らによる研究 [18] である. Kruengkrai らは, String Kernel を用いた類似度評価を行い, クラスタの重心からの距離の最小化あるいはサポートベクターマシンを用いて言語判別を行い, 非常に高い精度を得たとしている. しかしながら, この手の強力な機械学習の手法は非常に重たいことや, 目的関数を構成するために使いやすい値を得ることが難しいことから, 今回のような問題には向かないと考えている.

3 つ目は, Kikui による研究 [16] である. Kikui は東アジア圏の漢字文字コードの判別とヨーロッパの言語および HTML 実態参照判別を組み合わせた複合問題を扱った. 手法としてはアドホックな方法をとっており, まず, 2 バイト文字を分離し, 1 バイト文字は単語頻度, 2 バイト文字は文字頻度で判別する. 文字コード判別はかつて重要なタスクであったが, 現在では UTF-8 および Unicode がデファクトスタンダードとなったため本論文では扱わない. また, できるだけ言語ごとのアドホックな対応をしないことを目指しており, その点目指すところが異なる.

4 つ目として, Cilibrasi による研究 [7] である. 基本的には n-gram 頻度を用いたものであるが, 最適なモデルの比較に, 確率や尤度ではなく, カルバック・ライブラー距離を用いている点が珍しい.

3.3　データ圧縮に基づく言語判別

今回の問題設定では, 多数の少数言語に対応しなければならないため, その中でも大量の学習データや言語ごとのアドホックな手作業による調整を必要としないことが重要である. その点, データ圧縮を用いた言語判別は条件を満たしており, 本論文でも重視している.

データ圧縮を用いる手法は, 圧縮器の学習方法から,

1. 先に入力テキスト全体を 1 回走査して圧縮器を作成するオフライン型,
2. 入力テキストから逐次動的に学習するオンライン型

の 2 種類に分けることができる.

3.3.1　オフライン型データ圧縮を用いた言語判別

オフライン型のデータ圧縮は,

1. 入力テキスト全体を 1 回走査して圧縮器を作成し,
2. その圧縮器を用いて入力テキストを圧縮する

という 2 段階からなっている. ゆえに, これを用いた言語判別では, 1 段階目で入力テキストの代わりにあらかじめ用意した各言語の学習データを与えることで, それぞれの言語用の圧縮

器を作成することができ, これらで圧縮した際の符号長[*6]の最小値をとることで言語が判別できる. この符号長には, 確率論におけるクロスエントロピーの近似になっているというよい性質があり, 数学的な裏付けもあり扱いやすい. そのため, 本論文では, これらのオフライン型のデータ圧縮を利用した言語判別を元に, 多言語混合文書の分割を行う.

オフライン型のデータ圧縮を用いた研究としては, 統計型の Prediction by Partial Matching (PPM) [9] [5] を用いた Teahan らの研究 [24] [23] や, 辞書型の Lempel-Ziv [27] を簡略化したものとして, matching statistics の平均によるクロスエントロピー推定法 [11] [15] を用いた, Juola による研究 [15] などがある. 詳しくは 5 章で説明する.

3.3.2 オンライン型データ圧縮を用いた言語判別

オフライン型とは違い, あらかじめ学習データから圧縮器を作っておくことが難しいため, 異なる方法で類似度を計算する. 例えば Benedetto らによる場合 [6] は, 学習データと入力文字列を連結したものと学習データのみを, それぞれ実際にデータ圧縮して出力された符号長の差をとるということが行われる. この量は, 辞書サイズが有限であることによる忘却効果や, 入力文字列からも学習するという性質により, オフライン型の場合の学習データと入力文字列の総合的な差とは違い, 連結部における変化の大きさを強調した量になる. それゆえ, 入力文字列が長くなっても差がつきにくい, クロスエントロピーといったわかりやすい概念とは対応しなくなることもあって, 扱いづらさもあるため, 今回の問題に利用するのは難しいのではないかと考えている.

一方で, 学習データに含まれにくいが, 入力テキスト中で繰り返し登場しやすい, 専門用語や固有名詞などに強いと考えられ, オフライン型のデータ圧縮にこれらの要素を加えることは今後の研究課題である.

オンライン型のデータ圧縮を用いた研究としては, Benedetto らによる研究 [6] や, Cilibrasi らによる研究 [7] があり, 主に gzip が用いられている. なお, 既存のデータ圧縮ツールを使い, 実際に符号を出力しているということに起因する問題もあり,

- 実際に符号を出力するため非常に遅い,
- 1bit あるいは 8bit 単位の離散的な値しか得られない,
- ハフマン符号化を経ることによる近似損失がある,
- 途中状態を効率よく保存できない,

といった難点もある.

[*6] 実際には, いずれも符号の生成は行わず, 実数で得られる理論値のみを使用している.

3.4　その他

多言語混合文書の分割に近いことを行うものに Ehara らによる, PPM を使った言語切り替え予測機能付きの IME の研究 [10] がある. オンラインで処理するため, 文書全体について最適化することができないが, 一方でインタラクティブであるため, 正解のフィードバックを得ることができるという大きな違いがある. 特に今回の問題設定では得られない, 言語の切り替わりやすい文脈や遷移しやすい言語対を学習できるという点は非常に重要である. なお, 変換の単位にあわせて, 単語単位で言語の単位を判定している.

また, 単語単位での分割で今回の問題設定により近いものとしては, Alex らのドイツ語文中の英単語を指摘する研究 [3] [4] がある. 前処理として英単語を指摘することによってドイツ語文の構文解析精度を上げることに成功したと報告されている. しかしながら, 主言語がわかっていることが前提の形態素解析の情報や, 大規模な辞書の検索, 特徴的なスペルによるヒューリスティクスの情報などを用いて, ようやく高い精度を達成していることから, これを応用して多くの少数言語に対応することは難しい. このように全く別の問題となることがわかる.

第 4 章

符号長最小化問題としての定式化

この論文では, 本論文で提案する, 多言語混合文書の分割問題のデータ圧縮を用いた符号長の最小化問題としての定式化を具体的に行う.

4.1 問題設定

定式化の前に前提を確認する.
本論文では, 多言語混合文書の分割問題について,

1. 多言語混合文書に現れうる言語の集合は既知であり,
2. そこに含まれる各言語についてテキストデータが存在する

と仮定を置くことにする. すなわち, 分割は教師ありであるともいえる. ベイズを利用した教師なしの分割手法が最近の流行であるが, さらなる処理のために適切な解析器を選ぶといった処理をするためには, 教師ありであることが重要な意味を持つと考えているため, あえてこのようにしている.

4.2 表記

この章と 5, 6 章で使われる表記について説明する.

まず, X を多言語混合文書として, i-番目の文字を x_i, 長さを $|X|$ と表記することにすると, $X = x_1, \ldots, x_{|X|}$ と表記される. 次に, $\mathbb{B}$ を X の言語境界位置のリストとして, 境界の数を $|\mathbb{B}|$, i-番目の境界の位置, すなわち多言語混合文書の先頭から何文字目に境界があるかを B_i と表すことにすると, $\mathbb{B} = [B_1, \ldots, B_{|\mathbb{B}|}]$ と表記される. 『言語に関する文書分割』とは, この $\mathbb{B}$ を求める過程と言い換えることもできる. なお, 今回の分割問題では境界の位置を任意の文字間に取ってもよいことにしているため, B_i は単語中の位置を示すこともあるので注意してほしい. さらに, $\mathbb{X}$ を分割された多言語混合文書とすると, 連結すると X に一致する文書片のリストを用いて $\mathbb{X} = [X_0, \ldots, X_{|\mathbb{B}|}]$ と表すことができる. ここで, X_i は元の文書における

B_{i-1} 文字目から $B_i - 1$ 文字目の部分文字列になる. 最後に, 言語集合を $\mathcal{L}$, $L_i \in \mathcal{L}$ を文書片 X_i に用いられた言語とすると, $\mathbb{X}$ のそれぞれに対応する言語を $\mathbb{L} = [L_0, \ldots, L_{|\mathbb{B}|}]$ と表記される. ただし, 表現の冗長性をなくす意味から $\mathbb{L}$ の隣り合う要素対は同じ言語であってはならないとする.

4.3　定式化

多言語文書の分割問題の定式化にはいくつか方法があると思われるが, 本研究では, 最小記述長原理 [20] を参考に, データ圧縮を用いた符号長最小化問題として, 次のように定式化した. なお, Teahan とは異なり, データ圧縮手法に依存したものではないため, この定式化の下で, 任意のオフライン型データ圧縮を自由に使用することできる. [*1]

問題 4.1　多言語混合文書 X が与えられたとする. その最適な分割の個数と分割位置と言語の 3 つ組 $(\hat{K}, \hat{\mathbb{B}}, \hat{\mathbb{L}})$ は,

$$(\hat{K}, \hat{\mathbb{B}}, \hat{\mathbb{L}}) = \arg\min_{K,\mathbb{B},\mathbb{L}} \sum_{i=0}^{K} \{-\log_2 P_{L_i}(X_i) + \log_2 |X| + \log_2 |\mathcal{L}| + \gamma\} \tag{4.1}$$

と与えられる. ただし, $K = |\mathbb{B}| = |\mathbb{L}| - 1$, 任意の $i \in \{1, \ldots |\mathbb{B}|\}$ について $L_i \neq L_{i-1}$ でなければならない.

この定式化において $\sum$ 内の, 第 1 項 $-\log_2 P_{L_i}(X_i)$ は, 言語 L_i 用の圧縮器による X_i の符号長, あるいはクロスエントロピーであり, これにより, 最適な言語と分割位置が選択される. 詳細な計算法については, 次の 5 章で説明する. また, 第 2・第 3 の項は最小記述長原理を参考に[*2]モデルの記述長を加えたものであり, それぞれ, 分割位置と使用する言語を指定する項である. これによって最適な分割数を選択でき, 過剰な分割に対して損失を与え, 抑制する働きをしている. ただし, 過剰な分割を防ぐ損失項は, 実際には第 2・第 3 項だけでは足りないことがわかっており, それを補うための追加の損失項が γ である. これがどのような理論的意味を持つかに関してはいくつか仮説を立てているものの, 結論はいまだ出ておらず, 今後の課題である. しかしながら, これは実用上は定数でよく, 学習用データセットごとに適当な機械学習アルゴリズムで定めればよい.

この定式化された問題は, 動的計画法を用いて効率良く解くことができる. 詳しくは 6 章で説明する.

[*1] ただし, 計算量を気にしないとする

[*2] 普通の最小記述長原理ではデータ数に応じて離散化幅を小さくしていくが, この問題は分類問題のため, 離散パラメータの個数は変化させられない

4.4 Teahan の方法との比較

Teahan の方法で単一言語のみからなる文書で学習した場合と比較すると, データ圧縮を用いて最適な言語と分割位置を求める点や, 分割にコストがかかるようにすることで分割の個数を調節するという点で非常によく似ている. しかしながら, 以下のような違いもある.

- コストの根拠に, Teahan は PPM のゼロ頻度記号の出現確率予測機構を使っているが, 本研究は最小記述長の考え方に基づいている.
- Teahan と違って特定のデータ圧縮手法の機能に依存しないため, データ圧縮手法に選択の自由がある.
- Teahan と違ってパラメータが陽な形で現れているので調節が簡単である.
- コスト計算に, Teahan の方法では, 学習データ中の n-gram の出現頻度・学習データサイズ・出現しないアルファベットのサイズを陰に用いているだけなので, 各言語ごとに学習データのみから決めてしまっているため, 他の言語との近さや, 入力文字列の性質を全く考慮していないが, 提案手法では, 入力文字列の長さ・言語数・適宜定める定数を用いているため, それらに対応することができる.
- Teahan の方法と違い, 分割コストに文脈の情報を全く利用していないため, その点は劣る可能性がある.

第 5 章

データ圧縮を用いた符号長計算

本研究では, 符号長計算あるいは, クロスエントロピー推定のため, オフライン型データ圧縮手法の中でも, 代表として辞書型の MMS, 統計型の PPM の 2 つの手法を用いる. これらについて詳細と実装上のテクニックについて解説する.

5.1 Mean of Matching Statistics (MMS)

辞書型のデータ圧縮手法として Lemple-Ziv 77 圧縮 [27] がよく知られているが, その簡略化として matching statistics と呼ばれる, 方向性のある文字列間の統計量を用いる方法がある. 本論文では, これを “Mean of Matching Statistics” (MMS) と呼ぶことにする.

これは Farach らによって提案された DNA のエントロピー推定手法 [11] を元にしている. Farach は次のような定理を証明した.

定理 5.1（Farach の定理） 文字列 $X = x_1x_2 \ldots x_ix_{i+1} \ldots$ に対して, Len_i を

1. $x_{i+1}x_{i+2} \ldots$ の接頭辞で,
2. $x_1x_2 \ldots x_i$ の部分文字列でない

ような文字列の中で最短のものの長さと定義する. また, Len_i の平均を $E[Len]$ と表すことにする. 文字列 X が定常エルゴードな有限次マルコフ情報源により生成されたという条件の下,

$$\lim_{i\to\infty} \left| E[Len] - \frac{\log_2 i}{H(X)} \right| = 0 \quad \text{(確率収束)} \tag{5.1}$$

が成り立つ.

この定理を用いて, Farach は文字列 X の 1 文字あたりのエントロピー $\hat{H}(X)$ を

$$\hat{H}(X) = \frac{\log_2 i}{E[Len]} \tag{5.2}$$

という式により推定した.

Juola はこれを応用してクロスエントロピーを推定する方法を提案した.[15] 具体的には, Farach の定理において (2) をあらかじめ用意した学習用パターン文字列に置き換えることで, 学習用文字列に基づくモデルでのクロスエントロピーの推定に用いることができるという発想に基づいている. すなわち, 次のような予想となる.

予想 5.1（Juola の予想） 文字列 $X = x_1x_2 \ldots x_ix_{i+1} \ldots$ と $Y = y_1y_2 \ldots y_{|Y|}$ に対して, $Len_i(Y)$ を

1. $x_{i+1}x_{i+2} \ldots$ の接頭辞で,
2. $y_1y_2 \ldots y_{|Y|}$ の部分文字列でない

ような文字列の中で最短のものの長さと定義する. また, $Len_i(Y)$ の平均を $E\left[Len(Y)\right]$ と表すことにする. 文字列 X と文字列 Y が定常エルゴードな有限次マルコフ情報源により生成されたという条件の下,

$$\lim_{i\to\infty} \left| E\left[Len(Y)\right] - \frac{\log_2 |Y|}{H(X)} \right| = 0 \quad \text{(確率収束)} \tag{5.3}$$

が成り立つ.

Juola はこの予想に基づき,

$$\hat{J}_Y(X) = \frac{\log_2 |Y|}{E\left[Len(Y)\right]} \tag{5.4}$$

と 1 文字のあたりのクロスエントロピーを推定した.

しかしながら, 1 文字ごとのクロスエントロピー $\log_2 |Y|/Len_i(Y)$ に注目すると, この推定値は調和平均になっているために, 非線形であり, 加法的な性質を持たず, 実は本論文の定式化と相性が悪いことがわかっている. 具体的には,

- 境界付近の文字を割り当てるとき, グループの大小により, 同じ言語だとしてもコストが変わってくるため, 出力が非常に大きくずれてしまい, ほとんど使いものにならない.
- 6 章で説明する線形時間アルゴリズムを構成することが難しい,

といったものである. そのため, 本論文では少し推定式を変更して,

$$\hat{J'}_Y(X) = E\left[\frac{\log_2 |Y|}{Len_i(Y)}\right] \tag{5.5}$$

のように, 算術平均を用いたものを使うことにした. [*1]

なお, Juola の推定式も変更された推定式も, 証明されていない予想に基づいていて, 数学的な裏付けがないことに注意してほしい. しかしながら, 実験的には一定値に収束しており, 実用上は特に問題ない.

[*1] Matching statistics の平均の取り方に注目した場合は, 算術平均から調和平均への変更である.

ちなみに, この $Len_i(Y)$ が, 一般に matching statistics と言われる量であり, [*2] 名前の由来となっている.

PPM と比較すると, 仕組みが単純なので高速・軽量である.

5.2　MMS の実装

Matching statistics は, 一般に suffix link つき接尾辞木を用いて, 線形時間で計算できるアルゴリズム [13] がよく知られている. 具体的には, アルゴリズム 5.1 のようにすればよい.

Algorithm 5.1 接尾辞木を用いた matching statistics の計算

```
i ⇐ 0
j ⇐ 0
接尾辞木 s のカーソルを root にリセット
while i < |X| do
  while X_j が s にマッチする do
    j++
    s のカーソルを X_j の方向に動かす
  end while
  Len_i ⇐ j − i + 1
  s の suffix link をたどる
  i++
end while
return [Len_i]_i
```

あらかじめ学習データから, 接尾辞木を作成しておいて, 入力文字列ごとに matching statistics を計算することになる.

接尾辞木の構築には Ukkonen のアルゴリズム [25] など, 様々なアルゴリズムが知られている. 接尾辞木の消費メモリが気になる場合には, DAG にする, あるいは, 強化接尾辞配列 (enhanced suffix array) [1] を使うこともできる.

5.3　Prediction by Partial Matching (PPM)

辞書型とは別に, 統計型と呼ばれるデータ圧縮手法の大きな流派が存在する. その代表として Prediction by Partial Matching (PPM) [9] という非常に効率のよい手法が知られている. 直前の n-gram から次の文字を予測する, 文脈を考慮した方法であり, ゼロ頻度のデータに対してはエスケープ記号を介した, 低次のモデルによるスムージングを行う.[5] $n+1$-gram

[*2] 文献によってはここの定義よりも 1 小さい値を定義とする場合もある.

モデルの変種ともいえ, エスケープ記号を用いたスムージングは, 自然言語処理の分野では “Witten-Bell のスムージング” と呼ばれる.

式で表すと, 学習データを Y として, 入力文字列 X 中の x_i の確率は, 直前の文脈として n-gram が与えられたときの $n+1$-gram の条件付き出現確率 P_Y とエスケープ記号の確率 e を用いて,

$$P_Y(x_i|x_1 \dots x_{i-1}) = \begin{cases} \{1 - e_Y(x_{i-n} \dots x_{i-1}\} \times p_Y(x_i|x_{i-n} \dots x_{i-1}) & (x_{i-n} \dots x_i \text{ が } Y \text{ に存在するとき}) \\ e_Y(x_{i-n} \dots x_{i-1}) \times p_Y(x_i|x_{i-n+1} \dots x_{i-1}) & (x_{i-n} \dots x_i \text{ が } Y \text{ に存在しないとき}) \end{cases} \tag{5.6}$$

と見積もられ, 全体の符号長は

$$\log_2 \prod_{t=0}^{|X|-1} P_Y(x_t|x_{t-1} \dots x_{\max(1,t-n)}) \tag{5.7}$$

と計算される. なお, エスケープ確率の計算にはいくつかの流派があるが, 本論文では, 総記号数を n と, 記号の種類数 u を用いて,

$$e = \frac{u}{n+u} \tag{5.8}$$

と表される method C [19] を用いた. また, 次数は Cleary らの研究 [8] の結果を参考に 5 次に設定した.

MMS と比較すると, より精密なモデルを使っており, 性能が高いと期待される.

5.4 PPM の実装

PPM の実装には文脈の文字列と文字の頻度表をハッシュテーブルで関連付ける実装が簡単であるが, 次数が上がるにつれてメモリ消費量がかなり辛くなる. また, それによりキャッシュに載りきらなくなるためアクセス速度も低下する Cleary らの研究 [8] で触れられているように, Suffix link つきの接尾辞 Trie あるいは接尾辞木で, 各エッジに出現回数を保存する拡張をすると効率よく実装できる. [*3]

また, 高次のモデルに出現する記号の確率を 0 にする, 除外処理を実装するのが普通であるが, 高次のモデルに応じて頻度表を書き換えなければならないことから非常に重い処理である. しかしながら, エスケープで移動する先は suffix link で指定された先に一意に定まるので, あらかじめ用意しておくことができる. なお, 符号の出力を行うなら頻度表全体を持っていなければならないが, 今回必要なのは符号長のみなので, 除外処理後の記号総数と記号の種類数だけを持っておけば十分である.

これらを利用した, PPM による符号長計算はアルゴリズム 5.2 のようになる.

[*3] MMS とは異なり, 強化接尾辞配列を使う方法は頻度表との関連付けが複雑になるため, あまりが効果ないと思われる.

Algorithm 5.2 接尾辞 Trie を用いた PPM の符号長の計算

```
i ⇐ 0
接尾辞木 s のカーソルを root にリセット
while i < |X| do
  n ⇐ (今の文脈での記号総数)
  u ⇐ (今の文脈での記号種類数)
  L_i ⇐ 0
  while X_j が s にマッチしない do
    L_i ⇐ L_i+ (今の文脈でのエスケープ確率 (n, u))
    n ⇐ (1 文字減らした文脈での除外処理をした記号総数)
    u ⇐ (1 文字減らした文脈での除外処理をした記号種類数)
    s の suffix link をたどる
  end while
  L_i ⇐ L_i+ (今の文脈での文字の出現確率 (n, u, x_i))
  s のカーソルを X_j の方向に動かす
  if 文脈の長さが次数制限を超えた then
    s の suffix link をたどる
  end if
end while
return [L_i]_i
```

このアルゴリズムは, エスケープ方式として A, B, C, D を想定しているが, P, X などおおよその方式において, 微修正により同様にできる.

第 6 章

動的計画法に基づく線形時間アルゴリズム

第 4 章で定式化した多言語混合文書の分割問題を実際に解く方法について説明する. 動的計画法を用いることにより, 入力長の線形時間で計算することができる. [*1]

この章での説明では, 説明の煩雑さを避けるため, 隣り合うブロックの言語が異なっていなければならないという制約を省略した緩和問題で説明を行う. なお, 万が一, 一致してしまったときにはマージすればよい.

6.1 素朴な動的計画法

定式化された問題は組み合わせ最適化の問題であり, 総当たりによる方法では, 最適解を求める計算時間が指数爆発してしまうため, 現実的な定式化に見えないかもしれない. しかしながら, この問題にはよい性質があるため, 動的計画法を用いて多項式時間で解決できるクラスに属する.

一般に, 目的関数を最小化するようなリストを動的計画法を使って求める問題は,

1. 動的計画法を用いて目的関数の最小値を求め,
2. (1) の計算に用いた表を逆にたどって経路を復元する (バックトラック)

という, 2 段階で構成される. このうち, 1 段階目について説明する.

目的関数の最小値を求めるアルゴリズムを導出する. 動的計画法のアルゴリズムを構成するには, 漸化式を求め, そこから動的計画法の表の次元と, 表のマス間の遷移の集合とコスト関数を構成するのがよい. まず, 最小化の式は, $C(|X|) = \log_2 |X| + \log_2 |\mathcal{L}| + \gamma$ とおいて,

$$\min_{K,\mathbb{B},\mathbb{L}} \sum_{i=0}^{K} \{-\log_2 P_{L_i}(X_i) + C(|X|)\} \tag{6.1}$$

[*1] 実際の分割の個数には全く依存しない

と与えられる. [*2] これを漸化式の形に変形すると,

$$DP(0) = 0 \tag{6.2}$$
$$DP(i) = \min_{t\in\{0,\ldots,i-1\},L\in\mathcal{L}} \{DP(t) - \log_2 P_L(x_t \ldots x_{i-1}) + C(|X|)\} \tag{6.3}$$

となる. この式から,

表の次元　入力文字列中の位置 i
遷移　直前の分割位置 $t(\in \{0,\ldots,i\})$ と言語 $L(\in \mathcal{L})$
コスト関数　$[t,i)$ を L で圧縮した符号長

となる. これを用いると, アルゴリズム (6.1) のようになる.

Algorithm 6.1 素朴な動的計画法による分割アルゴリズム

```
DP(0) ⇐ 0
for i = 1 to |X| do
  DP(i) ⇐ ∞
  for t = 0 until i do
    for all L ∈ 𝓛 do
      DP(i) ⇐ min{DP(i), DP(t) − log2 P_L(x_t ... x_{i−1}) + C(|X|)}
    end for
  end for
end for
return DP(|X|)
```

ここでアルゴリズムの計算量を考えてみよう. これは, ループの数を見てもわかるが, 表の大きさが $|X|$, 遷移の数がそれぞれ $\mathrm{O}(|X| \times |\mathcal{L}|)$ であることから, 全部で $\mathrm{O}(|X|^2 \times |\mathcal{L}|)$ 回の遷移が行われることになる. ここで注意しなければならないことがあり, それぞれの遷移ごとに符号長を使用しているので, 素朴な方法では毎回 $\mathrm{O}(|X|)$ がかかってしまうということである. [*3]すなわち, 計算量は $\mathrm{O}(|X|^3 \times |\mathcal{L}|)$ となり, 多項式時間ではあるものの実用性に乏しい.

6.2　途中状態のキャッシング

しかしながら, 大抵の場合は, 簡単なテクニックにより符号長の計算をならしで $\mathrm{O}(1)$ にすることができる. それは, 圧縮器の途中状態をキャッシュするようにして, その時点での符号長

[*2] 5 章では 1 文字あたりのクロスエントロピーが出てきたが, ここで使うのは文字列のクロスエントロピーである. よって, 文字数倍する必要がある.

[*3] これは必要な計算量の下限である. もしかするとより計算量の必要な手法が存在するかもしれないが, データ圧縮としての実用性があるのかははなはだ疑問であり, ここでは考慮しない

を記録していくというものである. [*4] これを使うとアルゴリズム 6.2 のようになり, 計算量は $\mathrm{O}(|X|^2 \times |\mathcal{L}|)$ となる.

Algorithm 6.2 キャッシングを利用した動的計画法による分割アルゴリズム

```
DP(0) ⇐ 0
for i = 1 to |X| do
  DP(i) ⇐ ∞
end for
for t = 0 until |X| do
  for all L ∈ 𝓛 do
    圧縮器 c_L を用意する
    for i = 1 to |X| do
      c_L に x_{i-1} を投入する
      DP(i) ⇐ min{DP(i), DP(t) + (c_L の累積符号長) + C(|X|)}
    end for
  end for
end for
return DP(|X|)
```

かなり改善されたものの, やはりこの種のシステムは線形時間で処理できることが好ましい. しかしながら, これ以上の改善は一般的な性質を利用することが難しいため, ここからは PPM と, MMS に分けて説明する.

6.3 PPM を使った線形時間アルゴリズム

PPM は, 直前の文脈に依存するものの, その長さに制限を設けているため, モデルの最高次数以上の長さにおいては, 出力される 1 文字の符号長が同じになる. よって, これらをまとめることができる. これに基づくと, 漸化式を, PPM の次数を n として,

$$DP(i, L) = \min \left\{ \begin{array}{l} DP(i-1, L) - \log_2 P_L(x_i | x_{i-n} \dots x_{i-1}), \\ DP(i-n) - \log_2 P_L(x_{i-n} \dots x_i) \end{array} \right\} \tag{6.4}$$

$$DP(i) = \min \left\{ \begin{array}{l} \min\limits_{t \in \{i-n, \dots, i-1\}, L \in \mathcal{L}} \{DP(t) - \log_2 P_L(x_t \dots x_{i-1})\}, \\ \min\limits_{L \in \mathcal{L}} DP(i, L) \end{array} \right\} + C(|X|) \tag{6.5}$$

のように書き換えることができる. この式に従うと, アルゴリズム 6.3 のようになる.

[*4] ここで実際に符号を生成するツールを利用すると, 1bit / 1byte 単位で出力される都合や, 高速化のためのバッファリングなど種々の要因により, うまく値を得ることが難しいので, ありものを使うのは推奨できない.

Algorithm 6.3 PPM を使った線形時間アルゴリズム

```
DP(0) ⇐ 0
for i = 1 to |X| do
  DP(i) ⇐ ∞
  for all L ∈ ℒ do
    DP(i, L) ⇐ ∞
  end for
end for
for t = 0 until |X| do
  for all L ∈ ℒ do
    DP(t) ⇐ min{DP(t), DP(t, L) + C(|X|)}
  end for
  for all L ∈ ℒ do
    圧縮器 c_L を用意する
    for i = 1 to min{t + n, |X|} do
      c_L に x_{i−1} を投入する
      DP(i) ⇐ min{DP(i), DP(t) + (c_L の累積符号長) + C(|X|)}
    end for
    if t + n < |X| then
      DP(t + n, L) ⇐ min{DP(t + n − 1, L), DP(t) + (c_L の累積符号長)} + c_L(x_{t+n})
    end if
  end for
end for
return DP(|X|)
```

このアルゴリズムの計算量を見積もると $\mathrm{O}(n \times |X| \times |\mathcal{L}|)$ となり, 線形時間を達成することができる.

なお, このアルゴリズムは, 理論的背景が異なるものの, 式が似ている Teahan の分割手法 [24] に使われているアルゴリズムとよく似ていて, 実際, ある種の変形とみなすこともできる. 最小化したい目的関数や, 有向非巡回グラフ上で計算しているなど細かい違いがあるものの, 本質的にはほぼ同等の内容を実現している. $C(|X|)$ を適切に置き換えるだけで, 簡単に Teahan の手法のために転用することもできる.

6.4 MMS を使った線形時間アルゴリズム

MMS も PPM と全く同じ方法で線形時間アルゴリズムを達成できるかというと, そうではないのであるが,

- PPM が直前の文脈との一致を見ていた一方で, MMS が基礎としている matching statistics は, 直後の文脈との一致を見ていることに相当する,
- PPM と違って一致長にモデルのパラメータによる制限がついていないが, matching statistics そのものであり, この値は有限である,

といった PPM 都の共通点を見出すことができる. これを参考にすると, PPM とは逆向きに文字列の末尾から先頭に向かって, 動的計画法を行い, matching statistics を超えた長さに関して 1 つにまとめるといった方法で, 似たようなアルゴリズムを導出することができ, これが求めるべきアルゴリズムになっている.

計算量であるが, MMS では考慮すべき一致長がどの程度になるか自明ではないが, 平均一致長が 1 番長くなるのは, モデルが一致するときなので, 5 章の Farach の定理により, 高い確率で上限を $O(\log |Y|)$ (Y は学習データ) と抑えることができる. ゆえに, $O(\log |Y| \times |X| \times |\mathcal{L}|)$ となり, 通常は $\log |Y| \ll |X|$ であるため, この場合も線形時間である.

なお, 学習データと全く一致する入力を与えた場合など, 計算量を抑えることに失敗する場合も存在して, このようなとき非常に時間がかかってしまうという脆弱性が存在する. Farach の定理と大体 1 文字あたり 2 bit 以上の情報を含んでいるという観察結果から, 実際に使うときには $\log |Y|$ 以上の一致はあまりないとみなすことができる. ゆえに, それ以上の一致を切るといった近似処理を行なってもほとんど影響がなく, 少ない副作用で安全性を担保できる.

6.5 その他のデータ圧縮の場合

PPM, MMS のいずれも, 依存する文脈が入力に依存せず十分短いことが, 線形時間アルゴリズムの肝となっている. そのため, 現状では, 例えば, 固有名詞, 専門用語など局所的によく出る単語を考慮して補正する手法などには適用できない. しかしながら, 動的計画法の高速化として一般的なテクニックが幾つか知られており [17],[26], [2], この辺りの手法で $O(|X| \log |X|)$ にできるのではないかと予想している. これも今後の課題である.

6.6 隣りあう言語が異なるという条件について

ここまでの説明では, 説明を簡単にする意味もあって, 隣りあうブロックの言語が同じになってしまったときには, ブロックをマージして, どの隣り合う言語も異なるようにする, と説明してきた. これは厳密には第 4 章で定式化された問題を解いておらず, あくまで解いたのは

緩和問題である. 解が一致するかは保証されないので, 細かいことを言えば, 動的計画法を解くときに直前の言語の条件を増やして解くべきである.

しかしながら, 実は, 7 章の実験ではほとんどすべてのテストケースにおいて, 出力が完全に一致するという結果を得ている. すなわち, この条件に本質的な意味はないということになる. ゆえに, 実用上特に気にする必要はない.

なお, 隣り合う言語が異なるという条件をつけると, 素朴に考えれば, 直前の言語, という次元が 1 個増えてしまうため, 言語数の 2 乗の時間がかかるようになると思えるかもしれない. 実は, 最適解と 2 番めの最適解のみを保存するというテクニックにより, やはり言語数の線形時間で計算することができる. ゆえに, 漸近的な計算量の意味では手抜きと厳密な方にほとんど違いはなく, どちらでも好きな方を使ってよい. [*5]

[*5] ちなみに, 今回実験に用いた実装では, 文書内の言語数が多いときにはほとんど差がなかった

第 7 章

人工データを用いた多言語混合文書分割手法の評価実験

この章では, 人工の多言語混合文書を用いた実験とその結果について述べる.

7.1 データセット

多言語混合文書には, 分割位置や, 使用言語のタグがついた大規模データセットが存在していない. そのため, 実験には, 単一言語のみからなる文書で構成される多言語コーパスとして, 世界人権宣言[*1]と Wikipedia [*2]を用いて, 人工的に多言語混合文書を作成し, 実験に用いることにした.

言語のまとめは, 表 7.1 の通りであり, 詳細なリストは付録 A にある.

言語の列挙は,

- 10kB 程度データが入手できたものはできるだけすべて扱う,
- 狭義には独立言語ではなく方言とみなされるものも 1 つとカウントする,
- 文字種の異なる複数の表記法が存在する場合には, 基本的にそれぞれ別言語とみなす,
- 同じ文字種の正書法の違いは 1 つに統一する.

というポリシーで行った. 同じ言語が複数カウントされていたり, ケチュア語関連が非常に多く存在したりするのはそのためである.

なお, 全て文字種が異なる場合の言語セットは, 付録 A でいうところの各言語で番号が 1 となっているものたちと, その他に分類されているものを合わせたものを使っている.

注意してほしいことは, 文字種によって言語の多様性にばらつきがあるということである. ラテン文字・キリル文字は語族・地域・文化・歴史など, 言語の類似性に影響を与えそうなパラメータが異なるものが多くあり, データセット内の多様性が概ね確保されているが, デーヴァ

[*1] http://www.ohchr.org/EN/UDHR/Pages/Introduction.aspx http://unicode.org/udhr/

[*2] http://download.wikimedia.org/

表 7.1. 文字種とのべ言語数

文字種	世界人権宣言 (UDHR)	Wikipedia(Wiki)
ラテン文字	295 (298)	165 (172)
キリル文字	24 (25)	30 (33)
アラビア文字	7	13
デーヴァナーガリ	7	7
ヘブライ文字	2	3
カナダ先住民文字	3	1
ゲエズ文字	2	1
漢字	1	4
ベンガル文字	1	2
グルジア文字	1	2
ギリシア文字	1	2
チベット文字	1	2
その他	16	20
文字種	28	32
言語	361 (365)	253 (263)

ナーガリ, 漢字, ベンガル文字, グルジア文字, ギリシア文字, チベット文字に関しては, 距離が近いものが多くなってしまっている. 結果に見るときにはそこに注意してほしい.

7.1.1　世界人権宣言

今回用いたデータセットの 1 つである世界人権宣言は, 非常に内容が綺麗で, ほぼ理想的な場合のデータセットとして用意した. 具体的には,

- 各文書に書かれている内容が統一されていて,
- 文書内の文体・言葉遣いが統一されていて,
- 他の言語由来の単語を極力使わないように気をつけられている

といったものである. このため, 世界人権宣言における結果は, 各手法の性能の上限であると考えても構わない.

分量としては, 平均 10,000 文字程度であり, 比較的少ない. 本研究では, 『少量のテキストデータ』の基準としてこの値を想定する.

なお, データセットの入手元としては, OHCHR のサイトではなく, *UDHR in Unicode* を利用した. これは,

- OHCHR のサイトでテキストデータとして手に入るものは, ラテン文字の言語に極端に偏っている.
- 文書の論理構造を示した XML のデータが手に入り, OHCHR のものよりも間違いが少ない,
- Unicode になっているので扱いやすい,

といった理由からである. データセットとしては, XML のタグに従って, 前文, 条番号を除去し, 条文部分のみを使用した.

なお, 前述の XML には, 見出し部と本文のタグが逆になっているものなど, 間違いがいくつかあったので, それらは英語版とすりあわせつつ, 手作業で修正を行なってある. また, ペレネ・アシェニンカ語 [prq] とカシナワ語 [cbs] は内容が全く同じであり, 言語タグが信用できないことから除外することにした.

7.1.2 Wikipedia

今回用いたデータセットのもう 1 つは Wikipedia であり, 現実の言語使用を色濃く反映したデータセットとして用意した. これは, 本研究の適用対象として想定されている対象でもある.

Wikipedia からデータセットは次のように作成した.

1. XML のデータから ページ名に “:” を含まない記事について, wiki のテキストを抽出する.
2. テンプレート・見出し・表・リスト・表題・図題・参考文献・40 文字以下の段落・見出しのない記事は, 短文が多かったり, 他の言語で書かれていたり, URL・固有名詞等のみで成り立っていたりすることが多く, 使いづらいため除去する.
3. Wiki 記法によるマークアップを除去する
4. 各言語ごとに主な文字種によって記事を分類する. ただし, 主な文字種が異なるような段落が存在する記事は別にする.
5. UDHR を学習データに用いて, 記事単位でデータ圧縮に基づく言語判別を行い, 明らかに異なる言語が含まれているとなった場合には除去する. グレーの場合はそのままにする.
6. ランダムに 10,000 文字くらいになるように段落を集める. 5,000 文字をきるようであれば諦める.
7. 定型文が多いことからほとんど同じ文が複数含まれてしまう場合であったり, 取りきれなかった URL などのゴミが多くなりすぎる場合もある. そういったときは手作業で掃除をした上で作りなおす.

こうして念入りに洗浄したデータセットではあるが, やはり, 取りきれなかったノイズは存在する. 現実に即した非常に汚いセットである.

7.1.3　識別が困難な言語の扱い

本研究は, 言語判別の技術をベースにしていることから, データ圧縮による言語判別がほとんどできなかった言語は, ひとまとめにする. 付録 A でグループとされているのはこれによるものである. そのような言語としては,

- セルビア語・ボスニア語・クロアチア語・セルボクロアチア語のグループ,
- コンゴ語・キトゥバ語のグループ,
- マレー語・インドネシア語・バニュマス語のグループ,
- スペイン語・チャバカノ語のグループ,
- トルコ語・ガガウズ語のグループ,
- モンゴル語・ブリヤート語・カルムイク語のグループ

があった. また, 標準中国語は様々な中国語系言語の要素を包括してしまっているようなので, 少し恣意的ではあるが, データセットから取り除くことにした.

他にも系統的に近い言語において, 英語とスコットランド語など区別がやや難しい組があったが, これらは区別できないわけではないため, そのまま残した。

7.2　予備実験: 単一言語からなる文書の言語判別

多言語混合文書の分割実験を行う前に, より単純な設定で実験を行う. まず, 単一言語のみで書かれた文書の言語判別実験である.

実験は, 世界人権宣言, Wikipedia の各データセットごと, および PPM, MMS の各データ圧縮手法ごとに, 5 分割交差検定を用いて行う. [*3] 訓練パート全体を用いて学習した圧縮器を用い, 1 言語あたり 10 個ずつ生成した文字列の長さ 200 までの接頭辞について言語の正解率を求めた. なお, 正解率が高い方が望ましい結果である.

まず, 文字種が異なる場合の結果を示す. その場合の結果が, 図 7.1 である. グラフは横軸が入力の長さ, 縦軸が正解率である. 世界人権宣言, Wikipedia や PPM, MMS を問わず, いずれも数文字ですぐに正解率が 1 に達してしまっていることがわかる. このように文字種が異なる場合には非常に簡単であり, このような組み合わせを評価用データセットに入れていると性能評価が過大評価されてしまう. よって, ここからは基本的に文字種ごとに分けて扱うことにする.

次にラテン文字の場合であり, その結果は図 7.2 の通りである. 左上の 2 本が世界人権宣

[*3] n 交差検定とは, データセットの $(n-1)/n$ を学習に用い, $1/n$ を評価に用いる実験を n 回繰り返す検定方法である.

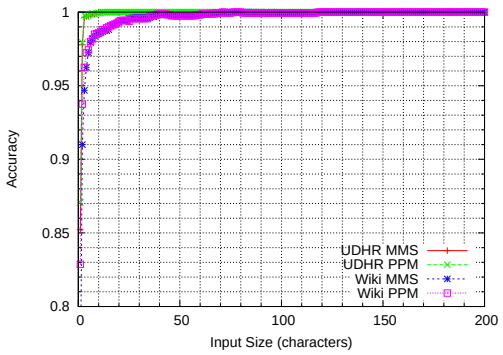

図 7.1. 文字種が異なる場合の言語判別性能

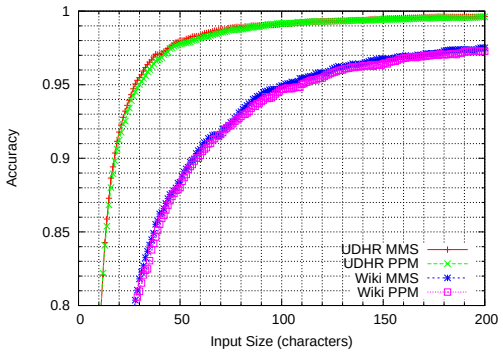

図 7.2. ラテン文字の場合の言語判別性能

言のグラフ, 右下側の 2 本が Wikipedia のグラフである. 言語数が 100 以上多いにもかかわらず, 世界人権宣言に関しての方が性能が良く, 大きな差が付いていることがわかる. 一方で, データ圧縮手法による差がほとんどなく, このタスクはデータ圧縮手法の違いにあまり依存していないということも言える.

最後に他の文字種の場合を述べる. PPM を使用したときの結果が, 図 7.3, 7.4 である. まず, 世界人権宣言を見る. 同じように, 文字種が同じ場合, といっても状況かなり違うことが分かる. おおよそ 4 グループが観察されて, 左上から, 非常によく判別できるヘブライ・ゲエズ・カナダ先住民文字のグループ, キリル文字, ラテン文字とアラビア文字, あまりうまく判別できていないデーヴァナーガリである. ただ, それでも 40 文字あれば 9 割程度判別できているので, まずまずではないかと思う.

また, Wikipedia の場合は, 言語数が少ない文字種が多く, 少し見づらいが, 左上からグルジア・ヘブライ, ギリシア・キリル, 残りとなっている. チベット文字が多少下にはみ出しているのが気になるが, 誤差の範囲ではないかと思う. 9 割のラインが 80 文字程度まで大きくなっているため, 多言語混合文書の分割もその分だけ困難になるのではないかと予想される.

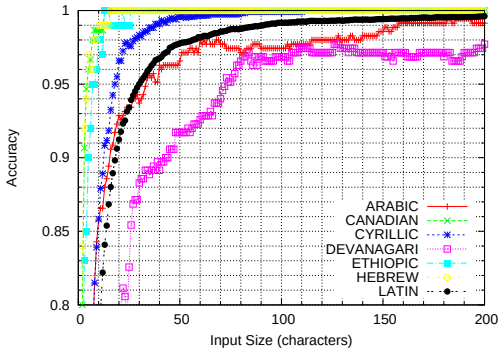

図 7.3. PPM を用いた, 世界人権宣言に関する文字種ごとの言語判別性能

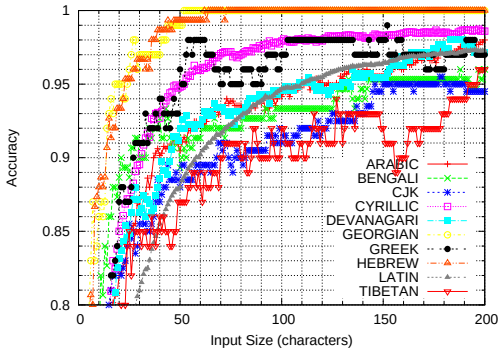

図 7.4. PPM を用いた, Wikipedia に関する文字種ごとの言語判別性能

7.3　予備実験: 境界の検出誤差

次に, もう 1 つ単純化した設定の実験として, 今度は, 多言語文書の言語に関する分割タスクのもう一方, 境界位置の推定精度について調べる. 言語が分かっているという前提の下で, そのために, 推定位置の偏差を求める実験を行った.

実験は, やはり, 世界人権宣言, Wikipedia の各データセットごと, および PPM, MMS の各データ圧縮手法ごとに, 5 分割交差検定を用いて行う. 訓練パート全体を用いて学習した圧縮器を用い, 1 言語あたり 10 個ずつ生成した長さ 100 の文字列を生成して, 長さ 100 の文字列 2 つを空白を挟んで連結させ, 簡略化された計算式,

$$\arg\min_t \{p_{L_1}(x_0 \dots x_{t-1}) + p_{L_2}(x_t \dots x_{|X|-1})\} \tag{7.1}$$

を求める実験を総当たりで, すなわち $100 \times |\mathcal{L}| \times (|\mathcal{L}| - 1)$ 回行い, $t - 100.5$ の累積分布を求めた. なお, グラフの 0 付近での傾きが急であるほど, 推定位置のばらつきが小さいことになり, よい結果である.

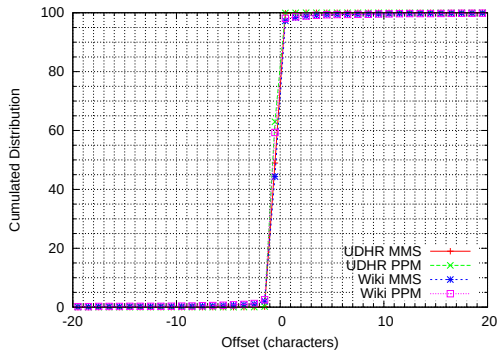

図 7.5. 文字種が異なる場合について境界推定位置の累積分布

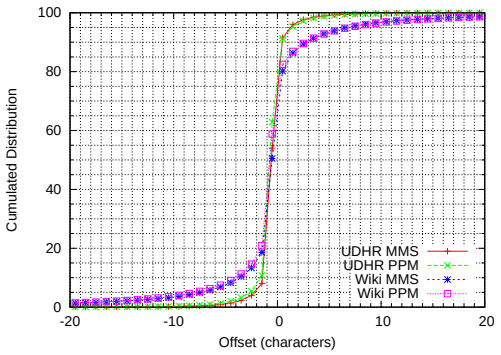

図 7.6. ラテン文字の場合について境界推定位置の累積分布

まず, 文字種が異なる場合の結果であるが, 図 7.5 のようになった. グラフは横軸が分割位置のオフセット, 縦軸が正解率である. 言語判別と同様に非常に簡単な問題であることがわかる.

また, ラテン文字の場合も, 図 7.6 のようになり, やはり言語判別のときと同様のことが言えて, データセットの差に比べてデータ圧縮手法の差が小さいということが言える.

他の文字については, PPM を使用したときの結果が, 図 7.7, 7.8 のようになる. 判別の性能が他と比べて低かった, デーヴァナーガリやチベット文字で傾きが緩やかであることや, 他の文字種はだいたい急であることから, 言語判別の性能との間にそれなりの相関が見られる. しかしながら, 判別性能が高かったはずのゲエズ文字や, ギリシア文字・グルジア文字の傾きが緩やかになるなど, 完全には同じでない. すなわち, 言語判別の性能だけでは, 分割の精度は議論できないということが言える.

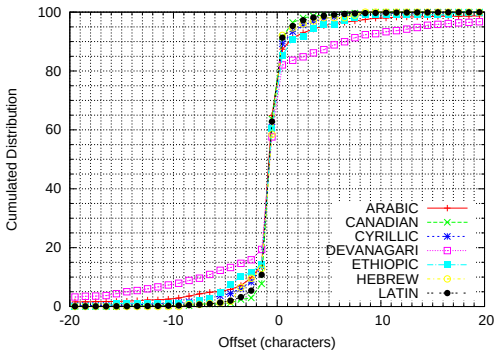

図 7.7. PPM を用いた, 世界人権宣言に関する文字種ごとの境界推定位置の累積分布

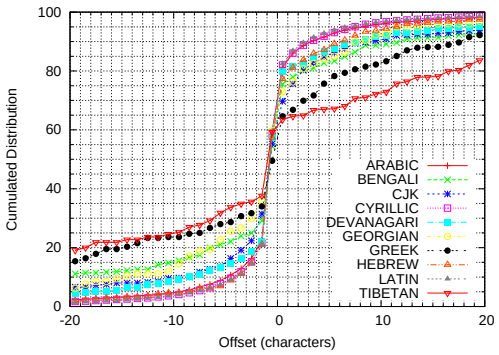

図 7.8. PPM を用いた, Wikipedia に関する文字種ごとの境界推定位置の累積分布

7.4　実験の設定

定量評価のできるような多言語混合文書のデータセットはないため, 人工的に多言語混合文書を作成し, それを用いて評価実験を行う.

なお, あくまで, 人工的に作成した多言語混合文書は, 多言語混合文書を分割するという目的に則った評価をするためだけに用いるものであり, 学習には素のままの単一言語のみからなる文書を使用する, という点を強調しておく. この点は, 実質的に多言語混合文書を必要とする Teahan の方法 [24] とは異なる.

なお, この実験でも 5-分割交差検定を行う.

7.4.1　多言語混合文書の作成法

実験に使う多言語混合文書は次のようにして, 1,000 個ずつ作成した.

1. 言語の数 k を 5–15 からランダムに選び,
2. k 個の言語を選んで並べ, (このとき同じ言語が隣接しないように気をつける)
3. それぞれ長さを $40, 60, 80, \ldots, 160$ からランダムに選び,
4. それぞれの言語のテスト部からランダムにその長さで切り出して,
5. 連結する.

なお, 実際の多言語混合文書は, ほとんどが $k = 2$ であり, 体感的にはべき乗則風の分布をしているように思われる. すなわち, 実態よりも言語数が多いが, これは境界の F 値の収束を早める都合である.

7.4.2　実験のパラメータ

まず, 提案手法では, 使えるデータ圧縮手法に任意性がある. ここでは, ここまで何度も使ってきた PPM と MMS の 2 種類を試すことにする. また, 他にも適当な定数 γ があり, 分割に使用する前にこれを決める必要がある. 実験では, パラメータとして γ の値を $0, 1, \sqrt{2}, 2, \ldots, 256$ を与えることにする. これらを提案手法のパラメータとする.

7.4.3　比較対象

実験で比較対象とする手法は基本的に 3 章で挙げた関連研究を用いることにする. すなわち,

1. 個別に判定, (Individual)
2. Text Tiling,
3. Teahan の方法

の 3 種類である.

まず個別に判定であるが, これは隣の候補点まで, すなわち今回は 1 文字ごとの判定となる. 言語判別の実験結果から見てもほとんど意味のある分割は無理だと思われるが, 分割問題の難易度評価としては使えるので, 追加することにする.

次に Text Tiling であるが, これもブロックが小さいため必然的にスムージングを行うことなる. 今回はそれを固定長とし, 実験のパラメータとして $\{10, 20, 40, 80, 160\}$ から与えることにした. また, 類似度の計算方法も与えなければならないが, 今回は他にあわせて, データ圧縮, すなわち PPM もしくは MMS を用いることにした. なお, 後処理の言語判別でも同じ手法を用いる. 具体的には, 注目点より前の部分で圧縮器を学習し, 後ろの部分の符号長を相違度として用いる. また, しきい値も実験のパラメータとして, $0, 1, \sqrt{2}, 2, \ldots, 256$ を与えることにした.

最後に Teahan の方法である. 多言語混合文書は用意しないことにしているので, 今回はアルファベットサイズを調節することにした. 実験では, データセット中に存在する文字のアルファベットサイズ, ユニコードとしてありえる範囲すべて, [*4] その 2 つの相乗平均, の 3 種類を用いる.

7.4.4　評価方法

今回の実験では,

1. 言語検出の F-値, (および Recall-Precision)
2. 境界検出の F-値, (および Recall-Precision)

[*4] 17×2^{16}

3. 編集距離に基づく正解率, (edit distance accuracy)

の 3 種類の評価基準を用いる.

それぞれの評価基準に言及する前に, まず F-値について説明する. F-値とは情報検索システムの評価で使われる指標である. 一般に情報検索システムには, 検索出力のうち実際に関係があったものの比率 (適合率: Precision) の高さと, 関係のある情報のうち検索出力に含まれていたものの比率 (再現率: Recall) の間にトレードオフの関係があり, これらのバランスをとった値としてそれらの調和平均である F-値がよく使われる.

ただし, これらの評価基準は, 正解と出力が集合であるものに対して適用されるものである. 今回の文書分割では, 例えば出現する言語は, 言語の順序に重要な意味があり, 正解も出力もリストで与えられるため, そのまま適用するには少々問題がある. そこで, F-値の定義を少し変更して, 正解集合と出力集合の積をとるところを, 正解リストと出力リストの最長共通部分列 (Longest Common Subsequence, LCS) をとることにした. 最長共通部分列は動的計画法により簡単に求めることができる量である. なお, リストの要素がユニークでソート済みのときは元の定義と一致する. 以降これを用いる.

これを踏まえて, 言語検出の F 値とは, 先ほど定義した, 含まれている言語のリスト, に関する F 値であり, 境界を多少間違えてもいいので, 言語の誤検出や見逃しをした箇所の少なさを示す指標であり, 発見できた混入の個数を示すものでもある.

また, 境界検出の F 値とは, 見つけられた言語境界の集合に関する普通の F-値である. ただ, 言語境界を 1 文字もずれなしに当てることはそれほど簡単なことではないので, 少しのミスですぐに 0 に近づいてしまう少々過敏な指標でもある. しかしながら, ズレが小さくても 0 でなければ修正の手間がかかるので, これも重要な意味を持つ指標である.

最後に編集距離に基づく正解率は, Teahan [24] が評価に用いた指標で, 正しい言語に分類された文字数の比率を用いたものである. 上の境界検出の F-値と比べると, 多少ずれても境界が検出できればその分を評価できるので, いくらか優しく, フェアな基準に思われる. ただ, 提案手法や Teahan の方法では境界の個数を制限するしくみを持っているため, 真の境界から大きく離れた位置で間違いが発生することを想定する意味はあまりない. 文字列の長さという分母がもっともなものであるかは少々疑問である.

7.5 実験結果

7.5.1 文字種が異なる場合

まずは, 文字種が異なる場合から述べる. このときの P-R プロットは, 言語検出についてが, 図 7.9, 7.10 であり, 境界の位置の検出についてが, 図 7.11, 7.12 であり, 3 つの評価基準それぞれの最高値が, 表 7.2 である.

図や表を見てわかるように提案手法や Teahan の方法で非常に簡単な場合なのであるが, この場合でも個別判定は, 再現率が高いものの適合率が非常に低く, ほとんど使い物になってい

表 7.2. 文字種が異なる場合

	手法	言語検出の F 値	境界検出の F 値	編集距離による正解率
世界人権宣言	提案手法	100.0%	97.4%	100.0%
	Teahan	99.8%	96.0%	100.0%
	TextTiling	98.2%	60.9%	99.0%
	Individual	7.3%	6.5%	86.0%
Wikipedia	提案手法	99.7%	92.8%	99.6%
	Teahan	95.0%	87.5%	98.9%
	TextTiling	96.3%	49.7%	97.1%
	Individual	6.7%	5.9%	82.2%

ない. また, Text Tiling も編集距離は近いものの境界を正確に当てることができておらず, あまり性能が良くないということが言える.

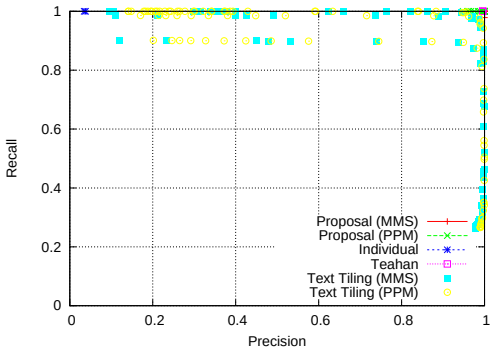

図 7.9. 文字種が異なる場合について, 世界人権宣言に関する言語検出の P-R プロット

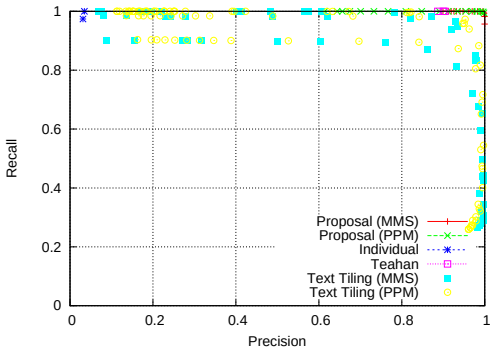

図 7.10. 文字種が異なる場合について, Wikipedia に関する言語検出の P-R プロット

7.5.2　ラテン文字の場合

次に 1 番よく使われる文字種である, ラテン文字のデータセットについての結果であり, このときの P-R プロットは, 言語検出についてが, 図 7.13, 7.14 であり, 境界の位置の検出についてが, 図 7.15, 7.16 であり, 3 つの評価基準それぞれの最高値が, 表 7.3 である.

図や表から読み取れるように, 提案手法は言語検出については非常によい性能を発揮していて, 世界人権宣言についてはほぼ完答, Wikipedia についてもかなり高い正解率である. 他の手法についてみてみると, 個別判定はもはや再現率すら 0 付近であり, Text Tiling も言語検出の F 値すらかなり低くて見劣りする.

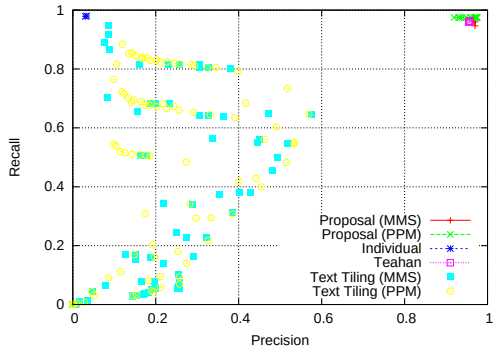

図 7.11. 文字種が異なる場合について, 世界人権宣言に関する分割位置の P-R プロット

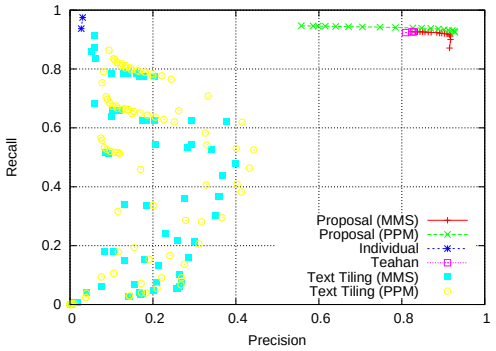

図 7.12. 文字種が異なる場合について, Wikipedia に関する分割位置の P-R プロット

表 7.3. ラテン文字の場合

	手法	言語検出の F 値	境界検出の F 値	編集距離による正解率
世界人権宣言	提案手法	98.8%	75.1%	98.6%
	Teahan	98.0%	66.2%	98.1%
	TextTiling	77.5%	9.5%	83.9%
	Individual	0.5%	1.7%	1.9%
Wikipedia	提案手法	92.3%	53.0%	91.7%
	Teahan	87.8%	44.0%	90.0%
	TextTiling	69.1%	9.1%	72.9%
	Individual	0.8%	1.7%	2.5%

興味深いことに, 2 つの提案手法と Teahan の手法が言語検出の P-R プロット上でほぼ同一の曲線に載っていることである. ここから推測するに, Teahan の手法はあと 1 歩提案手法に及んでいないが, Teahan の手法は損失項の大きさが足りていないことが考えられ, もう少し増やすことで, さら性能が向上して提案手法に肩を並べるということが予想される. しかしながら, 移動量からその大きさを見積もるとアルファベットサイズが `int` の範囲を超えてしまう可能性が高く, 実装上無理が出てくる可能性が出てくることに注意が必要である.

言語検出とは打って変わって, 分割の方を見るとあまり嬉しいとはいえない結果になっている. 文字単位で分割を行うため, 表 7.3 のように, ラテン文字の場合, F 値の最高が, 世界人権宣言ですら 75.1%, Wikipedia にいたっては 53.0% しかなく非常に厳しい結果である.

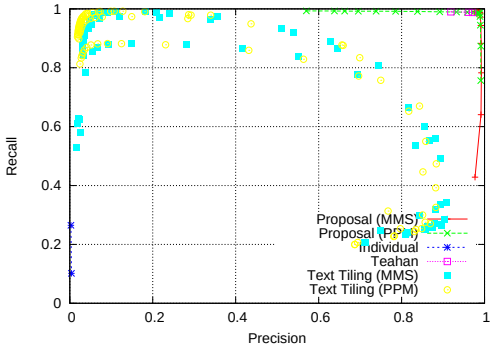

図 7.13. ラテン文字の場合について, 世界人権宣言に関する言語検出の P-R プロット

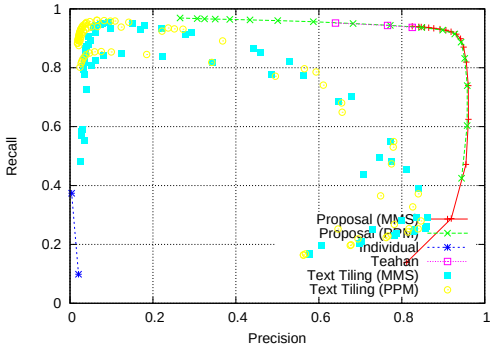

図 7.14. ラテン文字の場合について, Wikipedia に関する言語検出の P-R プロット

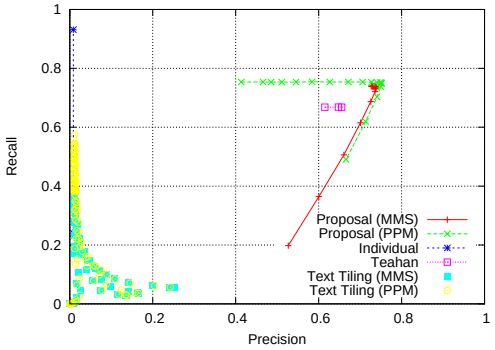

図 7.15. ラテン文字の場合について, 世界人権宣言に関する分割位置の P-R プロット

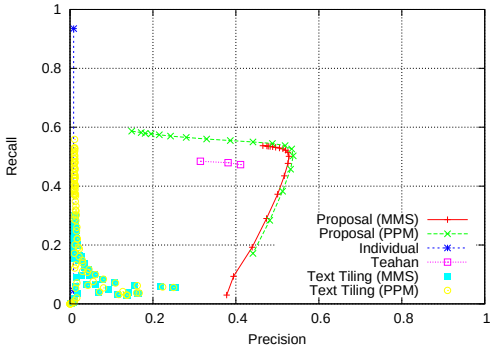

図 7.16. ラテンの場合文字について, Wikipedia に関する分割位置の P-R プロット

表 7.4. ラテン文字の場合 (単語単位)

	手法	言語検出の F 値	境界検出の F 値	編集距離による正解率
世界人権宣言	提案手法	98.9%	94.8%	98.9%
	Teahan	98.3%	92.7%	98.6%
	TextTiling	87.1%	41.3%	89.0%
	Individual	16.1%	14.8%	%
Wikipedia	提案手法	97.0%	79.2%	92.2%
	Teahan	89.5%	74.9%	91.2%
	TextTiling	85.4%	30.4%	76.9%
	Individual	14.4%	13.4%	48.4%

7.5.3 単語単位での分割

そこで, 文字単位で分割はやはり問題設定として厳しすぎるようなので, もう少し他の情報に頼ることにする. 現代では単語の区切りを空白文字で示す分かち書きが一般化していることから, 単語の区切り情報を使った分割実験も行うことにした.

このときの P-R プロットは, 言語検出についてが, 図 7.17, 7.18 であり, 境界の位置の検出についてが, 図 7.19, 7.20 であり, 3 つの評価基準それぞれの最高値が, 表 7.4 である.

すると, 性能はかなり良くなって, 境界検出の F が世界人権宣言で 94.2%, Wikipedia でも 79.2% を記録し, 実用化を考えても良い水準となった.

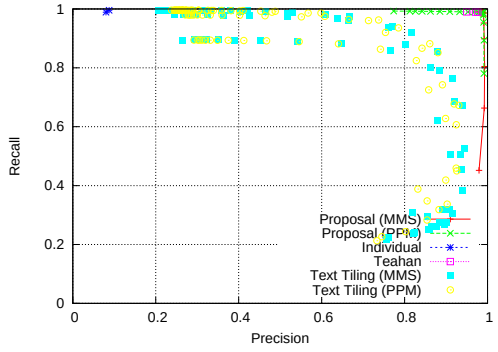

図 7.17. ラテン文字の場合について, 世界人権宣言に関する言語検出の P-R プロット

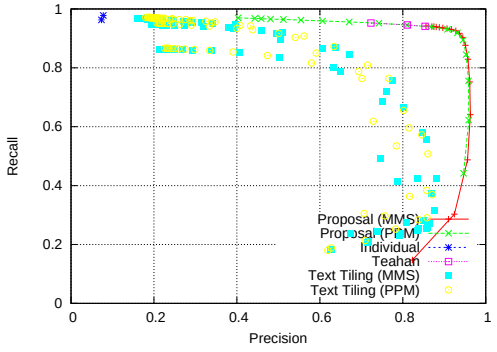

図 7.18. ラテン文字の場合について, Wikipedia に関する言語検出の P-R プロット

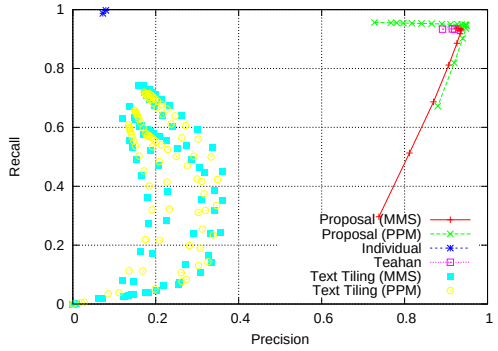

図 7.19. ラテン文字の場合について, 世界人権宣言に関する分割位置の P-R プロット

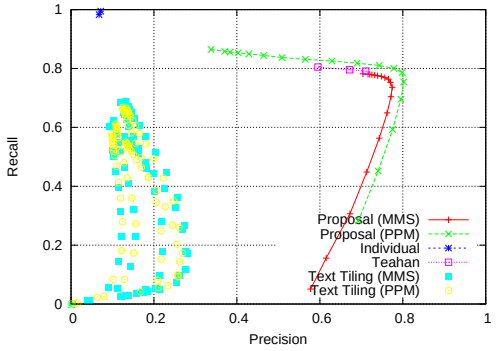

図 7.20. ラテン文字の場合について, Wikipedia に関する分割位置の P-R プロット

7.5.4 他の文字種の場合

さらに他の文字種についても単語単位で分割したときの結果についても述べる.

ただし, その前に気をつけなければないことがあり, それは, いくつかの言語では分かち書きをしないという点である. 本研究で扱う言語の中で該当するものは,

- 日本語 (かな)
- 中国諸語 (漢字)
- タイ語 (タイ文字)
- ラオ語 (ラオ文字)
- クメール語 (クメール文字)
- イ語 (ロロ文字)
- チベット語・ゾンカ語 (チベット文字)

であり, それらの文字に関しては文字単位にするという処置を行っている.

このとき, PPM を使った提案手法による, 分割位置に関する P-R プロットは, 言語検出についてが, 図 7.21, 7.22 である.

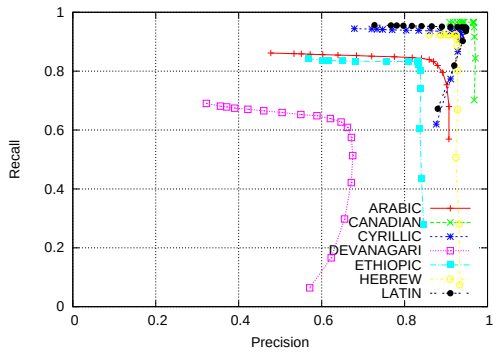

図 7.21. PPM を用いた, 世界人権宣言に関する, 単語単位での分割位置の文字種ごとの P-R プロット

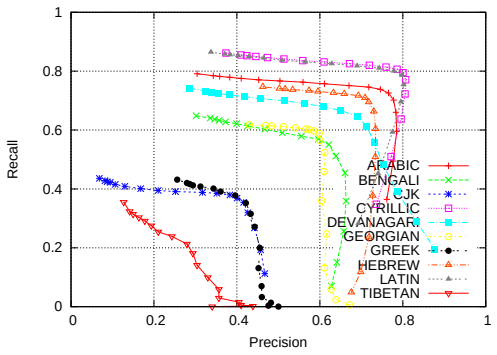

図 7.22. PPM を用いた, Wikipedia に関する, 単語単位での分割位置の文字種ごとの P-R プロット

グラフのように, 多くの文字種ではラテン文字同様, 良い結果を得られているといえ, 分割位置の推定誤差のない比率がとても高く, 非常に満足のいく結果である. しかしながら, デーヴァナーガリ, 漢字, ギリシア文字, そしてチベット文字などでは, 他の手法もやはり十分な性能がないものの, あまり満足できる結果とはなっていない. どのような文字種でも, 高い性能を発揮するように改善することが今後の課題であるといえる.

7.6　分割にかかる時間

実際, 実験に用いたプログラムがどの程度の速度動いているかを示す. 計測は Xeon5650 2.66-GHz CPU を積んだマシンで, 50 回計測しその平均をとった.

まず, 提案手法である. 世界人権宣言 (361 言語) と Wikipedia (253 言語) で実行したときのグラフが, 図 7.23 である. 上の 2 つが PPM によるものであり, 下 2 つが MMS によるも

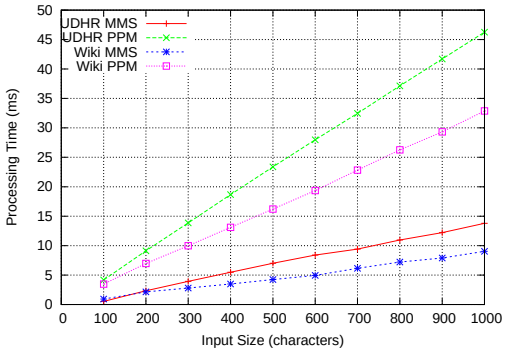

図 7.23. 文書の平均処理時間

表 7.5. 他の手法の速度

	Individual		Teahan	Text Tiling (40 文字)	
	MMS	PPM		MMS	PPM
UDHR	8.958	42.136	236.562	4.888	12.566
Wikipedia	6.004	29.988	161.186	3.418	8.072

のである. 理論上, 線形時間で動作すると述べたが, 実際グラフもそうなっていることが読み取れる. また, 1000 文字を処理する時間は MMS で 10ms 程度, PPM で 40 ms 程度となっていて, かなり高速に処理できるといえる. なお, 速度は言語数にも依存するため, 例えば 世界人権宣言のアラビア文字セット (7 言語) であれば, 0.3ms 程度となり, より高速にすることが可能である.

ちなみに, 今の実装で圧縮器を保持するのに必要なメモリは, 世界人権宣言のデータセットで 450MB 程度, Wikipedia のデータセットで 370MB 程度である.

参考までに, 性能の比較対象としている他の手法についても見てみよう. この実験で使った実装では, 表 7.5 の通りになった. 単位はいずれも ms である. おおよそどれも 10ms 前後となっていて, 提案手法が遅いということはない. Teahan が非常に遅いが, これは, 単に Teahan の手法における分割の損失を計算する部分の高速化を全くおこなっていないため, 理論的には PPM を使った提案手法なみになる見込みである.

第 8 章

実データを用いた実験

この章では, 実データを用いて行った実験結果を紹介する. 7 章で述べたように, 多言語混合文書の大規模データセットはないため, これらは自力かつ手作業で集め加工したものである.

8.1 データセット

実データとして集めたテストケースは, 表 Wikipedia から集めてきた, 8.1 に挙げた 20 個である. 表は左から, ケース番号, 文字種, どの言語版の Wikipedia の記事の一部か, 実際に使われている言語, 文字数, 2 章に挙げた多言語混合文書の分類における種別, 言語境界が段落内か段落外かである. また, 元 URL は付録 B に掲載してある. なお, いずれも Wikipedia のデータセットには使用されていない記事である. なお, できるだけ多様な事例を揃えるように気をつけたが, かなりの偏りがあるかもしれないという点には注意してほしい.

なお, 正解の区切りは, 今回の実験で使用した分割ツールや, いくつかの翻訳サイトの翻訳精度などから総合的に判断した.

これらについて, 7 章における実験と同様の設定で, 単語単位, 学習データには Wikipedia のセット (文字種を限定せず, すべて) を用いて分割を行った.

8.2 実験結果

総合での結果は表 8.2 のようになった. なお, それぞれ最も結果の良かったパラメータに関してのみ掲載してある. また P-R プロットは, 図 8.1, 8.2 のようになった.

結果の良い順から, 提案手法, Teahan の方法, Text Tiling となっていて, MMS を用いた提案手法が, $\gamma = 22.6$ と設定したときに, 言語検出の F 値において 90.74%, 境界検出の F 値において 50.0%, 編集距離に基づく正解率で 95.9% を記録している. 境界検出において, 50% しかないのは残念であるが, それでも他と比べるとかなり良い数字であることと言える.

個別のテストケースでの結果は, 表 8.3 の通りである. 今度は編集距離に基づく編集距離に基づく正解率だけを示している. また, P-R プロットは図 8.1, 8.2 の通りである.

表 8.1. 実データセットのリスト

番号	文字種	Wikiepdia の言語タグ	実際の言語	サイズ	種類	境界
1	ラテン	ルクセンブルク語	＋フランス語	30	引用	段落内
2	ラテン	マレー語	＋英語	795	不完全な翻訳	段落内
3	ラテン	タガログ語	＋英語	1145	不完全な翻訳	段落間
4	キリル	アブハズ語	＋ロシア語	248	不完全な翻訳	段落内
5	キリル	ヤクート語	＋ロシア語	1291	引用	段落内
6	キリル	チュヴァシ語	＋ロシア語	859	引用	段落内
7	キリル	タジク語	＋ロシア語	170	不完全な翻訳	段落内
8	ラテン	ヨルバ語	＋英語	427	不完全な翻訳	段落内
9	ラテン	ヨルバ語	＋英語	3741	不完全な翻訳	段落間
10	ラテン	トルコ語	＋英語	257	不完全な翻訳	段落内
11	ラテン	セブアノ語	＋英語	888	引用	段落内
12	ラテン	セブアノ語	＋英語	728	引用	段落内
13	ラテン	ハイチ語	英語のみ	11465	他言語話者に対する案内	-
14	ラテン	スンダ語	＋英語	423	不完全な翻訳	段落内
15	ラテン	スンダ語	＋英語	543	不完全な翻訳	段落内
16	ラテン	パンパンガ語	＋英語	882	不完全な翻訳	段落間
17	ラテン	パンパンガ語	＋英語	245	不完全な翻訳	段落間
18	グルジア	メグレル語	＋グルジア語	608	引用	段落内
19	ヘブライ	イディッシュ語	＋ヘブライ語	703	不完全な翻訳	段落間
20	ヘブライ	イディッシュ語	＋ヘブライ語	1426	不完全な翻訳	段落間

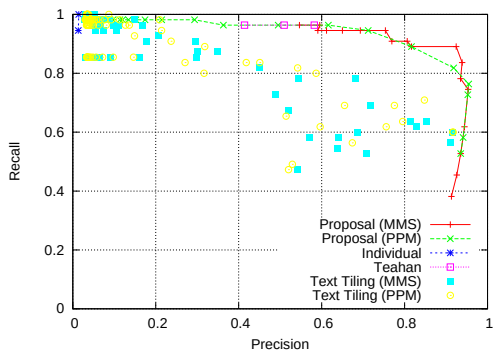

図 8.1. 言語の P-R プロット

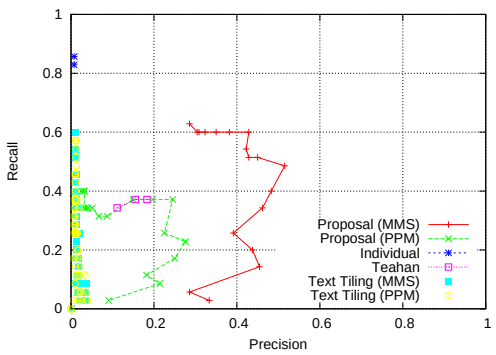

図 8.2. 分割位置の P-R プロット

表 8.2. 実データセットでの結果

手法	パラメータ	言語検出の F 値	境界検出の F 値	編集距離による正解率
提案手法	MMS, 22.6	90.7%	50.0%	95.9%
提案手法	PPM, 64.0	86.5%	25.0%	94.8%
Teahan	$17 \times 2^{16} + 1$	72.6%	24.5%	90.5%
TextTiling	PPM, 160, 128.0	77.2%	3.3%	88.9%
TextTiling	MMS, 160, 32.0	72.9%	2.6%	87.6%
Individual	PPM	2.9%	1.5%	28.6%

10, 15 など一部のケースでは大きく失敗してしまっているものの, 提案手法は大体安定して高い性能を出している. それに対して Text Tiling は全体的に性能が低めである. Teahan の手法は全体として, 提案手法と良い勝負である. 短くて難しい 1 で高い性能を発揮している点は興味深いが, その一方で 5 や 12 のような比較的簡単なケースで Text Tiling に劣るような性能しか出せない場合もあり, 少々汎用性に難があるようである.

個数が少ないため, あまり断定的なことは言えないが, 実データにおいても, 幾つかの場合を除いて提案手法が概ね最も優れており, しかも, 絶対的に見ても, 幾つかを除いてそれなりに良い性能を発揮することが確かめられたといえる.

表 8.3. 実データセットでの個別の結果

番号	提案手法 (MMS, 22.6)	提案手法 (PPM, 64.)	Teahan ($17 \times 2^{16} + 1$)	Text Tiling (PPM, 160, 128.)	Text Tiling (MMS, 160, 32.)
1	82.2%	82.2%	100.0%	71.4%	71.4%
2	99.1%	100.0%	100.0%	96.1%	76.0%
3	100.0%	98.3%	98.3%	88.7%	79.1%
4	100.0%	100.0%	100.0%	43.5%	43.5%
5	100.0%	100.0%	95.2%	97.1%	79.9%
6	97.7%	97.7%	92.7%	93.5%	77.1%
7	99.4%	99.4%	99.4%	74.1%	74.1%
8	100.0%	100.0%	100.0%	82.2%	76.8%
9	94.8%	85.3%	90.0%	78.7%	61.7%
10	0.0%	60.7%	31.9%	0.0%	0.0%
11	95.6%	95.6%	95.6%	62.0%	90.1%
12	99.6%	89.1%	78.6%	70.5%	81.3%
13	100.0%	100.0%	89.9%	95.5%	59.9%
14	96.9%	99.1%	100.0%	96.9%	86.5%
15	33.9%	33.9%	33.9%	48.8%	48.1%
16	100.0%	98.9%	98.4%	94.7%	80.4%
17	100.0%	100.0%	100.0%	49.8%	49.8%
18	92.1%	92.1%	92.1%	95.4%	97.7%
19	87.1%	87.9%	87.9%	94.0%	88.2%
20	98.5%	98.2%	96.8%	93.1%	95.9%

第 9 章

結論

9.1 まとめ

本研究では多言語文書の存在を示し, その分割手法について扱った. 論文中で示したようにWikipedia には大量の多言語混合文書が存在していることから, この研究により, 少数言語のコーパスづくりをはじめ, 様々な自然言語処理のタスクを遂行する上でのでの手助けになるのではないかと思う.

また, 分割問題を, 最小記述長原理を参考にしてデータ圧縮を利用した方法により, 最適な分割と言語の組みを求める最適化問題として定式化すると同時に, 2 種類のデータ圧縮手法それぞれについて, それを用いた最適化問題を解く線形時間アルゴリズムを提案した.

人工データによる実験結果も示した. デーヴァナーガリ・漢字・ギリシア文字・チベット文字などいくつかの文字種ではうまく行かなかったが, ラテン文字・キリル文字といった主要な文字種で, 各言語 10kB 程度と, 少量の限られた学習データにもかかわらず, 高い性能を発揮することを確かめることができた. 40–160 文字の文書片を 5–15 個組み合わせて作られた多言語文書でを用いて行った実験では, 世界人権宣言のラテン文字について言語の F 値の最高で 98.9%, 境界の F 値の最高で 94.8%, 編集距離による精度で 98.9% を達成している.

速度面でも優れており, 言語数 300 で, 秒間約 100,000 文字を処理できる.

数が非常に少ないものの, 実データでもそれなりに良い性能を発揮することを確認できた.

9.2 今後の課題

各章で触れたように, γ の理論付け, Juola の予想の証明など, 実用上は問題ないものの未解決の課題となっているものがいくつかあり, これらの解決を目指したいと考えている.

また, いくつかの文字種では十分な性能を発揮できておらず, その原因を究明したい.

この研究は, 前処理としての利用が想定されているため, 提案手法を使用することで, 自然言語処理のタスクの精度を向上させるかどうか, ぜひ確かめたいと考えている.

また, 提案手法で用いた定式化は, 言語による分割だけでなく, 他の文書分割タスクにも応用

出来るため, それらへの応用も考えたい.

この研究により, 多言語混合文書処理が一般的になる未来を願っている.

発表文献と研究活動

(1) 山口洋. “圧縮に基づく多言語文書の言語に関する分割に関する研究.” 東京大学音声・言語・コミュニケーション研究会, 2011.11.07.

(2) Hiroshi Yamaguchi and Kumiko Tanaka-Ishii. 2012. “Text Segmentation by Language Using Minimum Description Length.” In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, pages 969–978.

参考文献

[1] Mohamed Ibrahim Abouelhoda, Stefan Kurtz, and Enno Ohlebusch. Replacing suffix trees with enhanced suffix arrays. *Journal of Discrete Algorithms*, Vol. 2, No. 1, pp. 53–86, 2004.

[2] Alok Aggarwal, MariaM. Klawe, Shlomo Moran, Peter Shor, and Robert Wilber. Geometric applications of a matrix-searching algorithm. *Algorithmica*, Vol. 2, pp. 195–208, 1987.

[3] Beatrice Alex. An unsupervised system for identifying english inclusions in german text. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics, Student Research Workshop*, pp. 133–138, 2005.

[4] Beatrice Alex, Amit Dubey, and Frank Keller. Using foreign inclusion detection to improve parsing performance. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pp. 151–160, 2007.

[5] T.C. Bell, J.G. Cleary, and I. H. Witten. *Text Compression*. Prentice Hall, 1990.

[6] Dario Benedetto, Emanuele Caglioti, and Vittorio Loreto. Language trees and zipping. *Physical Review Letters*, Vol. 88, No. 4, 2002.

[7] Rudi Cilibrasi and Paul Vitányi. Clustering by compression. *IEEE Transactions on Information Theory*, Vol. 51, No. 4, pp. 1523–1545, 2005.

[8] J.G. Cleary, W.J. Teahan, and I.H. Witten. Unbounded length contexts for ppm. In *Data Compression Conference, 1995. DCC '95. Proceedings*, pp. 52–61, mar 1995.

[9] John G. Cleary and Ian H. Witten. Data compression using adaptive coding and partial string matching. *IEEE Transactions on Communications*, Vol. 32, pp. 396–402, 1984.

[10] Yo Ehara and Kumiko Tanaka-Ishii. Multilingual text entry using automatic language detection. In *Proceedings of the third International Joint Conference on Natural Language Processing*, pp. 441–448, 2008.

[11] Martin Farach, Michiel Noordewier, Serap Savari, Larry Shepp, Abraham J. Wyner, and Jacob Ziv. On the entropy of dna: Algorithms and measurements based on

memory and rapid convergence. In *Proceedings of the Sixth Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 48–57, 1994.

[12] Gregory Grefenstette. Comparing two language identification schemes. In *Proceedings of 3rd International Conference on Statistical Analysis of Textual Data*, pp. 263–268, 1995.

[13] Dan Gusfield. *Algorithms on Strings, Trees and Sequences: Computer Science and Computational Biology*. Cambridge Univ. Press, May 1997.

[14] Marti A. Hearst. Texttiling: Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, Vol. 23, No. 1, pp. 33–64, 1997.

[15] Patrick Juola. What can we do with small corpora? document categorization via cross-entropy. In *Proceedings of an Interdisciplinary Workshop on Similarity and Categorization*, 1997.

[16] Gen-itiro Kikui. Identifying the coding system and language of on-line documents on the internet. In *Proceedings of 16th International Conference on Computational Linguistics*, pp. 652–657, 1996.

[17] D.E. Knuth. Optimum binary search trees. *Acta Informatica*, Vol. 1, pp. 14–25, 1971.

[18] Casanai Kruengkrai, Prapass Srichaivattana, Virach Sornlertlamvanich, and Hitoshi Isahara. Language identification based on string kernels. In *Proceedings of the 5th International Symposium on Communications and Information Technologies*, pp. 926–929, 2005.

[19] Alistair Moffat. Implementing the ppm data compression scheme. *IEEE Transactions on Communications*, Vol. 38, No. 1, pp. 1917–1921, 1990.

[20] J. Rissanen. Modeling by shortest data description. *Automatica*, Vol. 14, No. 5, pp. 465 – 471, 1978.

[21] Nathan Sanders. Kde 4's sonnet will turbocharge language processing. `http://archive09.linux.com/articles/59963`.

[22] Nathan Sanders. Kde 4 の sonnet による高度な言語処理機能. `http://sourceforge.jp/magazine/07/02/13/0015223`.

[23] William J. Teahan and David J. Harper. Using compression-based language models for text categorization. In *Proceedings of the Workshop on Language Modeling and Information Retrieval*, pp. 83–88, 2001.

[24] William John Teahan. Text classification and segmentation using minimum cross-entropy. In *RIAO*, pp. 943–961, 2000.

[25] Esko Ukkonen. On-line construction of suffix trees. *Algorithmica*, Vol. 14, No. 3, pp. 249–260, 1995.

[26] F. Frances Yao. Efficient dynamic programming using quadrangle inequalities. In

Proceedings of the twelfth annual ACM symposium on Theory of computing, STOC '80, pp. 429–435, New York, NY, USA, 1980. ACM.

[27] Jacob Ziv and Abraham Lempel. A universal algorithm for sequential data compression. *IEEE Transactions on Information Theory*, Vol. 23, No. 3, pp. 337–343, 1977.

謝辞

真の指導教員である田中 (石井) 久美子先生には, 大変お世話になりました. この場を借りてお礼申し上げます.

また, 形式上の指導教員である千葉滋先生, および千葉研究室の皆様にもお世話になりました. さらに, 旧田中研の皆様, 早矢仕さんにもお世話になりました.

最後に, よい多言語データセットの素を提供して下さった, Wikipedia の編集者の皆様と UDHR の翻訳者の皆様, UDHR in Unicode の皆様にも感謝いたします.

ありがとうございました.

付録 A

全言語リスト

本研究で扱った全言語の日本語名と言語コードの一覧を載せる．なお, 併記する言語コードは, 世界人権宣言については *UDHR in Unicode*[*1]で採用されているもの, Wikipedia については Wikipedia 公式の言語コードであり, 概ね ISO 639 に従うが, 若干の差異がある．また, 異なる文字で書かれたものは別言語としてカウントしているのでいくらか重複がある．

A.1 世界人権宣言

A.1.1 アラビア文字

1. 標準アラビア語 [arb]
2. マレー語 [mly_arab]
3. 西ペルシア語 [pes_1]
4. 西パンジャーブ語 [pnb]
5. サライキ語 [skr]
6. ウイグル語 [uig_arab]
7. ウルドゥー語 [urd]

A.1.2 カナダ先住民文字

1. 湿原クリー語 [csw]
2. 東カナダ・イヌクティトゥット語 [ike]
3. 北西オジブウェー語 [ojb]

A.1.3 キリル文字

1. アブハズ語 [abk]
2. 南アルタイ語 [alt]
3. 北アゼルバイジャン語 [azj_cyrl]
4. ベラルーシ語 [bel]

[*1] http://unicode.org/udhr/index.html

5. ブルガリア語 [bul]
6. ショル語 [cjs]
7. エヴェン語 [eve]
8. エヴェンキ語 [evn]
9. カザフ語 [kzh]
10. ハルハ・ モンゴル語 [khk]
11. キルギス語 [kir]
12. カフカス語 [kjh]
13. マケドニア語 [mkd]
14. オセット語 [oss]
15. ロシア語 [rus]
16. ヤクート語 [sah]
17. タタール語 [tat]
18. タジク語 [tgk]
19. トゥルクメン語 [tuk_cyrl]
20. トゥバ語 [tyv]
21. ウクライナ語 [ukr]
22. 北ウズベク語 [uzn_cyrl]
23. 北ユカギール語 [ykg]
24. セルビア語グループ [srp_cyrl]

セルビア語グループ

1. ボスニア語 [bos_cyrl]
2. セルビア語 [srp_cyrl]

A.1.4　デーヴァナーガリ

1. ボージュプリー語 [bho]
2. ヒンディー語 [hin]
3. マガヒー語 [mag]
4. マイティリー語 [mai]
5. マラーティー語 [mar]
6. ネパール語 [nep]
7. サンスクリット語 [san]

A.1.5　ゲエズ文字

1. アムハラ語 [amh]
2. ティグリニャ語 [tir]

A.1.6　ヘブライ文字

1. ヘブライ語 [heb]
2. 東イディッシュ語 [ydd]

A.1.7　ラテン文字

1. サントメプリンシペ・ クレオール語 [007]
2. ギニアビサウ・ クレオール語 [008]
3. キンブンド語 [009]
4. ムブンドゥ語 [010]
5. テトゥン語 [011]
6. アチェ語 [ace]
7. アチュアラ・ ヒバロ語 [acu]
8. ダンメ語 [ada]
9. アフリカーンス語 [afr]
10. アグアルナ語 [agr]
11. アジャ語 [ajg]
12. アクアペン・ アカン語 [aka_akuapem]
13. アサンテ・ アカン語 [aka_asante]
14. ファンテ・ アカン語 [aka_fante]
15. トスク・ アルバニア語 [als]
16. アマワカ語 [amc]
17. ヤネシャ語 [ame]
18. アマラカエリ語 [amr]
19. アラベラ語 [arl]
20. マプーチェ語 [arn]
21. アストゥリアス語 [ast]
22. ワオラニ語 [auc]
23. オーベルニュ・ オック語 [auv]
24. 中部アユマラ語 [ayr]
25. 北アゼルバイジャン語 [azj_latn]
26. バンバラ語 [bam]
27. バリ語 (Bali) [ban]
28. バリバ語 [bba]
29. バオレ語 [bci]
30. 中部ビコール語 [bcl]
31. ベンバ語 [bem]
32. バリ語 (Bari) [bfa]
33. エド語 [bin]
34. ビスマラ語 [bis]
35. モン・ ジュア語 [blu]
36. ボラ語 [bora]
37. ベティ語 [btb]
38. ブレトン語 [bre]
39. ブギス語 [bug]
40. ガリフナ語 [cab]
41. 中部カクチケル語 [cak]
42. カタルーニャ語 (バレンシア語・ バレアレス諸島語) [cat]
43. カヤパス語 [cbi]
44. カシボ語 [cbr]
45. チャヤウィタ語 [cbt]
46. カンドシ語 [cbu]
47. 北チワン語 [ccx]
48. セブ語 [ceb]
49. チェコ語 [ces]
50. チャモロ語 [cha]
51. オヒトラン・ チナンテカ語 [chj]
52. チューク語 [chk]
53. チカソー語 [cic]
54. チョクウェ語 [cjk]
55. アンゴラ・ チョクウェ語 [cjk_AO]
56. 中部クルド語 [ckb]
57. ハカ・ チン語 [cnh]
58. アシャニンカ語 [cni]
59. コロラド語 [cof]
60. コルシカ語 [cos]
61. カキンテ語 [cot]
62. ピチス・ アシェニンカ語 [cpu]
63. セーシェル・ クレオール語 [crs]
64. チルテペク・ チナンテカ語 [csa]
65. テディム・ チン語 [ctd]
66. ウェールズ語 [cym]
67. ダバニ語 [dag]
68. デンマーク語 [dan]
69. デンディ語 [ddn]

70. 標準ドイツ語 [deu_1996]
71. 南ダガリ語 [dga]
72. 北西ディンカ語 [dip]
73. フォニィ・ ジョラ語 [dyo]
74. 東マニンカ語 [emk]
75. エミリア・ ロマーニャ語 [eml]
76. 英語 [eng]
77. エスペラント [epo]
78. エストニア語 [est]
79. バスク語 [eus]
80. エウェ語 [ewe]
81. フェロー語 [fao]
82. フィジー語 [fij]
83. フィン語 [fin]
84. ファラム・ チン語 [flm]
85. フォン語 [fon]
86. フランス語 [fra],
87. 西フリジア語 [fri]
88. プラール・ フラニ語 [fuc]
89. フリウリ語 [fur]
90. ガ語 [gaa]
91. ガガウズ語 [gag]
92. ボラナ・ アルシ・ グジ・ オロモ語 [gax]
93. ゴンジャ語 [gjn]
94. ギニア・ クペレ語 [gkp]
95. スコットランド・ ゲール語 [gla]
96. アイルランド・ ゲール語 [gle]
97. ガリシア語 [glg]
98. ワユー語 [guc]
99. パラグアイ・ グアラニー語 [gug]
100. ヤノマメ・ ヤノマミ語 [guu]
101. グアラヨ語 [gyr]
102. ハイチ・ クレオール語 [hat_kreyol]
103. ハイチ・ クレオール語 (口語) [hat_popular]
104. ニジェール・ ハウサ語 [hau_NE]
105. ハワイ語 [haw]
106. 北部黔東ミャオ語 [hea]
107. ヒリガイノン語 [hil]
108. マツ・ チン語 [hlt]
109. 南部黔東ミャオ語 [hms]
110. ミナ語 (ヒナ語, カメルーン) [hna]
111. ハニ語 [hni]
112. 上ソルブ語 [hsb]
113. 南東ワステコ語 [hsf]
114. ハンガリー語 [hun]
115. ベラクルス・ ワステコ語 [hus]
116. ムルイ・ ウィトト語 [huu]
117. サン・ ルイ・ ポトシ・ ワステコ語 [hva]
118. イビビオ語 [ibb]
119. イボ語 [ibo]
120. イド語 [ido]
121. イロカノ語 [ilo]
122. インターリンガ [ina]
123. インドネシア語 [ind]
124. アイスランド語 [isl]
125. イタリア語 [ita]
126. ジャワ語 [jav]
127. シュアール・ ヒバロ語 [jiv]
128. グリーンランド・ イヌクティトゥット語 [kal]
129. カビイェ語 [kbp]
130. マコンデ語 [kde]
131. カーボベルデ・ クレオール語 [kea]
132. ケクチ語 [kek]
133. カシ語 [kha]
134. ルワンダ語 [kin]
135. 中部カヌリ語 [knc]
136. コンゴ語グループ [kng]
137. アンゴラ・ コンゴ語 [kng_AO]

138. コンゾ語 (ウガンダ) [koo]
139. カオンデ語 [kqn]
140. クリオ語 [kri]
141. カレリア語 [krl]
142. アワ語 (クァイケル語) [kwi]
143. ラディーノ語 [lad]
144. ラテン語 [lat]
145. ラトビア語 [lav]
146. 中西部リンバ語 (シエラレオネ) [lia]
147. リンガラ語 [lin]
148. リトアニア語 [lit]
149. ラングドック・ オック語 [lnc]
150. ラムンソ語 [lns]
151. ロトゥコ語 [lot]
152. ロジ語 [loz]
153. ルクセンブルク語 [ltz]
154. ルバ語 [lua]
155. ルバレ語 [lue]
156. ガンダ語 [lug]
157. ルンダ語 [lun]
158. ミゾ語 [lus]
159. マドゥラ語 [mad]
160. マーシャル語 [mah]
161. 北マム語 [mam]
162. 中央マサワ語 [maz]
163. シャラナワ語 [mcd]
164. マツェス語 [mcf]
165. メンデ語 (シエラレオネ) [men]
166. ミクマク語 [mic]
167. ミナンカバウ語 [min]
168. ミスキート語 [miq]
169. マルタ語 [mlt]
170. マレー語 [mly_latn]
171. モシ語 [mos]
172. マオリ語 [mri]
173. モザラベ語 [mxi]
174. メトラトノク・ ミシュテカ語 [mxv]
175. イシュカトラン・ マサテク語 [mzi]
176. ナバホ語 [nav]
177. ニェンバ語 [nba]
178. ンデベレ語 [nbl]
179. ンドンガ・ オバンボ語 [ndo]
180. 低地サクソン語 [nds]
181. 中部ナワトル語 [nhn]
182. アオ・ ナガ語 [njo]
183. オランダ語 [nld]
184. ニーノシェク・ ノルウェー語 [nno]
185. ブークモール・ ノルウェー語 [nob]
186. ノマツィゲンガ語 [not]
187. 北ソト語 [nso]
188. チェワ・ ニャンジャ語 [nya_chechewa]
189. ニャンジャ・ ニャンジャ語 [nya_chinyanja]
190. ニャムウェジ語 [nym]
191. ニャンコレ語 [nyn]
192. ンゼマ語 [nzi]
193. メスキタル・ オトミ語 [ote]
194. パンパンガ語 [pam]
195. パラオ語 [pau]
196. パエズ語 [pbb]
197. ピカルディ語 [pcd]
198. ナイジェリア・ ピジン語 [pcm]
199. ピジン語 (ソロモン諸島) [pis]
200. 高原マダガスカル語 [plt]
201. ポーランド語 [pol]
202. ポンペイ語 [pon]
203. ブラジル・ ポルトガル語 [por_BR]
204. イベリア・ ポルトガル語 [por_PT]
205. ギニアビサウ・ クレオール語 [pov]
206. ピピル語 [ppl]

207. プロヴァンス・ オック語 [prv]
208. 中部キチェ語 [quc]
209. カルデロン高原ケチュア語 [qud]
210. チンボラソ高原ケチュア語 [qug]
211. アヤクチョ・ ケチュア語 [quy]
212. クスコ・ ケチュア語 [quz]
213. アンボ・ パスコ・ ケチュア語 [qva]
214. カハマルカ・ ケチュア語 [qvc]
215. ワマリエス・ ドス・ デ・ マヨ・ ワヌコ・ ケチュア語 [qvh]
216. マルゴス・ ヤロウィルカ・ ラウリコチャ・ ケチュア語 [qvm]
217. 北フニン・ ケチュア語 [qvn]
218. ウアイラス・ アンカシュ・ ケチュア語 [qwh]
219. チクィアン・ アンカシュ・ ケチュア語 [qxa]
220. 北コンチュコス・アンカシュ・ケチュア語 [qxn]
221. ラ・ ウニオン・ アレキパ・ ケチュア語 [qxu]
222. ラロトンガ・ クック諸島マオリ語 [rar]
223. バルカン・ ロマ語 [rmn]
224. ヴラックス・ ロマ語 [rmy]
225. ロマンシュ語 [roh]
226. ルーマニア語 [ron_2006]
227. ルンディ語 [run]
228. サンゴ語 [sag]
229. スコットランド語 [sco]
230. セコヤ語 [sey]
231. シルック語 [shk]
232. シピボ・ コニボ語 [shp]
233. スロバキア語 [slk]
234. スロベニア語 [slv]
235. 北サーミ語 [sme]
236. サモア語 [smo]
237. ショナ語 [sna]
238. ソニンケ語 [snk]
239. シオナ語 [snn]
240. ソマリ語 [som]
241. 南ソト語 [sot]
242. スペイン語 [spa]
243. ログドーロ・ サルデーニャ語 [src]
244. セルビア語グループ [srp_latn]
245. セレール語 [srr]
246. スワジ語 [ssw]
247. スクマ語 [suk]
248. スンダ語 [sun]
249. スス語 [sus]
250. スウェーデン語 [swe]
251. スワヒリ語 [swh]
252. タヒチ語 [tah]
253. ディタマリ語 [tbz]
254. ティクナ語 [tca]
255. テムネ語 [tem]
256. テトゥン語 [tet]
257. タガログ語 [tgl]
258. ティヴ語 [tiv]
259. トバ語 [tob]
260. トンガ語 (バントゥー語群) [toi]
261. トホラバル語 [toj]
262. トンガ語 (オーストロネシア語族) [ton]
263. パパントラ・ トトナカ語 [top]
264. トク・ ピシン [tpi]
265. ツワナ語 [tsn]
266. モザンビーク・ ツォンガ語 [tso_MZ]
267. プレペチャ語 (タラスコ語) [tsz]
268. トゥルクメン語 [tuk_latn]
269. トルコ語 [tur]
270. チャムラ・ ツォチル語 [tzc]

271. オスチュク・ツェルタル語 [tzh]
272. 中央アトラス・タマジクト語 [tzm]
273. ウイグル語 [uig_latn]
274. ウラリナ語 [ura]
275. 北ウズベク語 [uzn_latn]
276. ヴェネチア語 [vec]
277. ヴェンダ語 [ven]
278. ヴェプス語 [vep]
279. ベトナム語 [vie]
280. マクワ語 [vmw]
281. ワライ語 [war]
282. ワロン語 [wln]
283. ウォロフ語 [wol]
284. ワーマ語 [wwa]
285. コサ語 [xho]
286. カセン語 [xsm]
287. ヤグア語 [yad]
288. ヤオ語 (バントゥー語群) [yao]
289. ヤップ語 [yap]
290. ヨルバ語 [yor]
291. ユカタン・マヤ語 [yua]
292. ミアワトラン・サポテカ語 [zam]
293. サパロ語 [zro]
294. グイラ・サポテカ語 [ztu]
295. ズールー語 [zul]

セルビア語グループ

1. ボスニア語 [bos_latn]
2. クロアチア語 [hrv]
3. セルビア語 [srp_latn]

コンゴ語グループ

1. コンゴ語 [kng]
2. キトゥバ語 [ktu]

A.1.8 その他

アルメニア文字　アルメニア語 [hye]
ベンガル文字　ベンガル語 [ben]
漢字　中国語 [cmn_hans]
グルジア文字　グルジア語 [kat]
ギリシア文字　ギリシア語 [ell_monotonic]
グジャラーティー文字　グジャラート語 [guj]
グルムキー文字　東パンジャーブ語 [pan]
ハングル　韓国朝鮮語 [kor]
漢字かな交じり　日本語 [jpn]
カンナダ文字　カンナダ語 [kan]
クメール文字　中部クメール語 [khm]
ラオ文字　ラオ語 [lao]
マラヤーラム文字　マラヤーラム語 [mal]
ビルマ文字　ミャンマー語 [mya]
シリア文字　アッシリア現代アラム語 [aii]

タミル文字　タミル語 [tam]
ターナ文字　モルディブ語 [div]
タイ文字　タイ語 [tha]
チベット文字　中央チベット語 [bod]
ヴァイ文字　ヴァイ語 [vai]
口口文字　四川イ語 [iii]

A.2　Wikipedia

A.2.1　アラビア文字

1. アラビア語 [ar]
2. エジプト・ アラビア語 [arz]
3. アゼルバイジャン語 [az]
4. ソラニー・ クルド語 [ckb]
5. ペルシア語 [fa]
6. ギラキ語 [glk]
7. カシミール語 [ks]
8. マーザンダラーン語 [mzn]
9. 西パンジャーブ語 [pnb]
10. パシュトゥー語 [ps]
11. シンド語 [sd]
12. ウイグル語 [ug]
13. ウルドゥー語 [ur]

A.2.2　ベンガル文字

1. アッサム語 [as]
2. ベンガル語 [bn]
3. ビシュヌプリヤ・ マニプリ語 [bpy]

A.2.3　漢字

1. カン語 [gan]
2. 呉語 [wuu]
3. 古典中国語 [zh_classical]
4. 広東語 [zh_yue]

A.2.4　キリル文字

1. アブハズ語 [ab]
2. アヴァル語 [av]
3. バシキール語 [ba]
4. ベラルーシ語 [be]
5. ブルガリア語 [bg]
6. モンゴル語グループ [bxr] [mn] [xal]
7. 教会スラブ語 [cu]
8. チュヴァシ語 [cv]
9. カバルド語 [kbd]

10. カザフ語 [kk]
11. コミ・ ペルミャク語 [koi]
12. カラチャイ語 [krc]
13. コミ語 [kv]
14. キルギス語 [ky]
15. ラク語 [lbe]
16. レズギ語 [lez]
17. モクシャ語 [mdf]
18. 牧地マリ語 [mhr]
19. マケドニア語 [mk]
20. モルダビア語 [mo]
21. 山地マリ語 [mrj]
22. オセット語 [os]
23. ロシア語 [ru]
24. ルシン語 [rue]
25. ヤクート語 [sah]
26. セルビア語グループ [sh][sr]
27. タジク語 [tg]
28. タタール語 [tt]
29. ウドムルト語 [udm]
30. ウクライナ語 [uk]

モンゴル語グループ

1. ブリヤート語 [bxr]
2. モンゴル語 [mn]
3. カルムイク語 [xal]

セルビア語グループ

1. セルボ・ クロアチア語 [sh]
2. セルビア語 [sr]

A.2.5 デーヴァナーガリ

1. ビハール語 [bh]
2. ヒンディー語 [hi]
3. マラーティー語 [mr]
4. ネパール語 [ne]
5. ネワール語 [new]
6. パーリ語 [pi]
7. サンスクリット語 [sa]

A.2.6 グルジア文字

1. グルジア語 [ka]
2. メグレル語 [xmf]

A.2.7 ギリシア文字

1. ギリシャ語 [el]
2. ポントス・ ギリシア語 [pnt]

A.2.8　ヘブライ文字

1. ヘブライ語 [he]
2. ラディノ語 [lad]
3. イディッシュ語 [yi]

A.2.9　ラテン文字

1. アチェ語 [ace]
2. アフリカーンス語 [af]
3. アカン語 [ak]
4. アレマン語 [als]
5. アラゴン語 [an]
6. 古英語 [ang]
7. アストゥリアス語 [ast]
8. アイマラ語 [ay]
9. アゼルバイジャン語 [az]
10. バイエルン・ オーストリア語 [bar]
11. サモギティア語 [bat_smg]
12. ビコール語 [bcl]
13. ビスラマ語 [bi]
14. バンジャル語 [bjn]
15. バンバラ語 [bm]
16. ブルトン語 [br]
17. セルビア語グループ [bs][hr][sh][sr]
18. カタロニア語 [ca]
19. スペイン語グループ [cbk_zam][es]
20. 閩東語 [cdo]
21. セブアノ語 [ceb]
22. チャモロ語 [ch]
23. コルシカ語 [co]
24. クリミア・ タタール語 [crh]
25. チェコ語 [cs]
26. カシューブ語 [csb]
27. ウェールズ語 [cy]
28. デンマーク語 [da]
29. ドイツ語 [de]
30. ザザキ語 [diq]
31. 低ソルビア語 [dsb]
32. エウェ語 [ee]
33. エミリア・ ロマーニャ語 [eml]
34. 英語 [en]
35. エスペラント語 [eo]
36. エストニア語 [et]
37. バスク語 [eu]
38. エストレマドゥーラ語 [ext]
39. フラニ語 [ff]
40. フィンランド語 [fi]
41. ヴォロ語 [fiu_vro]
42. フェロー語 [fo]
43. フランス語 [fr]
44. アルピタン語 [frp]
45. 北フリジア語 [frr]
46. フリウリ語 [fur]
47. 西フリジア語 [fy]
48. アイルランド語 [ga]
49. トルコ語グループ [gag][tr]
50. スコットランド・ ゲール語 [gd]

51. ガリシア語 [gl]
52. グアラニー語 [gn]
53. ゴート語 [got]
54. マン島語 [gv]
55. ハウサ語 [ha]
56. 客家語 [hak]
57. ハワイ語 [haw]
58. フィジー・ ヒンディー語 [hif]
59. 上ソルビア語 [hsb]
60. ハイチ語 [ht]
61. ハンガリー語 [hu]
62. インターリングア [ia]
63. インドネシア・マレーシア語グループ [id][map_bms][ms]
64. インターリング [ie]
65. イボ語 [ig]
66. イロカノ語 [ilo]
67. イド語 [io]
68. アイスランド語 [is]
69. イタリア語 [it]
70. ロジバン語 [jbo]
71. ジャワ語 [jv]
72. カラカルパク語 [kaa]
73. カビル語 [kab]
74. コンゴ語 [kg]
75. キクユ語 [ki]
76. グリーンランド語 [kl]
77. ケルン語 [ksh]
78. クルド語 [ku]
79. コーンウォール語 [kw]
80. ラテン語 [la]
81. ラディノ語 [lad]
82. ルクセンブルク語 [lb]
83. ガンダ語 [lg]
84. リンブルフ語 [li]
85. リグリア語 [lij]
86. ロンバルド語 [lmo]
87. リンガラ語 [ln]
88. リトアニア語 [lt]
89. ラトガリア語 [ltg]
90. ラトビア語 [lv]
91. マダガスカル語 [mg]
92. マオリ語 [mi]
93. マルタ語 [mt]
94. ミランダ語 [mwl]
95. ナウル語 [na]
96. ナワトル語 [nah]
97. ナポリ語 [nap]
98. 低地ドイツ語 [nds]
99. オランダ低ザクセン語 [nds_nl]
100. オランダ語 [nl]
101. ニーノシュク・ ノルウェー語 [nn]
102. ブークモール・ ノルウェー語 [no]
103. ノヴィアル [nov]
104. ノルマン語 [nrm]
105. 北部ソト語 [nso]
106. ナバホ語 [nv]
107. ニャンジャ語 [ny]
108. オック語 [oc]
109. オロモ語 [om]
110. パンガシナン語 [pag]
111. パンパンガ語 [pam]
112. パピアメント語 [pap]
113. ピカルディ語 [pcd]
114. ペンシルベニアドイツ語 [pdc]
115. プファルツ語 [pfl]
116. ポーランド語 [pl]
117. ピエモンテ語 [pms]
118. ポルトガル語 [pt]
119. ケチュア語 [qu]
120. ロマンシュ語 [rm]
121. ロマ語 [rmy]

122. ルンディ語 [rn]
123. ルーマニア語 [ro]
124. タラント語 [roa_tara]
125. ルワンダ語 [rw]
126. スコットランド語 [sco]
127. 北サーミ語 [se]
128. サンゴ語 [sg]
129. スロバキア語 [sk]
130. サモア語 [sm]
131. ショナ語 [sn]
132. ソマリ語 [so]
133. アルバニア語 [sq]
134. スリナム語 [srn]
135. スワジ語 [ss]
136. 東フリジア語 [stq]
137. スンダ語 [su]
138. スウェーデン語 [sv]
139. スワヒリ語 [sw]
140. シレジア語 [szl]
141. テトゥン語 [tet]
142. トルクメン語 [tk]
143. タガログ語 [tl]
144. ツワナ語 [tn]
145. トンガ語 [to]
146. トク・ ピシン語 [tpi]
147. ツォンガ語 [ts]
148. タタール語 [tt]
149. トゥンブカ語 [tum]
150. トウィ語 [tw]
151. タヒチ語 [ty]
152. ウズベク語 [uz]
153. ベンダ語 [ve]
154. ヴェネツィア語 [vec]
155. ヴェプス語 [vep]
156. ベトナム語 [vi]
157. 西フラマン語 [vls]
158. ヴォラピュク語 [vo]
159. ワロン語 [wa]
160. ウォロフ語 [wo]
161. ヨルバ語 [yo]
162. チワン語 [za]
163. ゼーランド語 [zea]
164. 閩南語 [zh_min_nan]
165. ズールー語 [zu]

セルビア語グループ

1. ボスニア語 [bs]
2. クロアチア語 [hr]
3. セルボ・ クロアチア語 [sh]
4. セルビア語 [sr]

スペイン語グループ

1. チャバカノ語 [cbk_zam]
2. スペイン語 [es]

トルコ語グループ

1. ガガウズ語 [gag]
2. トルコ語 [tr]

インドネシア・マレーシア語グループ

1. インドネシア語 [id]
2. バニュマス語 [map_bms]
3. マレー語 [ms]

A.2.10 チベット文字

1. チベット語 [bo]
2. ゾンカ語 [dz]

A.2.11 その他

アルメニア文字 アルメニア語 [hy]
カナダ先住民文字 イヌクウティトット語 [iu]
チェロキー文字 チェロキー語 [chr]
ゲエズ文字 アムハラ語 [am]
ゴート文字 ゴート語 [got]
グジャラーティー文字 グジャラート語 [gu]
グルムキー文字 パンジャブ語 [pa]
ハングル 韓国語 [ko]
漢字かな交じり 日本語 [ja]
カンナダ文字 カンナダ語 [kn]
クメール文字 クメール語 [km]
ラオ文字 ラオ語 [lo]
マラヤーラム文字 マラヤーラム語 [ml]
ビルマ文字 ビルマ語 [my]
オリヤー文字 オリヤー語 [or]
ルーン文字 古英語 [ang]
シンハラ文字 シンハラ語 [si]
アラム文字 アラム語 [arc]
タミル文字 タミル語 [ta]
テルグ文字 テルグ語 [te]
ターナ文字 ディベヒ語 [dv]
タイ文字 タイ語 [th]

付録 B

実データセット

8 章で用いた実データの入手元の記事について URL を掲載する．なお, それぞれの特徴については, 8 章にある表を参照してほしい.

1. ルクセンブルク語版より [*1]

記事名	Vëlkermord an der Vendée
URL	http://lb.wikipedia.org/wiki/V%C3%ABlkermord_an_der_Vend%C3%A9e
分類	引用 / 段落内
ルクセンブルク語	Ausserdem notéiert hien datt
フランス語	les révolutionnaires n'ont pas cherché à identifier un peuple pour le détruire, regardant simplement la Vendée comme le symbole de toutes les oppositions à la Révolution,
ルクセンブルク語	a schreift a senger Konklusioun datt
フランス語	les atrocités commises par les troupes révolutionnaires en Vendée relèvent de ce qu'on appellerait aujourd'hui des crimes de guerre.

2. マレー語版より [*2]

記事名	Senarai reka cipta China
URL	http://ms.wikipedia.org/wiki/Senarai_reka_cipta_China

[*1] 図 2.1
[*2] 図 2.2

分類	不完全な翻訳 / 段落内
マレー語	Minuman ditapai: Ahli arkeologi telah menemukan sisa meniuman ditapai yang adalah 9,000-tahun dari tapak Neolithik di Jiahu, Henan.
英語	Ujian kimia (including gas and liquid chromatography-mass spectrometry, infrared spectrometry, and stable isotope analysis) have revealed a fermented beverage of hawthorn fruit and wild grape, beeswax associated with honey, and rice. Herbal wine and a filtered rice or millet beverage was found 5000 years later in sealed Shang and Western Zhou bronze containers and has been identified as containing specialized rice or millet, flavored with herbs, flowers, and possibly tree resins. It was found that the chemical composition of the samples is similar to those in modern rice, rice wine, grape wine, beehive wax, tannins, several herbal medicines and hawthorn.

3. タガログ語版より

記事名	Hanji
URL	`http://tl.wikipedia.org/wiki/Hanji`
分類	不完全な翻訳 / 段落間
タガログ語	papel na Korean o hanji ay ang pangalan ng tradisyonal na papel na gawang kamay mula sa Korea. Korea. Hanji ay ginawa mula sa kalooban na tahol ng Papel ng Mulberry, isang puno sa Korea na lumalaki ng mabuti sa mababato na bundok, na kilala sa Korean bilang dak. Ang mahalaga sa paggawa ng hanji ay ang uhog na lumalabas mula sa mga ugat ng Hibiscus manihot. Ang substansiya na ito ay tumutulong sa isuspinde ng mga indibidwal na fibers sa tubig.

英語	Papermaking methods that originated in China migrated to Korea and were likely well-developed by the 6th century. These methods are similar to those used in Japan to make washi but differ in sheet formation techniques (traditional hanji is made in laminated sheets using the we bal method, which allows for multi-directional grain) and calendering (dochim is a method of pounding finished sheets to compact fibers and lessen ink bleed).
タガログ語	Paggagawa ng papel na pamamaraan na buhat sa Tsina ay lumipat sa Korea at ito ay mahusay na binuo ng ika-anim na siglo. Ang mga pamamaraan ay katulad sa mga ginagamit sa Japan upang gumawa ng washi pero nakakaiba ito sa mga pamamaraan ng pagpormasyon ng sheet.

4. アブハズ語版より

記事名	Казан аметрополитен астанциақәа рсиа
URL	http://ab.wikipedia.org/wiki/%D0%9A%D0%B0%D0%B7%D0%B0%D0%BD_%D0%B0%D0%BC%D0%B5%D1%82%D1%80%D0%BE%D0%BF%D0%BE%D0%BB%D0%B8%D1%82%D0%B5%D0%BD_%D0%B0%D1%81%D1%82%D0%B0%D0%BD%D1%86%D0%B8%D0%B0%D2%9B%D3%99%D0%B0_%D1%80%D1%81%D0%B8%D0%B0
分類	不完全な翻訳 / 段落内
アブハズ語	Ари Казан аметрополитен астанциақәа рсиоуп. Казан аметрополитен — аметрополитен асистема Казанаҟны (Татарстан аресӡублика аҳтнлқалақь) ауп.
ロシア語	Первая и единственная линия была открыта 27 августа 2005 года и на данный момент состоит из семи станций.

5. ヤクート語版より

記事名	Виташевскай Николай Алексеевич
URL	
分類	引用 / 段落内

ヤクート語	Саха сирин этнографиятыгар саамай элбэх үлэни Н.А. Виташевскай 1892-1896 сс. Сибиряков экспедициятын саҕана оҥорбута. 1894 с. хас да улууһу кэрийбитэ, ол иһигэр Боотурусскай, Бороҕон, Дүпсү, Нам улуустарын; антропометрияҕа чинчийиилэри оҥорбута, уобалас экономикатын уонна суутун-сокуонун туругун («юридическая программа») үөрэппитэ.
ロシア語	Результатом деятельности Н.А. Виташевского в экспедиции стали его работы – «Способы разложения и сбора податей в 1 Якутской общине», «Основные правила распределения земли у якутов Дюпсинского улуса Якутского округа», «Якутские материалы для разработки вопросов эмбриологии права». Источниками для написания последнего исследования стали сведения, полученные от «родоначальников», личные наблюдения, архивные материалы, сообщения коллег по экспедиции (Э.К. Пекарский, В.М. Ионов, В.Ф. Трощанский). В этой работе Н.А. Виташевский рассмотрел вопросы землевладения, землепользования, обязательного права, брака и родства, системы родства, семейного права, наследственного права, уголовного права, судоустройства и судопроизводства. Учитывая научную значимость очерков о юридическом быте народов Сибири, работа Н.А. Виташевского «Якутские материалы для разработки вопросов эмбриологии права» была удостоена малой золотой медали Общества по отделению этнографии.

6. チュヴァシ語版より

記事名	Савкилта
URL	`http://cv.wikipedia.org/wiki/%D0%A1%D0%B0%D0%B2%D0%BA%D0%B8%D0%BB%D1%82%D0%B0`
分類	引用 / 段落内
チュヴァシ語	Савкилта çинчен пирĕн историк В.Д. Дмитриев хăйĕн "История Чувашии XVIII века"кĕнекинче уçăмлă çырать:

ロシア語	"Повстанцы-чуваши храбросражались при взятии Казани. Крестьянин-чуваш д. Бешпуляк Уфимского уезда Борис Савельев (Савгилда), с самого начала 1774 г. сражавшийся в армии Пугачева, 12 июля, когда пугачевцы отражали в Казани атаки отряда Михельсона, был послан для набора повстанцев в армию повстанцев..."
チュヴァシ語	Савкилта тăрăшнипе ĕнтĕ Пугачев Хусан хули патнелле пынă чухне кĕрешӳçĕсен çарне çĕршерен-çĕршерен Чулман Атăл таврашĕнчи чăвашсем хутшăнаççĕ. Çавна шута илсе, Пугачев хăйĕн шанчăклă çарпуçне татах çар пухма хушать. Савкита вара Исхак Ахметов тутар тусĕпе пĕрле Хусан уесне тухса каять... Анчах ăна патша салтакĕсем тытаççĕ. Ăна тискеррĕн асаплантараççĕ, юлашкинчен катăркăна ямалла тăваççĕ.

7. タジク語版より

記事名	Ҳомид Карзай
URL	http://tg.wikipedia.org/wiki/%D2%B2%D0%BE%D0%BC%D0%B8%D0%B4_%D0%9A%D0%B0%D1%80%D0%B7%D0%B0%D0%B9
分類	不完全な翻訳 / 段落内
タジク語	Дар Қандаҳори (Афғонистон) таваллуд шудааст.
ロシア語	Выходец из влиятельного пуштунского племени Популзай, представители которого правили Афганистаном в течение двух столетий.

8. ヨルバ語版より (1)

記事名	Gùyánà
URL	http://yo.wikipedia.org/wiki/G%C3%B9y%C3%A1n%C3%A0
分類	不完全な翻訳 / 段落内
ヨルバ語	Guyana tele je ibiamusin Holandi ati fun ogorun meji odun ti Ileoba Asokan.

英語	It is the only state of the Commonwealth of Nations on mainland South America and the only state in South America where English is the official language. Guyana achieved independence from the United Kingdom on 26 May 1966 and became a republic on 23 February 1970. In 2008, the country joined the Union of South American Nations as a founding member.

9. ヨルバ語版より (2) [*3]

記事名	Ìjálá
URL	http://yo.wikipedia.org/wiki/%C3%8Cj%C3%A1l%C3%A1
分類	不完全な翻訳 / 段落間
ヨルバ語	Ijala fun akoko kan nipa igbesi aye awujo kan ni eleyii. Asunjala tun le yi i pada bi o ba fe lo o fun awujo miiran.
英語	The ijala poems chanted with special regard to particular occasions reflect abundantly the way of life of the community to whom belongs the heritage of ijala-chanting tradition. It must be pointed out that these ijala poems are not a fixed stock from which an ijala-chanter makes wholesale quotations. The expert ijala-chanter while bearing in mind the traditional themes and the poetic clichés for a particular type of occasion, usually sets about improvising for the occasion in hand. That is to say, he composes a new poem of his own in honour of the celebration in progress....
ヨルバ語	Ijala eleyii, eranko ati awon eye ni o je mo. Awon nnkan ti o ti mo nipa awon nnkan wonyi ni asunjala menu ba ninu ijala naa.
英語	There are many ijala poems about the various birds and animals which the hunters hunt in the forst or in the savannah of Western Nigeria. An ijala-artist chants these poems usually when he is alone at work on his farm and he is giving himself some music to lighten his labours or when he is in the company this fellow-hunters and the occasion calls for reminiscences about game birds and game animals. ...

[*3] この記事は私用領域の文字を含んでいたため, □ に置き換えられている

ヨルバ語	Iha ti a ko si iru awon litireso alohun bi ijala ni onkowe menu ba ni isori yii.
英語	In Yorùbá literature, there are several other types of oral poetry apart form ijala, Of these, the most prominent are è□sà, rárà, o□fo□, ègè, and ìyè□rè□. In Igbo literature, Hausa literature, Fula literature, Efik literature, Nupe literature, Edo literature, Ijaw literature and Afenmai literature...to mention only a few of the other indigenous languages of Nigeria...there are also several distinct types of oral poetry, apart from other genres of spoken art. ...
ヨルバ語	Idanilekoo yii, ijala ni o da le lori. Ohun ti onkowe so ni ori yii ni pe litireso ti o ye ki a kaaramaasiki re ni litireso alohun.
英語	If you were to visit a school certificate class in any of the grammer schools in our country to-day you would, I am sure, hear the word 'literature' used by one pupil after another in answer to your request for information about the various subjects being offered by the pupils for the West African School Certificate Examination...
ヨルバ語	Awon ohun ti o maa n je akejaala logun ni o wa ni ori yii. Won maa n fi ijala ki eniyan a kii sii sabaa mo eni ti o seda iru ijala bee. Bi apeere, ijala ti o je mo oba Abiodun ni o wa ni ori yii. Lile ke ijala naa daadaa ni oriyin fun eni ti o ke ijala kii se ti pe boya ohun ni o seda re.
英語	In the repertoire of a master ijala-artist, there is usually a preponderance of ijala poems which are eulogies on individual progenitors or groups of progenitors. The praise poems are anonymous classics which date back to the halcyon days of the Yoruba kingdom during the reign of King Abio□dun (1770-1830). In rendering these classics, the ijala-chanter takes credit for his ability to recall the texts accurately from memory and for his ability to chant them in the special traditional style. The credit for the composition of the texts belongs to anonymous authors of bygone eras....

ヨルバ語	Eni kan ti akewi mo daadaa ni o ke ijala nipa re ni ori yii. Loooto, akewi yii yin in daadaa, sibe o so awon okodoro kan ti o koro nipa re.
英語	The theme of many an ijala poem is the charater of a particular personage who is well known to the poet. The poet presents a character-portrait of the personage as a verbal salute to the personage himself and so the portrait tends to be biased in favour of him, but nevertheless, some unpalatable truths about him may be mentioned in the poem....
ヨルバ語	DR. Adeboye Babalo□lá (1963), Ijala (A Form of Oral Poetry in Nigeria) Programme : October Lectures 1963, oju-iwe 1-37.

10. トルコ語版より

記事名	Xavi Torres
URL	`http://tr.wikipedia.org/wiki/Xavi_Torres`
分類	不完全な翻訳 / 段落内
	Xàbia, Province of Alicante'de doğmuştur, gençlik yıllarında iki klüpte yer almıştır; Villarreal CF, ki 2006'ya dek formasyonunu bu klüp altyapısında aldı.
	His professional debuts were made at local Alicante CF in the 2006–07 season, in Segunda División B.

11. セブアノ語版より (1)

記事名	Danny Sillada
URL	`http://ceb.wikipedia.org/wiki/Danny_Sillada`
分類	引用 / 段落内

セブアノ語	Otokan ug daghang talento, si Sillada usa ka magbabalak ug pilosopo nga nagmantala sa iyang mga sinulat sa magasin ug internet, nagtagik ug nagpasundayag sa iyang mga lumad nga kanta ug musika, hip-hop ug "ethno-techno" nga mga huni sa alternatibong mga lugar sa Metro Manila. Ginganlan og “Renaissance Man”, gitimbaya sa “research paper” sa University of Asia and the Pacific nga
英語	“the embodiment of a Filipino who defies the existing trend. His multi-faceted attribute in the humanities, as a Renaissance man, is identical with those of well-rounded historical figures during the Renaissance period in Europe. Sillada is a visual artist recognized in the Philippine art scene for his paintings and installation artworks, a literary writer who is into prose and poetry, a philosopher, whose writings are akin with existentialism, a first-rate performance artist, and also an art-critic."

12. セブアノ語版より (2)

記事名	Simbahang Bawtista sa Westboro
URL	`http://ceb.wikipedia.org/wiki/Simbahang_Bawtista_sa_Westboro`
分類	引用 / 段落内
セブアノ語	Ang Simbahang Bawtista sa Westboro maoy simbahang Independiyenteng Bawtista sa Estados Unidos. Si Fred Phelps, usa ka disbarred nga abogado, ang pundador ug dakodako niini. Kini kailhan sa anti-Hudaismo, Islamopobya, ug anti-Katolisismo niini, ug sa heneral niini nga galit sa tibuok kalibotan. Kini ginakonsidera isip kulto sa daghang organisasyong relihiyoso ug sekular, ug nga,
英語	"A number of Phelps’ critics have suggested that the actions of the Westboro Baptist Church are a ploy to receive attention above all else. Counter-protesting against the group, they suggest, gives them attention and incentive that they do not deserve; and a more effective response against Phelps would be to ignore his congregation completely."

13. ハイチ語版より

記事名	Aprann pale kreyòl ayisyen
URL	http://ht.wikipedia.org/wiki/Aprann_pale_krey%C3%B2l_ayisyen
分類	他言語話者に対する案内 / -
英語	Haitian Creole language (kreyòl ayisyen), often called simply Creole, is a language spoken in Haiti by about 8.5 million people (as of 2005), which is nearly the entire population, and via emigration, about 3.5 million speakers who live in other countries, including Canada, the United States, France, and many Caribbean nations, especially the Dominican Republic, Cuba, and the Bahamas. Haitian Creole is one of Haiti's two official languages, along with French. It is a creole based primarily on French, but it also contains various influences, notably the native Taíno, some West African and Central African languages, Portuguese and Spanish. The language has two distinct dialects: Fablas and Plateau. Guyane, Martinique, Guadeloupe as well as Saint Lucia and Dominica, also speak Creole, with some local variations. Haitian creole tends to move away from original creole under the influence of English introduced by Haitian working in USA. In part because of the efforts of Félix Morisseau-Leroy, since 1961 Haitian Creole has been recognized as an official language along with French, which had been the sole literary language of the country since its independence in 1804, and this status was upheld under the country's constitution of 1987. Its usage in literature is small but increasing, with Morisseau being one of the first and most significant examples. Many speakers are trilingual, speaking Haitian Creole, Spanish, and French. Many educators, writers and activists have emphasized pride and written literacy in Creole since the 1980s. Today there are numerous newspapers, as well as radio and television programs, in Creole.

Haitian Creole is used widely among Haitians who have relocated to other countries, particularly the United States and Canada. Some of the larger population centers include Montréal, Québec, where French is the official language, and parts of New York City, Boston, Central and South Florida (Miami, Fort Lauderdale, and Palm Beach). Various public service announcements, school-parent communications, and other materials are produced in this language by government agencies. Miami-Dade County in Florida sends out paper communications in Haitian Creole in addition to English and Spanish. Announcements are posted in the Boston subway system and area hospitals and medical offices in this language. HTN, a Miami-based television channel, is North America's only Creole-language television network. The Miami area also features over half a dozen Creole-language AM radio stations.

There is some controversy as to whether or not Creole should be taught in Miami-Dade County Public Schools. Many argue Creole is a peasant language which is not important, while others argue it is important for children to know their parents' native tongue.

Haitian Creole language and culture is taught in many Colleges in the United States as well as in the Bahamas. Indiana University has a Creole Institutefounded by Dr. Albert Valdmanwhere Haitian Creole, among other facets of Haiti, are studied and researched; the University of Kansas, Lawrence has an Institute of Haitian studies, founded by Dr. Bryant Freeman. Additionally, the University of Massachusetts-Boston and University of Florida offer seminars and courses every year under their Haitian Creole Summer Institute. More universities such as Brown University, Columbia University, and University of Miami offered numerous classes in Haitian Creole.

In the Americas, Haitian Creole is the second most spoken language in Cuba, where over 300,000 Haitian immigrants speak it. It is recognized as a language in Cuba and a moderate number of mestizo and mulatto Cubans speak it fluently. Surprisingly enough, most of these speakers have never been to Haiti and do not possess Haitian ancestry, but merely learned it in the communities they lived in. In addition, there is a Haitian Creole radio station operating in Havana The language is also spoken by over 150,000 Haitians (although estimates believe that there are over a million speakers due to a huge population of illegal aliens from Haiti) who reside in the neighboring Dominican Republic

Haitian Creole has ten vowels as opposed to standard French's twelve. This is primarily due to the loss of front rounded vowels. In Creole, these French phonemes are usually merged with their unrounded counterpart. Hence, becomes and becomes .

French's uvular rhotic either becomes an alveolar trill , or , or is elided altogether, depending on the environment. Being formed relatively recently, Haitian Creole orthography is mostly phonemic, and is similar to the International Phonetic Alphabet (IPA). The main differences are j = , y = , è = , ou = . Nasalization is indicated by a following n.

Most of the lexicon is derived from French, with significant changes in pronunciation and morphology. Often, the French definite article was retained as part of the noun. For example, the French definite article la in la lune ("the moon") was incorporated into the Creole noun for moon: lalin.

(*) A banana which is short and fat, not a plantain and not a conventional banana; regionally called "hog banana" or "sugar banana" in English.(#) The relationship shared between a child's mother and godmother.(^) The gap between a person's two front teeth.

Many trade marks have become common nouns in Haitian Creole (as happened in English with "aspirin" and "kleenex", for example).

The term nèg literally means a dark-skinned man and the word blan a white person, as in Gen yon nèg e gen yon blan. ("there is a black man and there is a white man"). However, nèg is generally used for any man, regardless of skin color (i.e. like "guy" or "dude" in American English). Blan is generally used for foreigner. It is not used to refer just to white foreigners, but foreigners of other races as well.

Etymologically, the word nèg is derived from the French "nègre" and is cognate with the Spanish negro ("black", both the color and the people)

There are many other Haitian Creole terms for specific tones of skin, such as grimo, bren, wòz, mawon, etc. However, such labels are considered offensive by some Haïtians, because of their association with color discrimination and the Haitian class system.

Haitian Creole grammar differs greatly from French and inflects much more simply: for example, verbs are not inflected for tense or person, and there is no grammatical gender — meaning that adjectives and articles are not inflected according to the noun. The primary word order (SVO) is the same as French, but the variations on the verbs and adjectives are minuscule compared to the complex rules employed by French.

Many grammatical features, particularly pluralization of nouns and indication of possession, are indicated by appending certain suffixes (postpositions) like yo to the main word. There has been a debate going on for some years as what should be used to connect the suffixes to the word: the most popular alternatives are a dash, an apostrophe or a space. It makes matters more complicated when the "suffix" itself is shortened, perhaps making only one letter (such as m or w).

There are six pronouns, one pronoun for each person/number combination. There is no difference between direct and indirect. Some are obviously of French origin, others are not.

(*)sometimes ou is written as w- in the sample phrases, w indicates ou.(**) depending on the situation. Nouns are pluralized by adding yo at the end.

Possession is indicated by placing the possessor after the item possessed. This is similar to the French construction of chez moi or chez lui which are "my place" and "his place", respectively.

The language has an indefinite article yon, roughly corresponding to English "a/an" and French un/une. It is derived from the French il y a un, (lit. "there is a/an/one"). It is placed before the noun:

There is also a definite article, roughly corresponding to English "the" and French le/la. It is placed after the noun, and the sound varies by the last sound of the noun itself. If the last sound is an oral consonant and is preceded by an oral vowel, it becomes la:

If the last sound is an oral consonant and is preceded by a nasal vowel, it becomes lan:

If the last sound is an oral vowel and is preceded by an oral consonant, it becomes a:

If the last sound is an oral vowel and is preceded by a nasal consonant, it becomes an:

If the last sound is a nasal vowel, it becomes an:

If the last sound is a nasal consonant, it becomes nan:

There is a single word sa that corresponds to French ce/ceci or ça, and English "this" and "that". As in English, it may be used as a demonstrative, except that it is placed after the noun it qualifies. It is often followed by a or yo (in order to mark number):

As in English, it may also be used as a pronoun, replacing a noun:

Many verbs in Haitian Creole are the same spoken words as the French infinitive, but they are spelled phonetically. As indicated above, there is no conjugation in the language; the verbs have one form only, and changes in tense are indicated by the use of tense markers.

The concept expressed in English by the verb "to be" is expressed in Haitian Creole by two words, se and ye.

The verb se (pronounced as the English word "say") is used to link a subject with a predicate nominative:

The subject sa or li can sometimes be omitted with se:

For the future tense, such as "I want to be", usually vin "to become" is used instead of se.

"Ye" also means "to be", but is placed exclusively at the end of the sentence, after the predicate and the subject (in that order):

The verb "to be" is not overt when followed by an adjective, that is, Haitian Creole has stative verbs. So, malad means "sick" and "to be sick":

The verb "to have" is genyen, often shortened to gen.

The verb genyen (or gen) also means "there is/are"

There are three verbs which are often translated as "to know", but they mean different things. Konn or konnen means "to know" + a noun (cf. French connaître).

Konn or konnen also means "to know" + a fact (cf. French savoir).

The third word is always spelled konn. It means "to know how to" or "to have experience". This is similar to the "know" is used in the English phrase "know how to ride a bike": it denotes not only a knowledge of the actions, but also some experience with it.

Another verb worth mentioning is fè. It comes from the French faire and is often translated as "do" or "make". It has a broad range of meanings, as it is one of the most common verbs used in idiomatic phrases.

The verb kapab (or shortened to ka, kap' or 'kab) means "to be able to (do something)". It refers to both "capability" and "availability", very similar to the French "capable".

There is no conjugation in Haitian Creole. In the present non-progressive tense, one just uses the basic verb form for stative verbs:

Note that when the basic form of action verbs is used without any verb markers, it is generally understood as referring to the past:

(Note that manje means both "food" and "to eat" – m ap manje bon manje means "I am eating good food").

	For other tenses, special "tense marker" words are placed before the verb. The basic ones are: Note: For the present progressive ("I am eating now") it is customary, though not necessary, to add "right now": A verb mood marker is ta, corresponding to English "would" and equivalent to the French conditional tense: The word pa comes before a verb (and all tense markers) to negate it:

14. スンダ語版より (1)

記事名	Fullerin
URL	http://su.wikipedia.org/wiki/Fullerin
分類	不完全な翻訳 / 段落内
スンダ語	Fullerin sarupa bentukna jeung struktur grafit, nu diwangun ku salambar cingcin héxagonal numbu, tapi cingcinnna péntagonal (atawa kadang héptagonal) nu nyegah lambaranana jadi planar. Fullerin nyolobong mindengna disebut tabungnano (nanotubes).
英語	The smallest fullerene in which no two pentagons share an edge (which is destabilizing — see pentalene) is C60(buckminsterfullerene), and as such it is also the most common.

15. スンダ語版より (2)

記事名	Akurasi jeung présisi
URL	http://su.wikipedia.org/wiki/Akurasi_jeung_pr%C3%A9sisi
分類	不完全な翻訳 / 段落内
スンダ語	Akurasi nyaeta tingkat kadeukeutan sedengkeun presisi nyaeta tingkat kamampuh dijieun deui. Analogi dipake di dieu keur nerangkeun beda antara akurasi jeung presisi nyaeta babandingan target.

	In this analogy, repeated measurements are compared to arrows that are fired at a target. Accuracy describes the closeness of arrows to the bullseye at the target center. Arrows that strike closer to the bullseye are considered more accurate. The closer a system's measurements to the accepted value, the more accurate the system is considered to be.

16. パンパンガ語版より (1)

記事名	Arung
URL	http://pam.wikipedia.org/wiki/Arung
分類	不完全な翻訳 / 段落間
パンパンガ語	Anatomically, ing arung makapatoto kareng vertebrate nung nu karin ya makabale ing nostril, nung nu karing ya mangisnawa (papalub at papalwal a angin) a ausan dang respiration keng English kayabe ne ing asbuk.
英語	In most humans, it also houses the nosehairs, which catch airborne particles and prevent them from reaching the lungs. Within and behind the nose is the olfactory mucosa and the sinuses. Behind the nasal cavity, air next passes through the pharynx, shared with the digestive system, and then into the rest of the respiratory system. In humans, the nose is located centrally on the face; on most other mammals, it is on the upper tip of the snout.
パンパンガ語	Keraklan kareng mammal, ing arung iya ing manimunang organ para keng pamamau. Kabang ing animal sisingap ya, ing angin daragus kapmialtan ning arung at kareng aliwang balangkas a ausan dang turbinate keng pidalan ning arung.

17. パンパンガ語版より (2)

記事名	University of the Assumption

URL	http://pam.wikipedia.org/wiki/University_of_the_Assumption
分類	不完全な翻訳 / 段落間
英語	Under his administration, he was able to construct a new High School building, whose enrollment was declining at the time.
パンパンガ語	Lalam ning kayang pamanibala, mika bayung gusaling High School, a mibababa ing bilang da reng magaral aniang panaung ita.

18. メグレル語版より

記事名	ლევან II დადიანი
URL	http://xmf.wikipedia.org/wiki/%E1%83%9A%E1%83%94%E1%83%95%E1%83%90%E1%83%9C_II_%E1%83%93%E1%83%90%E1%83%93%E1%83%98%E1%83%90%E1%83%9C%E1%83%98
分類	引用 / 段落内
メグレル語	თე ბორჯის იმერეთის მაფენდა გიორგი (1604–1739) ნამუხათ დადიანი სვანჯის ვარზენდა. დადიანქ რხხუ აკოკირა გურია-აფხაზეთწკმა დო თინეფს სინტკიცაშ ღანკით დემოჯგირა. უკული ლევანქ ქჯმიხუჯუ იმერეთიშ გოლინუან ფეოდალეფიშ დოყუნაფა დო 1623 წანას თიქ ილოშქა გიორგი -შ მეხა ქუთეშშა. თაქ მოხვადა კონწარი ჩხუპიქ გოჭოურაწკმა. გიმორძგა დადიანქ, ნამუქათ დოღინჩა მიარე ჭკორი დო მითხუ გიშარხებელი.
グルジア語	"მოვიდა ლევან დადიანი სპითა ადიშარ-აფხაზ-ჯიქითა და იქმნა ბრძოლა ძლიერი და მოწყდნენ მრავალნი. იძლია მეფე გიორგი და ილტვოდა: მაშინ დადიანმა შეიპყრა მრავალნი და წარჩინებულნი და მდაბიური, აღიღო ალაფი და იავარი და წარვიდა ოდიშს."

19. イディッシュ語版より (1)

記事名	לעזביאנקע
URL	http://yi.wikipedia.org/wiki/%D7%9C%D7%A2%D7%96%D7%91%D7%99%D7%90%D7%A0%D7%A7%D7%A2

分類	不完全な翻訳 / 段落間
イディッシュ語	פרויען וואס האבן נישט חתונה מיט מענער נאר וואוינען כדרך אישית מיט אנדערע פרויען אלס א לעבנסשטייגער ווערן אנגערופן לעזביאַנקעס.
ヘブライ語	דער רמב"ם אין זיין ספר משנה תורה הלכות איסורי ביאה פרק כא סימן ח־ט שרייבט אז דאס איז אסור. נשים המסוללות זו בזו–אסור, וממעשה מצריים הוא שהוזהרנו עליו: שנאמר "כמעשה ארץ מצריים . . . לא תעשו" (ויקרא יח,ג); ואמרו חכמים, מה היו עושים–איש נושא איש, ואישה נושאה אישה, ואישה נישאת לשני אנשים. אף על פי שמעשה זה אסור, אין מלקין עליו–שאין לו לאו מיוחד, והרי אין שם ביאה כלל; לפיכך אין נאסרות לכהונה משום זנות, ולא תיאסר אישה על בעלה בזה–שאין כאן זנות. וראוי להכותן מכת מרדות, הואיל ועשו איסור. ויש לאיש להקפיד על אשתו בדבר זה, ולמנוע הנשים הידועות בכך מלהיכנס לה ומלצאת היא אליהן.

20. イディッシュ語版より (2)

記事名	ישראל סאלאנטער
URL	http://yi.wikipedia.org/wiki/%D7%99%D7%A9%D7%A8%D7%90%D7%9C_%D7%A1%D7%90%D7%9C%D7%90%D7%A0%D7%98%D7%A2%D7%A8
分類	不完全な翻訳 / 段落間

イディッシュ語

ר׳ ישראל האט געהאט פיר זון און צוויי טעכטער זיין זון, יום טוב ליפמן (1875-1846), האט געגומען אקאדעמישע טרעניירונג אין קעניגסבורג און אין בערלין, נאך וואס ער האט פארלאזט זיין טאטעס הויז ביי די 15 יאר.ער האט פארענדיגט זיין דאקטאראט אין די אוניווערסיטעט אין וויען, און איז געווען באריימט מיט זיינע ערפינדונגען אין מעכאניק. ער איז געשטארבן נאך אין לעבן פון זיין פאטער. זיין זון, רבי יצחק ליפקין, האט געדינט אלס רב אין יאנוב, קרוז און פארשניץ און איז אין זיינע לעצטע יארן ארויף אויף ארץ ישראל און זיך באזעצט אין ירושלים. געשטארבן אין תרס״ג. אין דעם קונטרס ״חוט המשולש״ וואס איז געדרוקט געווארן אין ירושלים אין תרס״ד ווערט געברענגט פון זיינע חידושי תורה, און אזוי אויך פון זיין טאטנס און זיידענס חידושים. אין יאר תשמ״ח (1988) איז געדרוקט געווארן א ספר ״לוחות אבנים״ מיט זיינע דרשות און מאמרים. נאך א זון, רבי אריה לייב ליפקין־הורוויץ (כ״א אדר תרנ״ו, 1896), איז געווען א רב אין אפאר שטעטלעך און שפעטער אין בערזין. פארפאסער פון ספר ״חיי אריה״. זיין טאכטער, מלכה הינדא, האט חתונה געהאט צו רבי אליהו אליעזר גרודזינסקי, די שווער פון רבי חיים עוזר גרודזינסקי.

ヘブライ語

רבי ישראל מסלנט העמיד תלמידים רבים, ומתורת המוסר שלו יצאו שיטות מוסר שונות, ובהן שיטות המוסר של התלמוד תורה בקלם, ישיבת נובהרדוק (אשר הדגישה את שפלות האדם) וישיבת סלבודקה (אשר טענה שיש להדגיש את כוחו של רוח האדם, ולרוממה). שלושת תלמידיו העיקריים הם: הרב יצחק בלאזר - רבה של פטרבורג ומחבר הספר ״אור ישראל״ אודות רבו ושיטתו המוסרית רבי נפתלי אמסטרדם רבי שמחה זיסל זיו, ״הסבא מקלם״.