言語コーパス利用の中国語電子補語辞典編纂とその課題

砂岡和子（ 早稲田大学 ）、詹衛東 （北京大学）

於関東支部拡大例会 2007 年 3 月 17 日

［概要］

中国語動詞補語はその定義や用法自体が難しいばかりか、補語の使用条件について理屈にあった説明をするテキストが少なく、習得のネックとなっている。本報告では、北京大学汉语语言学研究中心(CCL)と同计算语言所(ICL)の言語コーパスを利用して現在構築中の、中国語動詞補語電子辞典の編集と、電子辞書の記述作業で明らかになった動詞補語の語義特性について報告する。

本電子辞書は結果補語を主に、一部の方向補語、可能補語、程度補語について、生の言語素材から学習者用にリライトした例文を収録し、実例に即し補語の語義特性を記述する。述部動詞（形容詞を含む）は HSK、既刊中国語補語辞典、ICL および CCL の頻度情報を参照し、述部と補語の組み合わせは作業者の語感とコーパスやインターネット言語資源を参照して選定した。個々の出現頻度情報の提示と補語の語義特性の記述に加え、検索の自在さは、従来の補語辞典に比べ、中国語動詞補語を活用するさいの有力な学習情報になると期待できる。

［キーワード］

中国語補語辞典、言語データ資源，総合型言語知識データベース

１　コーパスベースの辞書編纂メリット

コーパスとは言語研究に使用されることを前提に、実際に書かれたり話されたりする文の集合体を指す。コーパス言語学（Corpus Linguistics）は言語データの利用に不可欠な情報処理技術を基盤とし、緻密な事実観察による言語の実証的研究を行い、妥当な言語理論の構築および研究手法の確立を目指すものである。コーパス言語学は従来の言語研究に比べ、明示的で具体性に富み、電子化された膨大なデータを研究基盤とするため、科学技術への応用が迅速である。高速コンピューターやインターネットの普及に伴い、入手および記録可能なデータ量が飛躍的に増大し、コーパス言語学は言語研究に欠かせない一分野となっている。

一方、コーパスは自然言語処理研究のデータバンクとして需要が増し、コーパス建築とその言語研究への応用研究が活況を呈している。自然言語処理とは人間の言語運用プログラムを計算機上で実現することを目的に、インターネットなど大量の情報処理、機械翻訳、対話システム、要約など実用性の高い言語辞書の構築を行う技術である。

言語の理解とは何かをコンピューターで具体性に記述・提示する必要から、膨大かつ精度の高い実例データを集め、統語解析(syntactic analysis)や形態素解析(morphological analysis）を行い、出現頻度情報に基づき文全体に対して「もっともありそうな解釈」を求める構文解析研究がその最大の特色である。大量言語データ分析に基づいた網羅的かつ一貫性のある規則の抽出や、統計情報を用いた曖昧性解消の研究から、現在はコーパスを使った語義の獲得が自然言語処理研究の主流となっている。

一般にコーパス言語学は従来の言語研究にくらべ、以下の特徴を持つ。

A. 言語の普遍的特徴の解明よりも、個別言語の言語記述に中心を置く。
B. 言語研究における合理主義的立場より、経験主義的立場に中心を置く。
C. 質的な言語モデルのみならず、数量的な言語モデルも分析基準とする。
D. 言語能力よりも言語運用能力に中心を置く。

辞書編集へのコーパスの応用はその典型であろう。初めて本格的にコーパスを活用した

Collins COBUILD English Language Dictionary (Collins ELT, 1987) が発刊されて以来、英米の辞書や参考書はコーパスに基づく編纂が主流となり、コーパス辞書学という新しい研究分野も誕生した。その潮流は他の言語にも波及し、2002 年には日本でも英和・和英辞書『英辞郎』、『ジーニアス英和大辞典』、『ウィズダム英和辞典』など英語コーパス言語学の成果を取り入れた辞書の出版が相次いでいる。外国語辞書編纂がコーパス情報を利用するメリットはコーパス言語学と理念は同様であるが、学習者にとっての具体的なメリットは以下の点であろう。

1. コーパス解析結果に基づき高頻度の語義や用例を使用頻度順に配列が可能。
2. コーパスを精査することにより母語話者さえ気づかない言語事実の発見と、実態に即した詳細な語法分析結果の記述が可能。
3. 日常生活語彙を収集したコーパスソースに依拠し、従来の辞書から漏れがちなオーラルコミュニケーション対応型の語義や用法の充実が可能。
4. 従来のような○×式の単純な正誤情報にとどまらず、よりきめ細かな使用域を表示することで、平均的語構成規則とともに、異体的規則を示すことにより、言語直観の働かない外国人学習者が状況に応じた学習ができる。

コミュニケーション重視の外国語学習では初級段階から生きた言語との対面が避けられない。学習者は辞書にない語義や用法、新語や新語義、地域特有の表現や個人差による音声のバリエーションなど、従来の学校文法では教わらない言語現象にぶつかる。コーパスはこうした局面で母語話者に替わり、言語事実の適合性や語感の確認に辛抱強く応えてくれる。現在、Web を使いこなせる層は、インターネット検索や電子辞書を駆使して翻訳や閲読を行っている。今後、電子型辞書が多言語、百科知識の提供、自動音声などの機能を拡充すれば、この傾向はいっそう強まると予想される。中国語動詞補語辞典の編纂もまずは編集用コーパスを構築することから出発した。

２　語彙知識の獲得と記述の方法

語彙知識の提供を主目的とする辞書にとり、語彙知識の獲得と記述が重要であることは言を待たない。では言葉の意味はどのような枠組みで分析・記述できるのか？言語学では、実データから語の意味分析を行う記述的研究や、言語理論ベースの生成語彙意味論、語彙概念構造、フレーム意味論など、語彙意味論研究が多様性を見せながらめざましく発展している。

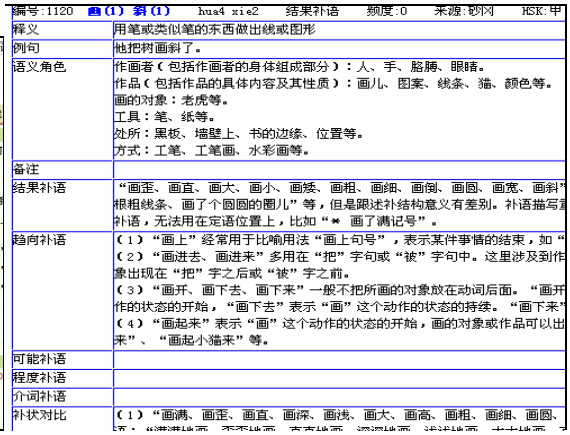
一方、自然言語処理(natural language processing)とよばれる応用研究の領域でも、従来の統語分析からより深いテキストの意味の獲得に向かっている。自然言語処理分野では多義解消が大きな課題で、ことに中国語のように有標の語法規則が非顕在的な言語の解釈には形式文法による分析だけでは不足である。個々の語の意味・形態・機能に焦点をあてた記述が多義解消に有効となる。

しかし語義を詳細に規定しようとするほど、語義の範疇の確定と意味体系をめぐる意見の対立を引き起こす。従来、辞書の意味の記述には、同義語、類語、反義語、あるいは上下位の概念からその言葉の属性を規定する方法が多く用いられる。これら語義は相互に巨大なネットワークを張り巡しているはずで、統合的な言語知識に依拠しない勝手な属性の定義は語彙記述間の矛盾となって現れる。近年、WordNet のように高度に形式化した概念構造に依拠した大規模同義語集の実現によって、語彙フレームワークの規定に基づいた語義コーパスの構築が試みられるようになり、これら実データに基づいた語彙のさまざまな

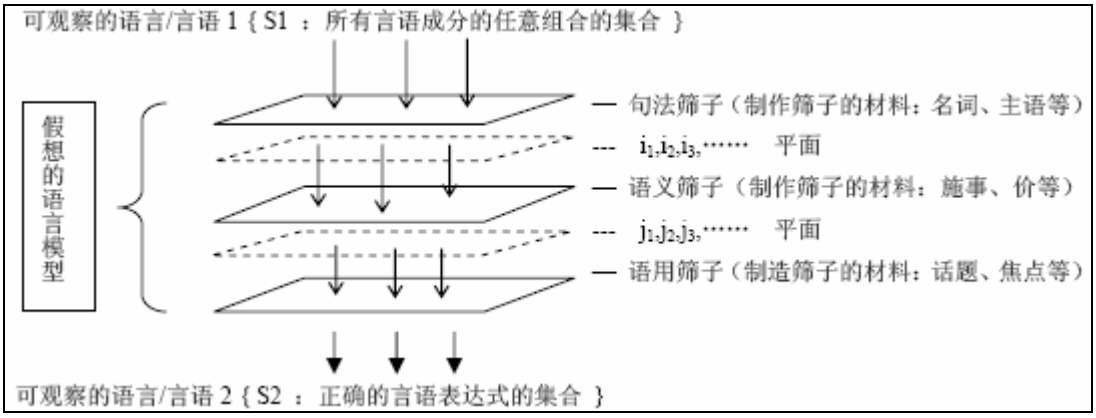
語法・語彙属性についての知識を統合、抽出、ルール化、形式化して知識データベースに組み込みんだシステムが、検索、機械翻訳、段落ダイジェストなどに威力を発揮している。

本発表の中国語電子補語辞典も北京大の現代中国語総合言語コーパス構築の一環であり（図 1）、随意な補語の意味羅列にとどまらず、外国人学習者向けの実用主義にのっとり、統計に基づいた語彙選定と、語彙属性の形式化をモットーに編集に当たっている。使用環境が規定されてこそ意味は現実のものとなり、語法的形式化によって文意解析と応用が可能との理念による。語義の解析は句法(統語規則)ならびに語用と不可分の関係にあり、用途に応じたフレームワークを使って言語事実を濾過し、相互補完的に分析観察することが望ましい（図 2）。

（図 1）



（図 4）



（図 2）詹卫东[确立语义范畴的原则及语义范畴的相对性]より

中国語動詞補語辞典は、動詞補語の語義の記述と同時に、補語の述部に対する具体的振る舞い（語義特性）を形式化することを目指している。

3　中国語補語辞典編纂の経緯

本中国語電子補語辞典（以後 DCVC〔Dictionary of Chinese Verb Complement〕と略）の原型は、2006 年に仮構築した Web 版中国語補語 CAI 教材である(以後砂岡旧辞典と略)。砂岡旧辞典は既刊の数種の書籍版中国語補語辞典の例文を参照し、見出し語約 2500 語、例文約 3 万条を電子版用にリライト後、日訳を付与し、さらに北京大学「現代漢語短語語法信息詞典」の解析データに従って、中国語動詞補語句の統語規則を書き込み、各種検索ツールをつけて、インターネット上で公開を試みた。しかし動詞補語の選定基準が不明瞭で、機械翻訳向け語法属性の転記だけでは学習者の認識・産出を支援できる内容に至らな

かった。2006 年 6 月より北京大学漢語語言学研究中心[Center for Chinese Linguistic PKU(以後 CCL と略)]と、同計算語言学研究所(ICL)の協力を得、それぞれの言語コーパスを利用し、統計的データに基づいて補語句を選別しなおし、その上で新たに語義の記述を開始した。以下、DCVC の概要を示す。

1　対象とする補語句の範疇：以下の（A）（B）を核にそれ以外の補語句は選別。

（A）述語（動詞、形容詞）＋結果補語

（B）述語（動詞、形容詞）＋方向補語

一般の可能補語は上記 AB の拡張形式とみなし、巻頭で派生規則を説明するに留める。ただし“V＋得”“V＋得了”“V＋不了”など、上記 AB で対応関係にない可能補語は収録する。時量・動量・介詞補語は個別記述の対象外とし、同様に程度補語も全般的説明に限定する。

2　補語句選別基準と補語句選別作業過程：

外国人学習者の便宜を考え、述語は HSK 所収の動詞と形容詞を優先して収録した。ただし HSK 甲乙丙丁に含まれる動詞 2900 語と形容詞約 1117 語、計約 4000 語がすべて補語句を取るとは限らず、また HSK 語彙が実際の中国語の使用状況を反映しているとは限らない。そこで ICL 各種言語データで動詞と形容詞の出現頻度、およびそれぞれの補語との組み合わせ頻度を調査した。たとえば信息詞典（約 7 万 6 千語収録）中の v+v、v+a、v+p 各補語句の出現数を調査したところ、大半が方向補語動詞で計 6282 回（同形語を含む；以下同様)、場所補語が 1172 回でこれに次ぎ、結果補語は 722 回とさほど多くない。同様に ICL コーパスを利用して、補語を取る形容詞について頻度情報とともに抽出作業を行った。

補語になる形容詞の選別はデータ不足であったが、CCL コーパスから手作業で検出し、CCL にない場合は母語話者の語感と Baidu や Google を検索し抽出した（図 3）。

word	comp	comp_type	freq_CCL	freq_Baidu	freq_Google	pos	hsk_level
单调						形容词	乙
单调	得很	程度	2		537,000		
单调	死了	程度	0	915	409,000		
单调	多了	程度	0	1910	1,630,000		
单调	得多	程度	0	1980	1,620,000		
单调	下去	方向	1	947	378,000		
单调	起来	方向	0	8740	422,000		
单调	下来	方向	0	804	434,000		

（図 3）述補語彙リスト選別作業“单调”（部分）

3　DCVC 編集ツールの構成

本辞書編集専用の作業ツールを開発した（図４）。作業プラットフォームは以下の①②③の 3 層から構成され、作業者が分担して記述を行う。

①述語と補語の基本情報（通し番号、ピンイン、語形、来源、頻度、語義特性など）

②補語の用法に関する情報（結果、方向、可能、程度、介詞の各補語の用法特色、補語と修飾語の対比など）

③当該補語句に関する情報（語彙説明、例文、注釈など）

このうち①の“语义角色（semantic role）”が編集の力点である。語義特性“语义角色”とは、述部の基本情報のことで、たとえば“吃”の語義特性は“吃”の前後に現れる名詞成分が担う役割について、次のように描写することができる。

摂食者の語義特性は[+人][+动物]、被摂食物の語義特性は[+固形食物][+一部飲料][+薬]さらに“吃”と共起する名詞成分には“用筷子吃”“用大碗吃”のように、食事道具や容器があるため、「道具」「容器」の語義特性を立てる。中国語の補語句は、(a)事態発生後の自然の帰結、もしくは(b)結果に対する話者の評価の描写に特色がある。述部の基本的語義特性は、事態発生前後の変化や方向性を判別する有効な情報となる。そのため語義特性の記述には“施事”“受事”などの述語を用いず、できるだけ“动作发出者”“可以吃的东西”のような具体的な用語を使う。さらに“吃”の語義の範疇(Semantic Category)を[某人、动物] 用 [工具、容器]吃 [食物、药物]、[某人、动物] 在 [处所、时间] 吃 [食物、药物]と形式化できると、学習者にとって明示的な語法ルールを提示できる。

一方、“挨”のように動作者との相互依存関係にあるタイプの動詞は[某人、某具体物]靠近或接触 [某人、某具体物]のように動作参与者と被参与者に対する語義特性を記述することになる。

挨(1)

语义角色	动作参与者 A：人、汽车、桌子等。 动作参与者 B：人、窗台、水库、枕头等。 （动作参与者 A 与动作参与者 B 在动作行为中存在互相依赖的关系，A 靠近、接触 B，则 B 同时也靠近、接触 B。）

編集用検索機能

多数による編集作業分担と、情報の共有、および全体の品質管理には、編集作業用の検索ツールが不可欠である。DCVC 用に編集専用の多機能検察ツールを開発した（図５）。本ツールによって複雑な語形の検索が可能となったと同時に、相互に入力内容を閲覧することで、情方の共有化ができ、作業効率向上と品質安定に威力を発揮できる。

（図５）

DCVC	例解詞典	搭配詞典
挨(1)定	有	無
爱透	無	無
吃瘦	無	有
吃舒服	無	有
吃得过来	例文は“吃不过来”のみ	無
单调起来	無	無

（図６）収録語彙比較

４　編集作業状況

現在、筆者２名の監修のもと、北京大中文系院生７名が Web 上で編集原稿の作成に当たっており、草稿完成後に、電子辞書としての体裁を整え、学習者向けの日本語訳などを付加して、CDROM 付書籍版として出版する。出版と平行し、補語データベースの構築を行う。書籍版の記述は以下のような形式となろう。

辞典記述例　吃（１）　ピンイン配列順(暫定)

詞形　　ピンイン[1]　　補語の種類　　頻度[2]

吃（１）

语义角色	动作发出者：人、动物等。 可以吃的东西：一般是固体形态的食品、药品，比如水果、米饭等。 吃东西用的工具：比如筷子、刀、叉、勺子等。 吃东西用的容器：比如碗、盘子等。

吃不过来　chi1bu5guo4lai2　　可能补语　　频度:2

释义	无法参加（或完成）吃东西的活动。
例句[3]	他的朋友很多，都要请他吃饭，根本吃不过来。 这么多样的小吃，一天根本吃不过来。
备注	甲的朋友多，朋友经常请甲吃饭。甲无法参加所有的吃饭活动，就可以说“吃不过来”； 食品的数量或种类多，在一定的时间内无法都吃下去，也可以说“吃不过来”。 注意：“吃不过来”中的“过来”不是表示物体的位移，而是趋向动词的引申用法，用“过来”表示“完成”、“能做完某件事情”。

吃不了　chi1bu5liao3　　可能补语　　频度:131

吃不得　chi1bu5de5　　可能补语　　频度:131

释义	不能吃，不可以吃。
例句	很多独生子女吃不得苦。 有毒的山菇吃不得。 老年人牙齿都掉了，吃不得硬东西。

吃得　chi1de2　　可能补语　　频度:8

释义	能吃，可以吃。
例句	老人牙齿一颗没掉，肉也吃得，酒也喝得。 一片吃得，整个的自然也吃得。 四季豆必须炒熟才能吃得，否则会有毒。
备注	“吃得”相当于“能吃、可以吃”，其中的“得”读音为 de2，不能读为 de5。

吃饱　chi1bao3　　结果补语　　频度:494

释义	人或动物进食后不再感到饥饿。
例句	让所有的孩子都能吃饱饭。 吃饱后立即进行剧烈运动对身体不好。 老先生喜欢吃饱饭后到院子里散散步。
备注	“吃饱”中的补语“饱”指人或动物在吃东西后到达“饱”的状态。 “吃饱”的前面可以出现动作发出者，比如“他吃饱了”。

---

1 編集ツールでは声調は数字で示される。出版のさいは記号付声調に変換する。

2 CCLにない補語句はBaiduやGoogleを検索して立項した。ただしあくまで参考値であり、出版時の表示方法とは別である。

3 出版時は語釈や例文には日本語訳をつける予定。

	“吃饱”的后面一般不能再带名词性成分作宾语，比如，不能说“*他吃饱了这个苹果、*他吃饱了两碗饭”。不过，也有个别名词可以出现在“吃饱”后面，比如“他吃饱了饭”“你们要吃饱肚子才能干活”等，但仅限于“饭”“肚子”等名词，且只能是单个名词的形式。这两个名词有时候也可以由“把”引出，放在“吃饱”的前面，比如：“你们得让老百姓把肚子吃饱”。除此之外，“吃饱”很少用于“把”字句中。

（中略）

吃到 chi1dao4 结果补语 频度:302

释义	有条件吃。
例句	现在可以吃到新鲜蔬菜。 园艺学家花了多少年的功夫，终于让大家吃到了无籽西瓜。

吃到 chi1dao4 介词补语 频度:68

释义	吃的动作行为完成后的时间或处所。
例句	今天的午饭一直吃到下午两点。 这种菜外形不好看，吃到嘴里滑滑的。
备注	“吃到”后接时间词时表示吃的动作行为完成后的时间点，如“吃到下午两点”。“吃到”后接处所词时是说明所吃的食物等所处的地方，相当于“吃在”，如“这种菜外形不好看，吃到嘴里滑滑的。”也可以“说这种菜外形不好看，吃在嘴里滑滑的”。但“吃在”后可以接动作发出者的处所，而“吃到”不可以。

吃得过来 chi1de5guo4lai2 可能补语 频度:2

释义	可以参加（或完成）吃东西的活动。
例句	天津的小吃很多，一次怎能吃得过来？ 蚂蚁吃大象，虽然上了身，却是不可能吃得过来的。
备注	“吃得过来”中的“过来”不是表示物体的位移，而是趋向动词的引申用法，表示能完成某件事情。 “吃得过来”常用于反问句或否定句中。

（後略）

５　電子補語辞典の特徴と編纂課題

上述のように、DCVC はコーパスに依拠して統計的に語句を選別し、頻度情報を提示し、例文を実際の言語から採っている。そのため、一般の補語辞典に比べ、収録語彙の実用価値が高く、既存の補語辞典にはない表現を見出すことができる（図６）。また規則的語法ルールは総論でまとめて行い、個別の語義特性の記述に重点を置くため、個々の語用法を具体的に知ることができる（例、上記“吃得过来”と“吃不过来”、結果補語と介詞補語の“吃到”、それぞれの用法と用法の違いの記述を参照）。

語義の描写も随意な記述にとどまらず、語義範疇の弁別と形式化を目指している。生成ルールが明示されれば、外国人学習者が中国語の補語の生成ルールを獲得する手がかりとなろう（例、上記“吃不过来”や“吃饱”の注釈欄を参照）。

しかし語義特性の弁別、ことに形式化は、当該言語全体を見渡す力量が必要である。以下の“ 挨(1) 不得(1) ”と“挨(1) 不了(1) ”は用法の違いを描写しているが、形式化には達していない。現在の草稿をさらに校訂し、出版に向けてより明確な語義範疇の記述を思考してゆく。

挨(1) 不得(1)　ai1 bu4de2　可能补语

释义	不能够靠近、接触。
例句	这地方到处是珍贵瓷器，挨不得碰不得。 我的脚疼得厉害，几乎挨不得地。

挨(1) 不了(1)　ai1 bu4liao3　可能补语

释义	不能够靠近、接触。
例句	我的脚疼得挨不了地，想坐下休息一会。

可能补语	“挨不得”、“挨不了”都是表示动作发生的可能性，“挨不得”是从动作发生的结果角度来说的，“挨”动作发生之后可能产生某种不好的结果，所以在人的主观愿望中不希望动作发生，如“这些珍贵瓷器挨不得碰不得”，“挨”的动作发生后，可能会破坏“瓷器”，所以人们的主观愿望中不希望这个动作发生；“挨不了”则是从动作发生的条件角度来说的，因为存在某种条件，所以动作发生的可能性很小，如“我的脚疼得挨不了地，想坐下休息一会。”，因为“脚疼”，所以“挨”这个动作发生的可能性很小。

コーパスデータ資源の性質に起因する例文の偏りも課題のひとつである。大多数のコーパスは書き言葉が主体で、口語で常用される補語句が出てこなかったり、頻度情報に偏りが見られる。語彙と例文の選別には複数のコーパスを参照検索せざるを得ない。

参考文献

中国

詹卫东 2001 确立语义范畴的原则及语义范畴的相对性，《世界汉语教学》第 2 期

詹卫东 2003 一个汉语语义知识表达框架：广义配价模式《第 5 届全国计算语言学联合学术会议论文集》

詹卫东，汉语语义分类系统及语义关系描述基本框架（设计草案）
http://ccl.pku.edu.cn/doubtfire/semantics/973_Beida/index.htm

北京大学汉语语言学研究中心(CCL) コーパス　http://ccl.pku.edu.cn/default_E.asp

北京大学计算语言所コーパス(ICL)　ライセンス契約必要

俞士汶 2003、基于语言数据资源的知识挖掘与语言资源库的集成

俞士汶他 2005 北京大学『現代漢語語法信息詞典』

候精一等編著 2001『中国語補語例解』商務印書館

孟琮等 2003『漢語動詞用法詞典』商務印書館

王硯農等編 1987『漢語動詞―結果補語搭配詞典』北京語言学院出版社

缪锦安 1987《汉语的语义结构和补语形式》上海外语教育出版社

Zhu Hong, Yang Liu October,2006；“MCD;A Chinese-Korean-Japanese Multilingual Concept Dictionary”2nd International Symposium on Knowledge Processing and Service for China, Japan and Korea: METADATA and ONTOLOGY”, Beijing, China

尹明，砂岡和子，成田誠之助 2003「多国語 Windows 作業系統下基于 Corpus 的中国語教学課件的開発」The 3ndInternational Conference on Internet Chinese Education

日本

郡司隆男 2004『意味』岩波書店『言語の科学』巻４
斎藤俊雄他編 1998『英語コーパス言語学』大修館書店
㈱アルク 2002 英和・和英辞書『英辞郎』
大修館書店 2002『ジーニアス英和大辞典』
三省堂 2002『ウィズダム英和辞典』
张国宪 2006『中国語の補語』《补语的句位探索-关于非可控义》日中対照言語学会白帝社
砂岡和子、劉揚 、朱虹 2007［中韓日 Multilingual Concept Dictionary 構築の現況］
言語処理学会第 13 回年次大会(NLP2007)発表論文
砂岡・尹明 2003「コーパス利用による中国語教育（１） 中国語コーパス利用とデータ変換」
早稲田大学政治経済学部紀要『諸学教養』紀要 114 号