外国人名対訳辞書の自動編纂：現状と展望

佐　藤　理　史

名古屋大学大学院工学研究科電子情報システム専攻
ssato@nuee.nagoya-u.ac.jp

1. はじめに

外国人名対訳辞書の自動編纂の研究を開始してから、約 2 年半が経過した。これまで、ウェブからの人名対訳の自動収集を中心に研究を進め、2007 年 3 月の時点で 4.2 万件（精度 87%）[1]、2008 年 3 月の時点では 19 万件（精度 90%）[2] と、収集対訳数およびその精度の向上を図ってきた。

この研究の過程で、次のような疑問が湧いてきた。

- ウェブから対訳を収集する現在の方法で、どれくらいの規模の外国人名対訳辞書を作成可能か？
- そもそもウェブには、どれくらいの数の外国人名対訳が存在するのか？
- 外国人名対訳辞書の自動編纂の最終到達目標をどこに設定すべきなのか？

本稿では、2009 年初頭における外国人名対訳辞書の現状を報告するとともに、上記の疑問に対する現時点の推測および暫定的な解を示す。

2. 対訳ファインダと対訳クローラ

我々が対象としている外国人名対訳とは、ラテン・アルファベット（アクセント記号を伴った主要な文字含む）で表記された原綴と、そのカタカナ訳（綴）からなるペアである。姓や名という人名の構成要素単体ではなく、それらが２つ以上連なって形成される人名を対象とする。

辞書の自動編纂は、まずウェブ上に存在する外国人名対訳を大量に自動収集した後、収集した対訳をある編集方針に従って取捨選択し、最後にそれらを辞書形式にパッケージ化するという手順で実現する。このうち、対訳の自動収集を担当する対訳クローラは、機能的にはほぼ完成している。この対訳クローラの中核部は、対訳ファインダと名付けたプログラムを呼び出す無限ループである。以下では、対訳ファイダ、対訳クローラーの順に説明する。

2.1 対訳ファインダ

対訳ファインダは、与えられた人名（対訳の片側、つまり原綴またはカタカナ綴）から、対訳のもう一方を推定する。まず、与えられた人名をクエリとして検索エンジンを引き、上位 n 件（最大 100 件）のスニペットを得る。こうして得られたスニペットに対して、次の３つのフィルタを順次適用し、すべてをパスした候補を出力する。

2.1.1 S フィルタ

S フィルタは、テキストから、スタイル制約を満たす部分文字列（人名候補文字列）を抽出する。ここで、スタイル制約とは、人名が文字列として満たすべき条件群のことで、原綴およびカタカナ綴のそれぞれに対して定義する。たとえば、原綴に対しては、構成する語の先頭文字は（少数の例外を除いて）大文字でなければならない、最後の語はイニシャルであってはならない、などの条件を定義している。

現在の実装では、スタイル制約を文字列に対する正規表現として定義しており、ある文字列がスタイル制約を満たすか否かは厳密に判定可能である。しかし、ある長い文字列からスタイル制約を満たす部分列を抽出するためには、正規表現による定義に加えて、語の区切りの認定やオーバーラップした場合の扱いなどを別途定める必要がある。S フィルタは、これらの部分に各種のヒューリスティックを用いているため、スタイル制約を満たす全ての部分列を抽出するわけではない。たとえば、“J. D. Salinger” と “D. Salinger” は、いずれもスタイル制約を満たすが、前者の出現に対して、その部分列である後者は抽出しない。

2.1.2 T フィルタ

T フィルタは、S フィルタによって抽出された人名候補文字列のうち、与えられた人名と翻字関係にあるものだけを残すフィルタである。翻字関係のチェックは、２つの文字列間に定義した翻字スコアとそれに対する閾値によって実現されている [1]。

2.1.3 R フィルタ

R フィルタは、T フィルタをパスした候補を、スニペット中の出現数で順位付けし、出現数と翻字スコアを利用した絞り込みを行なう。具体的には、順位 1 位の候補は必ず残し、2 位以下の候補は、その出現数が 1 位の候補の出現数の 10%以上であり、かつ、翻字スコアが 1 位の候補の翻字スコアの 95%以上である場合のみ残す。ここで、2 位以下の候補も残すのは、最も良く使われている綴り以外の綴り（異表記）も収集するためである。

2.2 対訳クローラ

ウェブから大量の人名対訳を収集する対訳クローラを構成するためには、上記の対訳ファインダに加え、対訳ファインダの入力となる人名を収集する機構が必要となる。

ある人名をクエリとして取得したスニペットには、そ

の人名以外の人名が存在することが多い。そこで、まず、対訳ファインダが取得したスニペットに S フィルタを適用し、スタイル制約を満たす文字列をすべて抽出する。次に、人名フィルタ [3] を適用し、人名の可能性の高いもののみを残す。こうして得られた人名候補を次回以降の対訳ファインダの入力とすれば、全体として、人名対訳収集と人名収集が同時に行えることになる。

現在の実装では、(1) 対訳ファインダの出力として得られた人名候補（原綴またはカタカナ綴）と、(2) スニペットから上記の方法で収集したカタカナ綴の人名候補をプールし、次回以降の対訳ファイダの入力として用いる方法を採用している。ここで (1) は、ある方向で見つかった対訳に対して、逆方向からも見つかるかどうかをチェックすることを意味する。(2) でカタカナ綴しか抽出しないのは、最近まで原綴の人名フィルタが未実装だったという理由による。

対訳クローラはすでに 5ヶ月以上稼働しており、これまでに、カタカナ綴 132 万件、原綴 36 万件に対して対訳ファインダを実行し、56 万件の対訳候補を収集した。但し、この間、システムのデバッグと調整を何度も行なっており、収集済みの対訳候補には、最新のフィルタをパスしないものも含まれている。

3. 外国人名対訳辞書『紬 1.0』

これまでに収集した対訳候補集合から外国人名対訳辞書『紬 1.0』を編纂した。

3.1 対訳の推定頻度

辞書の編纂では、各対訳 $\langle en, ja \rangle$ がウェブ上に何件ぐらい存在するかという推定頻度 f を重要な指標として用いる。まず、原綴 en とカタカナ綴 ja の両方をクエリとしていわゆる AND 検索を行ない、そのヒット数 h と n 件（最大 50 件）のスニペットを得る。次に、それぞれのスニペットに対して、原綴 en とカタカナ綴 ja が実際に存在するかどうかを調べ、両者がともに存在するスニペット数 b を求める。これらの値より、推定頻度 f を次式で計算する。

$$f = \begin{cases} \dfrac{bh}{n} & \text{if } h \geq 50 \text{ and } n \geq 50 - \alpha \\ b & \text{otherwise} \end{cases} \qquad (1)$$

ここで α は、スニペット取得エラーに対するマージンで、現在は $\alpha = 10$ を採用している。推定頻度 f が 1 以上ということは、対訳（ペア）で検索エンジンを引き、その対訳を含むスニペットが実際に確認できたことを意味する。

3.2 編 纂 方 法

以下の手順で、外国人名対訳辞書を編纂した。

(1) 対訳クローラ（のデータベース）から、収集済みの対訳候補を取得する。2009 年 1 月 5 日の時点で、561,327 件の対訳候補が得られた。

(2) 最新の S フィルタと T フィルタを適用し、これらをパスする候補を求める。558,883 件の候補が残った。

表 1 対訳のタイプと推定頻度 f

f	type1	type2	type3	total
$f \geq 100$	19880	1821	34	21735
$100 > f \geq 10$	69470	11837	278	81585
$10 > f \geq 3$	125451	39776	1520	166747
$f = 2$	49238	27172	1587	77997
$f = 1$	49375	47761	3896	101032
$f = 0$	11975	892	38	12905
all	325389	129259	7353	462001

(3) それぞれの対訳に、順方向および逆方向のそれぞれの方向から見た場合の順位を付与する。たとえば、順方向から見た場合は、原綴 en を固定し、en を含む対訳候補集合 $\bigcup_i \langle en, ja_i \rangle$ の各要素を、推定頻度で順位付けする。順位 1 位は残し、2 位以下は R フィルタと同じ方法で残すものを決め、それ以外のものは削除する。

(4) 上記の結果を用いて、対訳を次の 3 種類に分類する。

タイプ 1: 順位が両方向とも 1 位である対訳。このタイプの対訳は、原綴とカタカナ綴がどちらも標準的な綴り（代表表記）と考えられる対訳である。

タイプ 2: 片方向の順位が 1 位で、他方の順位が 2 位以下である対訳。原綴とカタカナ綴のいずれか一方が異表記と考えられる対訳である。

タイプ 3: それ以外の対訳。

分類結果を表 **1** に示す。

(5) 推定頻度 $f = 0$ の候補とタイプ 3 の対訳候補をすべて削除する。これにより、441,781 件（タイプ 1 は 313,414 件）の候補が残った。

(6) 最新の人名フィルタを適用する。これにより、406,416 件の候補が残った。

3.3 規　　模

『紬 1.0』の収録対訳数は、406,416 件（タイプ 1 が 286,047 件、タイプ 2 が 120,369 件、原綴の異なり 316,259 件、カタカナ綴の異なり数 379,504 件）である。これに対して、既存の辞書・辞典等の収録数は、次の通りである。

- コンサイス外国人名事典 第三版. 三省堂, 1999. 収録人名項目数 約 18,000 名
- 岩波=ケンブリッジ 世界人名辞典. 岩波書店, 1997. 収録人名項目数 約 15,000 名
- 岩波西洋人名辞典 増補版. 岩波書店, 1981. 収録人名項目数 約 25,000 名
- 現代外国人名録 2008. 日外アソシエーツ, 2008. 収録人数項目数 13,268 名（カタログデータに基づく）
- DSC-西洋人名辞書. 日外アソシエーツ. 収録数 17.5 万人（カタログデータに基づく）

ウィキペディアの言語間リンクから作成できる、英日の固有名詞対訳の数は、HeiNER[4] によれば 12.5 万件であり、そのなかに含まれる人名対訳の数は、我々の調査によれば 4 万から 5 万件である。これらの数字から、『紬

表 2　雪駄を用いた調査

AND 検索のヒット数	対訳数 T	発見数 F	発見率 F/T	紬 1.0 に収録済 type1	type2	total(C)	被覆率 C/T	収集率 C/F
$h \geq 100$	195	189	0.97	164	17	181	0.93	0.96
$100 > h \geq 10$	212	188	0.89	127	38	165	0.78	0.88
$10 > h \geq 3$	113	80	0.71	40	17	57	0.50	0.71
$h \geq 3$	520	457	0.879	331	72	403	0.775	0.881
$h = 2$	55	34	0.62	10	8	18	0.33	0.53
$h = 1$	83	27	0.33	6	4	10	0.12	0.37
$h \geq 1$	658	518	0.789	343	84	431	0.655	0.832
$h = 0$	685	–	–	1	0	1	–	–
total	1343	–	–	348	84	432	–	–

1.0』は、少なくとも収録対訳数においては現存する最大の外国人名対訳辞書と考えてよさそうである。

3.4 精　　度

前節の規模の比較より明らかなように、『紬 1.0』の精度評価に利用できるリファレンスブックは存在しない。このため、ある対訳 $\langle en, ja \rangle$ が正しいか否かを判定するために、評価者は、原綴 en がある特定の人物の名前であり、かつ、カタカナ綴 ja がその訳として実際に用いられていることを、検索エンジンを駆使して確認せざるを得ない。

我々が 2008 年 12 月に行なったタイプ 1 の対訳に対する評価では、サンプル調査した 458 件中、正しい対訳は 393 件、誤りが 61 件、不明が 4 件であった。このときの母集団は、3.2 節のステップ (5) に相当し、その大きさは 229,049 件であった。不明を除く 454 件に人名フィルタを適用すると 420 件が残り、その内訳は正しい対訳 390 件 (93%)、誤り 30 件となる。このデータを単純に外挿すると、『紬 1.0』のタイプ 1 の対訳の精度は 93%程度と推測される。(この結果、これまでと同じように辞書の規模と精度を記述すれば、28 万件（精度 93%）となる。)

4. 雪駄を用いた母集団推定

4.1 雪　　駄

人名対訳に関して各種の調査を行なうために、テストデータが必要である。このため、英語の書籍（原著）4 冊 [5)~8)] とその日本語訳から人名対訳ペア計 1,362 件を収集し、これを整理して、外国人名対訳評価用データ『雪駄』を作成した。対訳の異なりは 1,343 件、原綴は 1,339 件、カタカナ綴は 1,342 件である。

4.2 雪駄を用いた調査

上記の雪駄データを用いて、以下の調査を行なった。

(1) 各対訳（ペア）に対して、AND 検索のヒット数 h を調べた。h は原綴とカタカナ綴の両方を含むページ数を表す。$h \geq 1$ となった 658 件が、特定のウェブページを見ることによって、それが対訳であることを確認できる上限となる。

(2) この 658 件の対訳に対して、対訳ファインダが原綴 en からカタカナ綴 ja を見つけられるかどうかを調べた。発見できたのは 518 件 (79%) であった。

(3) 各対訳が『紬 1.0』に収録されているかどうかを調べた。518 件中 431 件 (83%) がすでに収録済みであった。

これらの結果を表 **2** に示す。ここでは、次の 3 つの指標を併せて示した。

発見率　雪駄に含まれる対訳のうち、どれだけの割合の対訳を対訳ファインダは見つけることができるか。

被覆率　雪駄に含まれる対訳のうち、どれだけの割合が『紬 1.0』に含まれているか。

収集率　対訳ファインダで収集可能な対訳のうち、どれだけの割合の対訳を『紬 1.0』は収集済みか。

上記の定義より明らかなように、発見率は対訳ファインダ単体の性能指標となる。これに対して、収集率は対訳クローラの性能指標および収集状況指標となる。この 2 つの指標のかけ算となる被覆率は、対訳をどれだけ網羅的に収録できているかを表す指標であり、情報検索における recall に相当する。

我々はウェブの情報はそれほど信頼できないことを知っている。対訳が 1 例しか見つからない場合は、その対訳を無条件に信じたりはしない。AND ヒット数 h がそのまま対訳の実例数とはならないことを考慮し、我々は $h \geq 3$ を 1 つの基準として設定する。つまり、「ウェブ上に存在するすべての人名対訳を対象とするのではなく、AND ヒット数が 3 以上の人名対訳を対象とする」。このように対象を設定した場合、現在の状況は、発見率 87.9%、収集率 88.1%、被覆率 77.5%となる。

4.3 母集団推定

ここで、「雪駄データがウェブ上の人名対訳集合の適切なサンプル集合となっている」ことを仮定しよう。このような仮定の下で、次式によって、ウェブ上に存在する人名対訳の異なり数を推定することが可能となる。

$$\text{ウェブ上の対訳数} = \frac{\text{紬の収録数}}{\text{雪駄に対する紬の被覆率}} \quad (2)$$

この式の計算結果を表 **3** に示す。なお、計算は行単位（横方向）で行なった。

当然のことながら、上記の仮定は成立しない。これは、表 3 の T 欄と E 欄を比較すれば一目瞭然である。しかしながら、この仮定の下で得られる数字は、現時点において我々が知りうる、それなりの根拠を持った唯一の推定値である。

この推定値を用いて、ウェブ上に存在する人名対訳数

表 3　母集団推定

	雪駄			紬 1.0			ウェブ
	対訳数	被覆数	被覆率	type1	type2	total	推定数
	T	C	C/T			E	ET/C
$h \geq 100$	195	181	0.93	27453	4434	31887	34353
$100 > h \geq 10$	212	165	0.78	106906	28311	135217	173733
$10 > h \geq 3$	113	57	0.50	89714	38493	128207	254165
$h \geq 3$	520	403	0.775	224073	71238	295311	381046
$h = 2$	55	18	0.33	29944	17112	47056	143782
$h = 1$	83	10	0.12	32030	32019	64049	531607
$h \geq 1$	658	431	0.655	286047	120369	406416	620468

を大胆に予想してみよう。

(1) $h \geq 100$ の人名対訳の推定値は 3.4 万である。被覆率はかなり高い (0.93) なので、対訳数は 4 万件程度と考えてよいだろう。

(2) $100 > h \geq 10$ の人名対訳の推定値は 17 万である。対訳数は 20–25 万の範囲に収まると予想される。

(3) $10 > h \geq 3$ の人名対訳の推定値は 25 万である。対訳数はおおよそ 30–35 万の範囲に収まるのではないかと推測される。

これらの予想の下限と上限をそれぞれ合計すれば、54 万–64 万という数字が出てくる。これが現時点で私が予想する、$h \geq 3$ の人名対訳の異なり数である。さらにざっくりと言ってしまえば、60 万となる。これが、1 節に示した 2 番目の疑問「そもそもウェブには、どれくらいの数の外国人名対訳が存在するのか？」に対する、$h \geq 3$ の対訳に限った場合の現時点の答である。

5. 展　　望

残りの 2 つの疑問に対する答も示そう。

- ウェブから対訳を収集する現在の方法で、どれくらいの規模の外国人名対訳辞書を作成可能か？

先の予想値の 80%を収録できると仮定するならば、収録対訳数 44 万–52 万の辞書が作成可能である。

- 外国人名対訳辞書の自動編纂の最終到達目標をどこに設定すべきなのか？

ウェブの成長を考慮に入れるならば、収録対訳数の絶対値を目標とするのは適切ではない。ウェブ上に存在する外国人名対訳に対する被覆率を目標とすべきである。雪駄データ ($h \geq 3$) に対して被覆率 77.5%という結果が得られているが、実際の数値はそれよりも低い 50–65%に留まっている可能性が高い。

しかしながら、結局のところ、計測できない値は目標とはできず、雪駄のようなテストデータに対する被覆率を目標として掲げざるを得ない。これまでの経験から予想すると、テストデータに対する被覆率 80%はおそらく達成できる目標である。被覆率 85%の達成はかなり困難と考えられるが、不可能ではないかもしれない。この値が現時点の最終到達目標である。

一方、精度に関しては、タイプ 1 の対訳に対する精度 95%が到達目標である。現在の推定精度は 93%でその差は 2%であるが、被覆率を維持したまま、この 2%を上げることはなかなか手強いと予想している。

これまでの研究によって、十分な量の人名対訳が収集できるようになってきたため、辞書形式へのパッケージ化の研究を本格的に開始できる状況になってきた。ここでは、見出し語集合の選定が主な研究課題となるが、これに関しては、『紬 1.0』の編纂で行なった、推定頻度 f を用いた選択を採用する方針である。この他にも、異表記の扱い（イニシャルを含む人名の扱い、アクセント記号の有無の同一化）や検索・ブラウジング機構など、いくつかの研究課題がある。

謝辞　雪駄データの一部の入力は、影浦峡准教授（東大）の研究グループによる（科学研究費補助金基盤研究 (A)「翻訳者を支援するオンライン多言語レファレンス・ツールの構築」課題番号 17200018 の支援を得ている）。この研究は、栢森情報科学振興財団の助成（「ウェブを利用した対訳辞書の自動編纂」）を受けて遂行された。

参　考　文　献

1) 榊原洋平, 佐藤理史. 2007. ウェブを用いた外国人名事典の自動編纂. 言語処理学会第 13 回年次大会発表論文集, pp.879-882.
2) 榊原洋平, 佐藤理史. 2008. 外国人名対訳辞典の大規模化 —１５万件の自動編纂— . 言語処理学会第 14 回年次大会発表論文集, pp.833-836.
3) 開出紗代子, 佐藤理史. 2009. 生起確率の差を用いた人名判定. 言語処理学会第 15 回年次大会発表論文集, A1-3.
4) Wolodja Wentland, Johannes Knopp, Carina Silberer and Matthias Hartung. 2008. Building a Multilingual Lexical Resource for Named Entity Disambiguation, Translation and Transliteration. Proceedings of the Sixth International Language Resources and Evaluation (LREC'08).
5) Malcolm Gladwell. 2002. *The Tipping Point: how little things can make a big difference*. Back Bay Books, paperback edition.
6) Jonathon Green. 1996. *Chasing the Sun: Dictionary Makers and the Dictionaries They Made*. Henrt Holt and Company.
7) Jeremy Leggett. 2005. *Half Gone*. Portobello.
8) Lawrence Lessig. 2004. *Free Culture*. Penguin.