Original Paper
原著論文

日本語と中国語における確率的言語知識構造の比較研究

The comparative study of probabilistic knowledge structure between Japanese and Chinese

張 寓杰[1]，董 媛[1]，太 蘭[1]，寺井 あすか[1]，中川 正宣[1]
[1] 東京工業大学社会理工学研究科人間行動システム専攻

Yujie Zhang[1], Yuan Dong[1], Lan Tai[1], Asuka Terai[1], and Masanori Nakagawa[1]
[1]Tokyo Institute of Technology Graduate School of Decision Science and Technology,
Department of Human System Science
2-12-1 Ookayama, Meguro-ku, Tokyo, JAPAN 152-8550

キーワード：確率的言語知識構造，日本語，中国語，比較
Key words：Probabilistic knowledge structure, Japanese, Chinese, Comparison

抄録

本研究の目的は，日本語，中国語の大規模データに言語統計解析を用いて確率的言語知識構造を構築し，個々の言語の表面的な違いを超えて両言語の背景にある文化や社会システムを比較考察することである．本研究では，従来用いられている潜在意味分析（Latent Semantic Analysis）等の問題点を解消した，より厳密な確率的方法として，Kameya & Sato[6] のアルゴリズムを用いる．具体的には，同手法を用いて推定された両言語における多数の潜在クラスから，特に分析結果の解釈に基づき，中国語と日本語で同じネーミングになったクラスを選んで，比較考察した．

本研究の方法を用いて，日本語，中国語に限らず，より多くの他言語，たとえば英語の大規模データベースを用いて，同様の確率的言語知識構造を構成し，多角的な比較研究を進めていくことも今後の重要な課題の一つである．

1．はじめに

1.1. 確率的言語知識構造とは

ここでいう言語知識構造とは，語彙を意味的に分類し，体系的に整理した一種の構造的データベースを意味する．特に確率的言語知識構造とは，そのような知識構造にさらに単語間の確率的関係を付与したものである．言語知識を体系的に表現する方法は，シソーラス（類語辞書）の編纂など過去からの多くの蓄積がある．しかし，このような方法では多くの人的資源や時間を要し，その客観性も十分保証されていない．最近では，統計的な解析手法に基づいたテキストマイニングや潜在意味分析（Latent Semantic Analysis）と呼ばれる方法が注目されている[1]．潜在意味分析では，類似した意味を持つ単語は類似した文脈の中に現れるという仮定に基づき，文書-単語間の頻度行列に対する特異値分解を用いて構成した潜在的意味空間内のベクトルとして，単語の意味を表現する．

しかし，潜在意味分析ではスパースネス問題（学習・解析データに出現しなかった単語，または単語組の生起頻度が 0 となってしまう問題）を解消できない．さらにこの方法では，特異値分解で得られる空間の次元の意味を一意的に決めることは困難である．また，この方法においては文書内での単語間の指示関係を考慮しておらず，一定範囲内（文や節など）での単純な出現頻度を用いている．そのため，この方法では意味的に結びつきがたい単語との共起もカウントしてしまい，ノイズを多分に含んだ結果を出す危険性が高い．

これらの問題を解消するため，近年では潜在意味分析の方法を確率的な形式で表現した PLSI（Probabilistic Latent Semantic Indexing）を始めとする，確率的潜在意味分析の手法が提案されており[2,3,4]，概念表現の精度において優れていることが報告されている[3,4,5]．Hofmann の提案した PLSI では，ある文書内における単語の生起の背後に，単語間の共起を支える潜在意味クラスがあると仮定し，ある単語の生起は，潜在意味クラスそ

のものの生起確率，および潜在意味クラスが生起した前提での文書, 単語の出現確率（条件付確率）によって確率的に表現されると仮定している．したがって，この手法における文書 d 内の単語 w の出現確率は以下の式で表現される．

$$P(d,w) = \sum_{c \in C} P(d|c)P(w|c)P(c) \qquad (1)$$

ここでの P(c) は潜在クラスの出現確率である．この手法では，文書の内容は k 個の意味クラスに基づき確率的に表現でき，情報の圧縮も可能になる．上記 PLSI を含めた確率的潜在意味分析では，潜在意味分析と同様，単語同士の共起関係の背後に潜在変数が媒介すると仮定する．ただし，確率的潜在意味分析においては，まず，単語―潜在意味クラス間の関係強度を両者の生起確率（及び条件付確率）で表現し，その確率関係に基づき単語の意味を潜在意味クラスによって分類，表現する．

しかし， PLSI の方法は，主に文書の分類に有効な方法である． 一方，人間は文書から単語を連想するよりも単語から単語を連想すると仮定する方が自然である．つまり，人間の言語知識を表現する場合，上記のように文書と単語という関係を前提とするのではなく，単語同士の関係で記述する方が適している．たとえば，Pereira et al.[2]は上記 PLSI と同様の潜在意味クラスの仮定に基づき，単語間の共起確率を表現する手法を提案している．同様の仮定に基づき，より厳密な確率的方法を用いてパラメータを推定する手法として，Kameya & Sato[6]によるアルゴリズムが挙げられる．この手法では，名詞と形容詞もしくは動詞の共起（係り受け）確率 $P(n_i, a_j)$ は以下の式で表現される．

$$P(n_i, a_j) = \sum_k P(n_i|c_k)\, P(a_j|c_k) P(c_k) \qquad (2)$$

上記式 (2) における a_j は動詞もしくは形容詞，n_i は名詞，c_k は潜在意味クラスを表す．$P(c_k)$ は潜在意味クラスの生起確率，$P(a_j|c_k)$，$P(n_i|c_k)$ は意味クラスが生起した前提での形容詞もしくは動詞，および名詞の出現確率（条件付確率）を表す．$P(c_k)$ の値は，EM アルゴリズムを用いて尤度最大にするように推定される（方法の詳細は Kameya & Sato [6]参照）．この方法では，k 個の各潜在意味クラスに対する k 次元の帰属確率の分布として，各単語の意味が表現される．この方法は，確率的な制約条件下での尤度の最大化という客観的基準に従ってパラメータ推定を行い，潜在意味クラスを一意に決定できる．また，この手法ではスパースネス問題を解消できる（具体的方法は付録参照）．これらの点からも，潜在意味分析に比して情報理論的な根拠と，より高い客観性を備えた手法であると言える．さらに，この手法では，文の係り受け解析に基づき単語間の共起頻度を計算しており，前期の無意味な共起が抽出される問題も解消されている．

本研究では中国語，日本語場合ともに各々の確率的言語知識構造の構成に，この Kameya & Sato の方法を用いる．同様の方法は，阿部[7]，Terai[8,9]，Sakamoto（日本語）[10]と，多くの先行研究で用いられており，各研究における計算モデルの前提として，同じ形式の確率的言語知識構造が構成されている．また，これらの確率的言語知識構造の妥当性や信頼性については，日本語の場合は間接的ではあるが，阿部の研究において確認されており，中国語の場合は心理評定との高い相関が示されている[11]．

1.2. 本研究の目的

前記の日本語での先行研究では，主に毎日新聞10年分（1993－2002）の言語データを対象に分析を行った．しかし，これらの研究はすべて日本語の場合に限られており，日本語以外での確率的言語知識構造の構築については，考慮されていない．またこれらの研究で，確率的言語知識構造の構成に用いた大規模言語データは新聞データに限られている．そのため用いられている語彙に偏りがある可能性がある．

本研究では，上記の問題点を考慮し，まず中国語の大規模データに言語統計解析を用いて確率的言語知識構造を構築する．さらに，日本語のデータを前記の先行研究で用いられたたものと同様の新聞データに，文学，評論等のデータを加えて拡張し，確率的言語知識構造を再構築する．

本研究では，日本語，中国語ともに，以下の同じ手順に従って確率的言語知識構造を構成している．

①形態素解析
②係り受け解析
③単語間共起頻度の抽出

④Kameya & Sato[6]のアルゴリズムに基づくクラスタリング

特に，この手続きの③，④は各言語の文法構造に依存しない，共通の手法を用いている．その結果，この方法で構成される確率的言語知識構造もやはり，個々の言語の文法構造には依存していないと言える．

本研究の目的は，このようにして構成された両言語での確率的言語知識構造を用いて，個々の言語の表面的な違いを超えて両言語の背景にある文化や社会システムを比較考察することである．

２．中国語における確率的言語知識構造の構成方法

中国語の言語データとして以下の表 1 に示したコーパスを用いる．これらのコーパスはすべて一般公開されており，新聞記事や文学作品を含んでいて，政治，経済，社会，スポーツ，犯罪，あるいは文学，芸術等，中国語のさまざまな言語知識領域をカバーすることができる．

形態素解析に関しては，中国科学院計算技術研究所が制作した形態素解析ソフトウェア ICTCLAS を用いる．本研究では ICTCLAS2011 を用いて，タグなしコーパスに対し，文を単語ごとに区切り，品詞をつける作業を行った．

表 1. 本研究で用いた中国語コーパス
（サイズ: 651.44MB）

コーパスの種類	サイズ（MB）
ChineseTreebank4.0(2010 取得)	2.34
人民日報タグ付きコーパス(1998)	23
新京報電子版(2010 取得)	21.1
文学作品の電子テキスト(2010 取得)	605

係り受け分析に関しては，CNP Parser を利用した．CNP Parser は情報通信研究機構により，2010 年 8 月に新しく開発公開された高精度の中国語係り受け解析システムである．CNP Parser の解析結果と人間による構文解析結果を比較すると，全体の一致率は 89.8%であることが報告されている[12]．ただし，ICTCLAS の品詞タグが 99 個であるのに対し，CNP Parser の品詞タグは 33 個（Chinese Treebank の表記と同様）である．その為，ICTCLAS で分析された結果を CNP Parser で解析する前に，品詞情報の変更が必要である．その対応関係の一部が表 2 になる．

表 2　品詞対応関係の例

ICTCLAS	CNParser	意味
1　d, uyy, uls, u,	1　AD	副詞
2　uzhe, ule, uguo	2　AS	アスペクト，動詞が表す行為の過去
3　pba	3　BA	～でもって．～を使って
4　cc	4　CC	等位接続詞
5　m,mq	5　CD	数詞
6　c	6　CS	縦位接続詞

CNP Parser では名詞（主語）－動詞（述語）と名詞（目的語)－動詞二種類の係り受け解析が可能であり，本研究ではその各々の名詞と動詞の組み合わせで共起頻度を抽出した．共起頻度データに，Kameya & Sato[6]の解析方法を適用し中国語の確率的言語知識構造を構築した．これらの結果は前述のように心理評定との高い相関が示されている[11]．

本研究に用いた全ての中国語コーパスを係り受け解析した結果，抽出された名詞と動詞の数は表 3 のようになった．

また，抽出するクラスの数を約「名詞数÷200」とすると，各クラスへのメンバーシップ値（各名詞が各クラスに属する程度を示す確率 $P(n_i|c_k)$）が十分高い名詞の数が 200 程度となり，各クラスの意味が明解になりやすいことが経験的にわかっている．また，付録の α の値を変化させて分析した結果，クラス数が 200 以上では，潜在意味クラスの生起確率がほとんど 0 になることも確認された．よって，本研究でもクラス数を約「名詞数÷200」とする基準を採用し，「目的語－動詞」，「主語―動詞」の場合共に抽出するクラス数を 100(計 200)とした．

表 3. 中国語の分析対象となる
コーパスの異なり語数

	名詞（目的語）-動詞	名詞（主語）-動詞（述語）
名詞数	23,882	24,090
動詞数	13,843	14,776
クラス数	100	100

３．日本語における確率的言語知識構造の構成方法

日本語の確率的言語知識構造は中国語と同じ手順を用いて構成している．

今回用いた拡張された日本語コーパスの内容と

サイズは以下の表 4 に示す通りである．

これらのコーパスに形態素解析，係り受け解析の順に実行し単語間の共起頻度を求めた．

まず形態素解析には日本語の形態素解析ツール MeCab を用い，係り受け関係，すなわち「名詞（主語）」が「動詞（述語）」，「名詞（目的語）」を「動詞（述語）」の係り受けの抽出には日本語の係り受け解析ツール CaboCha [13]を使用した．日本語作文支援システム「なつめ」[14]で用いられている抽出ルールに従った．共起頻度データに，Kameya & Sato [6]の解析方法を適用し日本語の確率的言語知識構造を構築した．本研究に用いた全部のコーパスから係り受け解析で抽出された名詞と動詞の数，およびクラスの抽出されたクラスの数は表 5 に示した通りである．ここでのクラス数の決定方法は中国語の場合と同様である．

表 4. 本研究で用いた日本語コーパス
（サイズ: 2.6GB）

コーパスの種類	サイズ（MB）
毎日新聞（1991-2008）	2195
学研国語辞典(2004)	21
小学館百科事典(2004)	119
小学校国語教科書(2006 取得)	4
書籍・新聞記事日英対応付け(2009 取得)	38
自然言語論文(2009 取得)	2
KNB コーパス Version1.0(2009)	1
インターネット図書館青空文庫(2009 取得)	158
フォーマルな会話のコーパス(2009 取得)	1
現代日本語書き言葉均衡コーパス: 国会議事録，政府系白書(2008)	119
インフォーマルな談話を含むコーパス (2009 取得)	4
大学理系基礎科目教科書(2009 取得)	4

表 5. 日本語の分析対象となる
コーパスの異なり語数

	名詞（目的語）-を-動詞（述語）	名詞（主語）-が-動詞（述語）
名詞数	38,816	34,668
動詞数	92,567	70,398
クラス数	200	200

4．構築された中国語と日本語の確率的言語知識構造の比較と考察

すでに述べたように，本研究では，日本語と中国語の確率的言語知識構造を用いて，日本，中国両国の文化や社会の比較を試みる．確率的言語知識構造を比較検討する際，①内容がかなり一致しているクラス，②一見似たようなクラスであるが，内容が異なっているクラスという 2 つの分類に分けた．これらの分類は具体的には以下の方法に従った．

(1) 構築された確率的言語知識構造について，クラスごとに帰属確率の高い単語をリストアップし，それに基づいて各クラスの解釈をおこなった（クラスに名前をつけた）．

(2) 中国語と日本語で同じクラス名になったものについて，それぞれのクラスの帰属確率の高い単語の違いを検討した．

(3) その結果，

①クラスの概念（帰属確率の高い単語リスト）が比較的一致していると思われるクラス:「感情」関連クラス，「金銭」関連クラス，「政府機関」関連クラス，「経済」関連クラス．

②クラスの概念（帰属確率の高い単語リスト）が異なっていると思われるクラス:「権力」と「権利」関連クラス，「古代文化」関連クラス，「闘争」関連クラス「理論」と「学問」関連クラスが存在することがわかった．

(4)そこで，①内容が一致していると思われるクラス，②内容が異なっていると思われるクラスに分けて考察を行った．

以下の表 6.1～表 13.2 にその一部の具体例を示す．これらの表では，特定の一つの潜在意味クラスについて，名詞，動詞ともに単語-潜在意味クラスと間の条件付確率の高い順に上位 20 個を示した．

4.1. 内容がかなり一致しているクラス

表 6.1，6.2 に示すクラスは，日本語，中国語ともに，クラスに含まれる上位の名詞群，動詞群の意味内容から，感情関連クラスと考えることができる．この感情関連クラスの名詞群では，日本語側に「憂さ」，「競争心」，「不安感」，「危機感」，「劣等感」，「不信」等，競争社会で内面に抑圧されたストレスを象徴するような単語が多く見られる．一方，中国語では「好奇心」，「恐怖」，「欲望」，「虚栄心」，「激怒」，「恨み」など，直接的な感情を表す単語が多く，それぞれ興味深い．

さらに，このクラスの動詞群を見ると，日本人と中国人では各々の感情への反応が異なっていることがわかる．

表 6.1 中国語における「名詞（目的語）-動詞（述語）」：「感情」関連クラス

名詞 日本語訳	名詞	動詞 日本語訳	動詞
趣味	兴趣	抱える	怀着
反響	反响	断念する	打消
好奇心	好奇心	喚起する	激起
気兼ね	顾忌	感じる	感
恐怖	惧色	心から感じる	发自
寒さ	寒意	抱える	怀有
下心	鬼胎	述べ表す	抒发
疑問	疑虑	我慢できない	按捺不住
偉大な波	波澜	生じる	萌生
偉大な乱れ	轩然大波	災いを招く	惹祸
情欲	情欲	作る	产生
気分	情绪	発散する	宣泄
虚栄心	虚荣心	心に抱く	心怀
内心	内心	深める	加深
好感	好感	拘束する	克制
警戒心	戒心	引き起こす	引起
波紋	涟漪	湧かす	泛起
疲れ	倦意	芽生させる	萌发
公憤	公愤	緩和する	缓和
恨み	芥蒂	存在する	存有

表 6.2 日本語における「名詞（目的語）-動詞（述語）」：「感情」関連クラス

名詞	動詞
大志	募らせる
鬱憤	抱かせる
憂さ	そそられる
不審	抱き始める
疑念	もたれる
競争心	持たれる
もやもや	抱く
不安感	つのらせる
恋心	払しょくする
食欲	そそる
しわ	いだく
危機感	いだかせる
無念	募らす
妄想	ぶちまける
ギャザー	持たれかねる
劣等感	ぬぐい切れる
いらいら	ぬぐいきれる
不信	払拭する
憎悪	ぬぐえる
悪心	ふっしょくする

日本語の動詞クラスには「払拭する」，「ぬぐい切れる」，「拭える」など，感情を拭い去ることを表す単語が多く存在しているが，中国語にはそのような意味の単語がほとんどみつけられない．

逆に，中国語の動詞クラスには「述べ表す」，「我慢出来ない」，「生じる」，「湧く」など，名詞の場合と同様，感情をそのまま直接的に表現し，外へ押し出すことを表す単語が多く存在している．

その差異の原因として以下のように考えることもできる．

歴史的に，日本においては，いわゆる「武士道」という言葉で象徴されるように，日常生活においてすら，安易に不平不満を並べ立てない克己の精神を訓練することが行われてきた．

すなわち，自己の悲しみ，苦しみを外面に表して他人に無用の心配をかけることがないように教育され，感情を顔に表すことは男らしくないと考えられてきた．立派な人物を評するとき，「喜怒を色に表さず」という言葉がよく用いられた[15]．つまり，長い年月にわたる克己の訓練や教育を通じて，日本人は常に感情を抑える傾向があるのではないかと考えられる．

一方，中国においては，古典的文献の代表としての唐詩，四書五経など，多くの作品では，著者の思考と感情を文学的に表現している．中国人は感情を抑えることではなく，むしろ含蓄のあるかつ自然な感情表現を美しいと考えている．現代の中国人はそれを受け継いできたが，中国社会の急激な変化に応じて，自らの意見や気持ちをある程度，直接的に伝えることも求められていると言える．このような両国の歴史や社会的評価の違いが，感情を表現する言語知識構造の差異として表れていることは非常に興味深い．

表 7.1，7.2 に示すクラスは，所属する各単語の意味から，「金銭関連クラス」と名付けることができる．名詞のほうでは日本語も中国語も概ね金銭に関する単語が見られる．ただし，中国語には「許可書」，「証明書」，「面積」，「管理手数料」，「設備」，「手数料」など特に不動産に関連すると考えられる単語も多く含まれている．

現在，中国においては，これらの不動産に関係する事例は各種の資金や費用に深く関わっていて，金銭と同じくらい重要と考えられているのかもしれない．

表 7.1 中国語における「名詞（目的語）-動詞（述語）」：「金銭」関連クラス

名詞日本語訳	名詞	動詞日本語訳	動詞
経費	开支	合格する	考上
許可書	许可证	集める	筹集
面積	面积	融資する	筹措
要素	要素	受け取る	收取
ローン	贷款	納める	缴纳
金額	金额	節約する	节约
資金	资金	支払う	交纳
費用	费用	交換する	兑换
債券	债券	払う	支付
外国投資	外资	占用する	占用
証明書	证件	年産する	年产
国債	国债	自分で集める	自筹
物資	物资	売る	抛售
支出	支出	使用する	使用
管理手数料	管理费	圧縮する	压缩
会費	会费	取り消す	吊销
税金	税款	不正流用する	挪用
設備	设备	輸入する	进口
ワクチン	疫苗	換算する	折合
手数料	手续费	採用する	考取

表 7.2 日本語における「名詞（目的語）-動詞（述語）」：「金銭」関連クラス

名詞	動詞
私財	取り崩す
日銭	脅し取る
私費	つぎ込む
巨費	使い込む
大枚	還流させる
白票	融通する
公金	用立てる
資金	貸し付ける
小金	全額保護する
積立金	脅し取られる
国費	拠出する
巨額	自己調達する
利ざや	借り入れる
外貨	融資する
身の毛	無心する
公費	投入される
預貯金	引き出される
預金	つぎこむ
原資	儲ける
印税	せびる

動詞に関しては，「融資する」，「自己調達する」等，基本的な用語に両言語での一致した表現が見られる．さらに，中国語で「不正流用する」，日本語で「使い込む」，「脅し取る」，「脅し取られる」と表現されているような，金銭に関する不正行為が，現在，中国においても日本においても問題視されていることがわかる．

表 8.1，8.2 は，このクラスに所属する単語の意味から，「政府機関関連クラス」と名付けることができる．これらの単語を見ると，名詞のほうでは中国と日本の各々，独特な政府機関の名前が示され，動詞のほうも対応していると分かる．

中国語には「三令五申」という，中国独特の熟語表現が含まれている．この動詞表現がこの「政府機関関連クラス」の上位に含まれていることは，政府機関の命令が再三繰り返されている現在の中国の社会状況を表していると考えることもできる．

表 8.1 中国語における「名詞（主語）-動詞（述語）」：「政府機関」関連クラス

名詞 日本語訳	名詞	動詞 日本語訳	動詞
政府	政府	登場する	出台
人民政府	人民政府	発行する	下发
明確な命令	明令	直接属する	直属
省庁	部委	公布する	颁布
地方政府	省政府	承認する	批准
国務院	国务院	投稿する	发文
市政府	市政府	策定する	制定
人事部	人事部	統計する	统计
中央政府	中央政府	明示的に規定する	明文规定
建設部	建设部	発表する	发布
商工業局	工商局	記録に載せる	备案
部門	部门	何度も繰り返し命令する	三令五申
上級機関	上级	受け入れる	采纳
当局	当局	実行する	执行
中央軍事委員会	中央军委	抗議する	抗诉
予算	预算	コメントを返す	批复
科学技術委員会	科委	命令する	责令
教育委員会	教委	授ける	授予
文化省	文化部	授与する	颁发
市党委員会	市委	有効になる	生效

表 9.1，9.2 のクラスは，このクラスに所属する単語の意味から，「経済関連クラス」と名付けることができる．動詞に関して，中国語では，経済的上昇を示す用語が多く見られる一方，日本語では対照的に，経済の下降を示す用語が多い．このことは，今回用いた両国の言語データ，特に新聞データに対応する時期の両国の対照的な経済状況を反映しているようで興味深い．

表 8.2 日本語における「名詞（主語）-動詞（述語）」：「政府機関」関連クラス

名詞	動詞
統計局	答申する
総務庁	まとめる
内閣府	建議する
総理府	主管する
経済企画庁	特別手配する
国土庁	全国調査する
労働省	発表する
総務省	最終答申する
アットホーム	追加指定する
国税庁	抽出調査する
農林水産省	研究委託する
気象庁	内示する
資源エネルギー庁	ヒアリングする
財務省	丸抱えさせる
通商産業省	中間発表する
中小企業庁	改訂する
リサーチ	毎月公表する
審議会	公表する
北海道開発庁	中間答申する
警察庁	調整保管する

表 9.1 中国語における「名詞（主語）-動詞（述語）」：「経済」関連クラス

名詞日本語訳	名詞	動詞日本語訳	動詞
総額	总额	以下になる	低于
総量	总量	少しづつ増える	递增
純利益	纯收入	超える	超过
売り上げ	销售额	高齢化になる	老龄化
総生産値	生产总值	激減する	锐减
金額	金额	激増する	激增
増幅	增幅	下がる	下降
貿易額	贸易额	増大する	增大
一人当たりの収入	人均收入	少なくなる	少于
カバー率	覆盖率	激増する	猛增
総面積	总面积	激増する	剧增
値上がりの幅	涨幅	増加した	增至
エリア	面积	増加する	增加
生産量	产量	到達する	达到
生産値	产值	目標を超える	超标
合計	总数	減少する	减少
客の量	客流量	到達する	达
成長率	增长率	上昇する	上升
入学率	入学率	少しづつ減る	递减
割合	比重	下げる	降低

名詞のほうでは，日本語には株や債券，相場等，個人資産に関する単語が多く見られるが，中国語には総生産値，貿易額，成長率等，むしろ社会全体の経済統計を示す単語のほうが多く含まれている．

9.2 日本語における「名詞（主語）-動詞（述語）」：「経済」関連クラス

名詞	動詞
先物	買い戻される
ペソ	急落する
渡し	暴落する
債券	下落する
株	続伸する
マルク	買い進まれる
銘柄	大幅下落する
円	続落する
ドル	反落する
株価	売り込まれる
労賃	高騰する
相場	買い進められる
平均	大幅続落する
ウォン	乱高下する
ポンド	大幅上昇する
ダウ	急騰する
国債	額面割れする
バーツ	騰貴する
世尊	反騰する
リラ	売り浴びせられる

4.2. 一見似たようなクラスであるが，内容が異なっているクラス

表 10.1，10.2 のクラスでは，名詞に関して，日本語と中国語ともに，両言語で共通した漢字「権」を含む単語が多く含まれている．しかし，実際に現れた単語の意味は両国語できわめて対照的に異なっている．

日本語では明らかに，個々人の「権利」を示す表現が多く，中国語では政府や国家の「権力」を意味する単語が多く含まれている．

さらに，動詞に関しても，日本語では「権利」の保証や付与を示す用語が多いが，中国語では「権力」を主体とした，土木，建築工事の遂行を意味する用語が多く見られる．

このような，両国の言語意味構造に現れた，同じ漢字「権」を含む単語表現の意味の違いは，やはり現在の両国の社会状況の違いを色濃く反映していると考えることもできる．

表 11.1，11.2 のクラスは，このクラスに所属する単語の意味から，「古代文化クラス」と名付けることができる．中国語では，自国，中国自身の古代皇族を表す単語が多く見られる．

表 10.1 中国語における「名詞（主語）-動詞（述語）」:「権力」と「権利」関連クラス

名詞 日本語訳	名詞	動詞 日本語訳	動詞
主要な権力	大权	品評する	品
リスク	险	握る	在握
主導権	主动权	他人の手に落ちる	旁落
突破性	突破性	高まる	昂
タスク	任务	不意になる	泡汤
著作権	版权	思い上がる	冲昏头脑
王位	皇位	工事を終える	完工
軍に対する指導権	军权	完成する	完成
王位	王位	竣工する	竣工
文化的な遺物	文物	工事を始める	开工
財産権	产权	工事を始める	动工
遺産	遗产	合理化する	合理化
軍隊の指揮権	兵权	貯水する	蓄水
主権	主权	当てが外れる	落空
5年計画	五年计划	実施する	付诸实施
プロジェクト	工程	渡す	移交
元利	本利	請負に出す	发包
利益	利益	空いている	空缺
農地	农田	終わる	告竣
決定権	决定权	保護する	保护

表 10.2 日本語における「名詞（主語）-動詞（述語）」:「権力」と「権利」関連クラス

名詞	動詞
基本権	与えられる
人権	移譲される
スト権	侵害される
立法権	保障される
老け役	行使される
プライバシー	委譲される
肖像権	付与される
主権	侵されかねる
私権	出願される
全権	奪われかねる
三権	全権移譲される
旧領	一部委譲される
栄誉	保障されとる
日照権	一部制限される
幣帛	侵害されかねる
西南	附与される
自治権	独占されかねる
一切合切	奉られる
行政権	侵される
自由	束縛される

表 11.1 中国語における「名詞（主語-動詞（述語）」:「古代文化」関連クラス

名詞日本語訳	名詞	動詞日本語訳	動詞
皇帝の乗り物	銮	想像する	试想
朝廷	朝廷	自ら政治を行う	亲政
大臣	臣	即位する	登基
漢代の皇帝の名前	高祖	ご光臨する	驾临
貴下	足下	帝王が自ら出征する	亲征
皇帝	天子	召喚する	召见
孔子，または高名な儒学者に対する呼称	夫子	戴冠する	加冕
戦国時代の楚国の王	楚王	視察に出る	出巡
紳士	君子	協定する	和议
国王陛下	陛下	退位する	退位
皇帝	君王	容赦する	开恩
古代の青銅器と石刻	金石	命令する	敕
皇帝	皇上	君主に謁見する	朝见
皇太后	皇太后	頓首する	顿首
皇帝	帝	訪問する	亲临
官吏が自分のことを卑下して言った言葉	卑职	厚く信頼する	倚重
漢の武帝	汉武帝	恩恵を与える	恩典
水に関することを支配する神	龙王	在位する	在位
皇帝	皇帝	よく考える	三思
世代または親族関係で目下の者	晚辈	国のために尽くしてその恩に報いる	报国

しかし，日本語には皇帝だけではなく，「儒家」，「隋」，「魏」，「ケルト人」，「ピューリタン」など，中国の古代文化を含む，外国の文化を表す単語が含まれている．

歴史的には，日本は中国から儒教，仏教，道教という，現在も日本の中心的な精神文化を担う思想や宗教を取り入れてきた．一方，明治以降は，日本はキリスト教を主体とする西洋の科学技術や思想，文化の影響のもとに，今日に至るまで発展してきた．

つまり，東洋文化と西洋文化を結びつけることで，現在の日本独特の文化が成立したと言える．この日本語のクラスの内容はそのような日本文化の混血的一面を示しているようにも見える．

表 11.2 日本語における「名詞（主語）-動詞（述語）」：「古代文化」関連クラス

名詞	動詞
隋	即位する
諸人	出御する
魏	行幸する
天子	崩御する
父祖	封する
公家	入植する
エスキモー	封ぜられる
ケルト人	造営する
衆生	退位する
国師	寺する
儒家	共立される
帝	収取する
修験者	遷都する
平家	来住する
僧正	読経する
ピューリタン	創建する
嫡流	配流される
僧尼	住する
忠実	滅ぼされる
明朝	混住する

表 12.1 中国語における「名詞（目的語-動詞（述語）」：「闘争」関連クラス

名詞日本語訳	名詞	動詞日本語訳	動詞
艦隊	舰队	率いる	率领
敵軍	敌军	敗滅させる	击溃
軍隊	大军	寄せ集める	调集
強力な軍隊	重兵	集結する	集结
選り抜きの強兵	精兵	阻止する	阻击
調査団	调查组	参加する	加入
分隊	分队	消滅する	歼灭
部隊	部队	増加派遣する	增派
軍隊	军队	派遣する	调遣
生きている戦闘力	有生力量	撃退する	击退
軍隊	人马	組み込まれる	编入
ソ連軍	苏军	移動する	调动
主力部隊	主力	派出する	派出
侵略軍	侵略军	攻撃する	歼击
軍隊の組織単位の一つ	纵队	強化援助する	增援
残党	残部	検閲する	检阅
軍隊の組織単位の一つ	中队	組み込む	收编
警察隊	警力	整え改編する	整编
警察	军警	分割する	分成
軍団	军团	滅ぼす	消灭

表 12.1，12.2 のクラスは，このクラスに所属する単語の意味から，「闘争クラス」と名付けることができる．ただし，日本語と中国語ではその闘争の意味内容がかなり異なっている．中国語では明らかに，軍隊を中心とした，実際の戦争に関する用語が多く見られる．

表 12.2 日本語における「名詞（目的語-動詞（述語）」：「闘争」関連クラス

名詞	動詞
出はな	繰り広げる
舌戦	展開する
熱戦	くじかれる
争奪戦	展開し始める
出鼻	くり広げる
力戦	武力鎮圧する
合戦	くじく
バトル	繰りひろげる
持論	展開し続ける
論旨	武力弾圧する
デッドヒート	客演指揮する
銃撃戦	展開させる
前哨戦	激化させる
ゲリラ戦	過熱させる
攻防	くりひろげる
選挙戦	鎮圧する
激戦	展開し出す
カーチェイス	余り続ける
白兵戦	封殺する
神経戦	かい摘む

一方，日本語では実際の戦争を意味する用語の順位は低く，上位には，むしろ，「舌戦」や「持論」を「繰り広げる」，「論旨」を「展開する」といった，言語に基づく論争を意味する表現が多い．あるいは，日本語には，「熱戦」や「争奪戦」などスポーツやゲームでの対戦を意味する用語も含まれている．このような，同じ「闘争クラス」における両国語での内容の相違も，やはり両国の現在の社会情勢の違いを「闘争」という側面から，端的に映し出していると言える．

表 13.1，13.2 にあるように，日本語では明らかに，具体的な各学問分野を含む，学問関連クラスが存在するが，中国語にはそのような，包括的な学問関連クラスは見られない．表 13.1 の中国語のクラスでは，一応，「法医学」，「社会学」，「天文学」，「社会経済学」などの個々の学問分野が含まれている．しかし，クラス全体としては単なる「学問」だけではなく，より広く「学科」や「理論」の意味を含んでいると考えるほうが適切であろう．中国では新聞等で各学問が単独で話題になることは比較的少ないのかもしれない．ちなみに現在，「鄧小平理論」は中国では独自の理論分野として考え

られている．

表 13.1 中国語における「名詞（主語-動詞（述語）」：「理論」と「学問」関連クラス

名詞日本語訳	名詞	動詞日本語訳	動詞
法医	法医	鑑定する	鉴定
興奮剤	兴奋剂	臆断する	臆断
現場	实地	検証する	检验
考古学	考古	検出する	检测
暗箱	暗箱	掘削する	发掘
静電気	静电	ガイドする	指引
特集	专题	指導する	指导
実行性	可行性	自動化する	自动化
鄧小平理論	邓小平理论	テストする	试验
婦人科	妇科	研究する	研究
漢学	汉学	検討する	研讨
顕微鏡	显微镜	クローンする	克隆
電子	电子	操作する	操作
社会学	社会学	測定する	测定
実証	实证	描く	绘制
法医学	法医学	検眼する	验光
科学	科学	干渉する	干扰
職位	职称	評定する	评定
天文学	天文	観測する	观测
ネットワーク	台网	アラームする	告警

表 13.2 日本語における「名詞（目的語-動詞（述語）」：「理論」と「学問」関連クラス

名詞	動詞
経営学	専攻する
心理学	学ぶ
政治学	修める
漢学	習得する
中国語	伝授する
動物学	学び始める
蘭学	教え始める
建築学	伝授される
法学	勉強する
英学	マスターする
オランダ語	学べる
読み書き	習う
工学	教わる
日本語	習得させる
地質学	独学する
薬学	師事する
ロシア語	独習する
物理学	学ばせる
声楽	教授する
生物学	大成する

以上のように，今回用いた確率的言語知識構造に基づく日本語と中国語の比較では，各々の言語が属する国の，社会的，文化的，歴史的特徴が相対的に映し出され，サピア・ウォーフ[16]的な意味合いでも興味深い．また，これらの結果は，この方法を用いた今後の多国語，多国間比較研究に興味深い課題を提供していると言える．

５．今後の課題

本研究の分析においては，日本語で 400 クラス，中国で 200 クラスが抽出されている．本論文では，両言語におけるこれらの多数のクラスから，特に分析結果の解釈に基づき，中国語と日本語で同じネーミングになったクラスを選んで，①内容がかなり一致しているクラス，②似たようなクラスであるが，内容が異なるクラス，という二つのグループに分類し，日本と中国の文化や社会システムを比較考察した．しかし，このようなクラスの分類は必ずしも客観的な基準に基づく厳密な分類ではない．今後，客観的な方法に基づく両言語でのクラスの一致度を測定する方法を考える必要があると言える．

一方，今後，今回取り上げなかったクラス群の考察はもちろん，両言語とも分析対象としての言語データベースを量的にも，質的にも拡張していく必要がある．特に日本語にあるクラスが中国語では存在しない場合を考慮すると，両言語での十分な比較考察のためには，少なくとも中国語のデータベースを拡張し，そのクラス数 200 を日本語でのクラス数 400 まで増やす必要がある．

さらに，本研究の方法を用いて，日本語，中国語に限らず，より多くの他言語，たとえば英語の大規模データベースを用いて，同様の確率的言語知識構造を構成し，多角的な比較研究を進めていくことも今後の重要な課題の一つである．

６．参考文献

[1] 市村由美・長谷川隆明・渡部勇・佐藤光弘.テキストマイニング— 事例紹介—. 『人工知能学会誌』, 2001, 16 (2), 192–200.
[2] Pereira, F., Tishby, N., & Lee, L.. Distributional Clustering of English Words. The Proceedings of 31st Meeting of the Association for Computational Linguistics, 1993, 183–190.
[3] Hofmann, T.. Probabilistic latent semantic indexing, Proceedings of the 22nd Annual ADM Conference on Research and Development in Information Retrieval 1999, 50–57.
[4] 持橋大地・松本裕治.意味の確率的表現.『情報処理学会研究報告』, 自然言語処理研究会, 2002-NL-147, 77–84.
[5] Hagiwara, M., Ogawa, Y., Toyama, K..PLSI Utilization for Automatic Thesaurus Construction. Lecture Notes in Computer Science, 2005, 334–345.
[6] Kameya, Y., & Sato, T. Computation of probabilistic relationship between concepts and their attributes using a statistical analysis of Japanese corpora. Proceedings of Symposium on Large-scale Knowledge Resources, 2005, 65-68.
[7] 阿部慶賀, 中川正宣. 言語統計解析を用いた確率的言語知識の構築とその心理学的妥当性の検証. 認知科学. 2007, 14(1), 91-117.
[8] Asuka Terai, Masanori Nakagawa. A meural network model of metaphor understanding with dynamic interaction based on a statistical language analysis: targeting a human-like model. International Journal of Neural Systems, 2007, Vol. 17, No. 4, 265–274.
[9] Asuka Terai, Masanori Nakagawa. A corpus-based computational model of metaphor understanding consisting of two processes. Cognitive Systems Research. 2012.
[10] Kayo Sakamoto, Asuka Terai, Masanori Nakagawa. Computational models of inductive reasoning using a statistical analysis of a Japanese corpus. Cognitive Systems Research, 2007, 8, 282-299.
[11] 太蘭. 中国語における確率的言語知識構造の構築—動詞と名詞の関係を中心として—. 2011, 東京工業大学, 修士論文.
[12] 情報通信研究機構. 中国語解析技術と機械翻訳の評価. http://www.congre.co.jp/imttsympo/2009/program/pdf/project02.pdf.
[13] Taku Kudo, Yuji Matsumoto (2002). Japanese Dependency Analysis using Cascaded Chunking, CONLL 2002 in TAIPEI.
[14] Bor Hodoscek, 阿辺川武, Andrej Bekes, 仁科喜久子 (2011). レポート作成のための共起表現算出支援— 作文支援ツール「なつめ」の使用効果—. 専門日本語教育研究. 13. 33-40.
[15] 新渡戸稲造 (著), 奈良本辰也 (翻訳). 武士道. 三笠書房,1993.
[16] 大堀寿夫. 認知言語学. 東京大学出版会, 2002.

付録

以下に本研究で用いたKameya & Sato(2005)の手法における, EMアルゴリズムの詳細を記載する.

E (xpectation)-step:

$$P(c_h \mid n_i, a_j) = \frac{P(n_i \mid c_k)P(a_j \mid c_k)P(c_k)}{\sum_k P(n_i \mid c_k)P(a_j \mid c_k)P(c_k)},$$
$$E[c_h] = \sum_{i,j} F(n_i \mid a_j)P(c_h \mid n_i, a_j),$$
$$E[n_i \mid c_h] = \sum_j F(n_i \mid a_j)P(c_h \mid n_i, a_j),$$
$$E[a_j \mid c_h] = \sum_i F(n_i \mid a_j)P(c_h \mid n_i, a_j),$$

M (axmization)-step:

$$P(c_h) = \frac{E[c_h] + \alpha}{V^t + \alpha V^c}$$
$$P(n_i \mid c_h) = \frac{E[n_i \mid c_h] + \alpha}{E[c_h] + \alpha V^n}$$
$$P(a_j \mid c_h) = \frac{E[a_i \mid c_h] + \alpha}{E[c_h] + \alpha V^a}$$

$F(n_i, a_j)$は名詞と形容詞もしくは動詞の共起頻度数, V^tは全共起頻度数で$V^t = \sum_i \sum_j F(n_i, a_j)$, V^cは潜在クラス数, V^nは名詞数, V^aは形容詞数もしくは動詞数, αは平滑化パラメータで, この方法に基づきスパースネス問題($F(n_i, a_j)$=0 の場合)が解消される.

EM アルゴリズムにおいては, 係り受け共起

頻度データD が与えられたときの，パラメータ θ （ここでは上記Kameya & Sato のモデルにおける分布パラメータ$P(n_i|c_h), P(a_j|c_h), P(c_h)$の総称として$\theta$ とする）の事後確率P(θ |D) を考える．ここでは式(4) により，尤度P(D| θ) の代わりに事後確率P(θ |D) の最大化を行う．なお，事前分布P(θ) にはこのような場合，一般的に用いられるDirichlet 分布を仮定する(Blei, Ng & Jordan, 2003; Buntine& Jakulin, 2004).

$$P(\theta|D) = \frac{P(D|\theta)P(\theta)}{\int P(D|\theta)P(\theta)d\theta} \propto P(D|\theta)P(\theta)$$

Abstract

The purpose of the present study is to compare the cultures and social systems behind languages, based on probabilistic knowledge structures constructed from large scale language data in Japanese and Chinese. The present study applies the probabilistic algorithm developed by Kameya & Sato, more strict than the existing method like LSA (Latent Semantic Analysis).

Practically, we compares interesting classes which have same names between Japanese and Chinese as the result of interpretation, selected from the latent classes estimated using the above algorithm. It is an important future plan to construct the same knowledge structures using other foreign languages like English.

（受付日：2014 年 7 月 11 日，受理日：2014 年 10 月 22 日）

張　寓杰（チョウ　ユージェー）
現職：東京工業大学大学院社会理工学研究科人間行動システム専攻
博士課程在学
首都大学東京大学教育センター　特任助教

2010 年首都師範大学(中国)教育研究科教育心理学専攻修士課程修了.
2011 年から東京工業大学大学院社会理工学研究科人間行動システム専攻博士課程在学.
現在は日本語と中国語の言語統計解析に基づく計算モデルを用いて，日本語と中国語における帰納的推論の比較研究を行っている.

主な論文：張寓杰，寺井あすか，董媛，王月，中川正宣(2013). 日本語と中国語における帰納的推論の比較研究―言語統計解析に基づく計算モデルを用いて―. 「認知科学」，20(4), 439-469.