日中韓各語の
知的情報検索に於ける
辞書を用いた異表記処理

春遍雀來
日中韓辭典研究所
The CJK Dictionary Institute
〒352-0001　埼玉県新座市東北2-34-14
jack@cjk.org
COLING 2002発表

目次

要旨

日中韓各語の表記法の複雑さは、計算言語学処理用の各種ツール、特に**知的情報検索用**ツールの開発者にとって、大きな課題となっている。この難しさは、これら各語に標準的な正書法がないことで増幅され、特に表記のゆれが著しい日本語に於て顕著である。本稿では日中韓各語に於ける表記のゆれの類型論に焦点を合わせ、その言語学的な諸問題について手短に分析し、異表記処理に於て語彙データベースが中心的役割を果たす理由について論じる。

1　序論

日中韓各語に於ける情報検索の難しさには種々の要因がある。真に「知的」検索を実現するには、多くの課題を克服しなければならない。主な課題には以下のようなものがある：

１．標準的正書法の欠如。極めて多数の異表記体(特に日本語の)と文字種を処理するには、**表記法間検索**(Halpern 2000)のような高度な情報検索技術を備えていることが必要である。
２．中国語の簡体字(SC)と繁体字(TC)相互間の正確な変換。一見容易そうに見えるが、実は非常に難しい計算処理課題である(Halpern and Kerman 1999)。
３．日本語と韓国語の形態論上の複雑さは正確な形態素解析ツールの開発にとって非常に困難な課題となっている。形態素解析ツールは、形態素レベルで、基準化、語幹抽出(活用語尾の除去)、基本形復元(複数活用形を単一の語に還元)を行う。
４．正確な単語分節が困難であること。特に分ち書きを使用しない日本語と中国語に於てこの困難が顕著である。これは辞書検索や索引作成の際、一連のテキストを有意の意味単位に分解し、各単語の境界を識別するものである。この分野ではかなり研究が進んでいることが Emerson(2000)と Yu *et al.*(2000)によって報告されている。
５．雑多な検索技術。これは語彙素に基づく検索(例　'took off his jacket'から 'take off" + 'jacket')、統語表現(「研究をした」から「研究する」のような)を突きとめること、同義語の拡張、言語間情報検索(CLIR)(Goto *et al.* 2001)等の技術を含む。
６．各種の技術的な要求。これは複数文字セットと符号化方式間の相互変換、ユニコードや IME のサポート等を含む。これらの問題の殆どは Lunde(1999)で報告されているように十分に解決されている。
７．固有名詞は情報検索ツールにとってとりわけ大きな問題である。つまり、固有名詞は極めて数が多く、辞書無しには発見が困難であり、また表記がゆれている等がその原因である。
８．用語とその異表記の自動認識。本稿では取り扱わないが複雑なテーマである。それについては Jacquemin(2001)でヨーロッパの言語に関して詳細に述べられており、当研究所は現在、中国語と日本語に関して調査中である。

上記のどれをとってもそれ一つで論文に値する重要な課題である。ここでは、焦点は**異表記処理**に置く。異表記処理とは日中韓各語の表記のゆれの認識、基準化、変換のことである。本稿では日中韓各語に於ける表記のゆれの類型論を概観し、その言語学的諸問題を簡略に分析し、異表記処理に於て、どのように語彙データベースが中心的役割を果たすかについて論じる。

2. 中国語に於ける表記のゆれ

2.1. １つの言語、２つの文字体系

中華人民共和国に於ける戦後の言語改革の結果、数千に及ぶ漢字の字形が大胆に簡略化された(Zongbiao 1986)。簡略化された字形で書かれるこの中文は、**簡体中文(SC)**と呼ばれる。台湾、香港及び大半の華僑は**繁体中文(TC)**と呼ばれる旧字体で複雑な字形を使用し続けている。

中国語の書記体系の複雑さはよく知られている。この主な原因は、一般に用いられる文字数の多さ及びその字形の複雑さ、対応関係が多次元に亘る場合の、繁体中文と簡体中文の間で生じる大きな相違、繁体中文に於ける多数の表記のゆれの存在等が挙げられる。数多くの表記のゆれと、簡繁間変換の難しさは中国語の情報検索アプリケーションに特に重要な意味を持っている。

2.2. 中国語対中国語変換

簡繁間自動変換は**中国語対中国語変換**と呼ばれ、困難と陥穽に満ちている。その言語学的諸問題については Halpern and Kerman(1999)に、符号化方式及び文字セットに関しての技術的諸問題については Lunde(1999)に詳述されている。変換工程は３つのレベルで行うことができる。以下にそれらを精緻度順に手短に述べる。

2.2.1.コード変換方式

最も簡単であるが、最も信頼性の低い簡繁間変換は、符合位置対符合位置による変換方式であり、下記の表に示したように、対応表の中で変換前の符号位置を検索することによって変換を行う。これは**コード変換方式**と呼ばれている。おびただしい数の一対多の多義性(これは簡繁・繁簡の両方向で生じる)があるため、誤変換率は許容範囲を遥かに超えている。

表1 コード変換

簡体字	繁体字 1	繁体字 2	繁体字 3	繁体字 4	備考
门	們				一对一
汤	湯				一对一
发	發	髮			一对多
暗	暗	闇			一对多
干	幹	乾	干	榦	一对多

2.2.2.表記変換方式

簡繁間変換に於ける次の精緻化レベルは、変換処理の対象となる項目が文字セットのコードというよりはむしろ表記上の単位であるため、これを「表記変換方式」と呼ぶ。つまり、これらは、一定の意味を有する言語単位であり、特に複数漢字による語彙素である。コード変換方式が文字コード毎に変換するので、一意的でないのに対して、表記変換方式では表記対応表が単語レベルでの変換を可能にするので、より良い結果を得ることができる。

表2 表記変換

英語	簡体	繁体1	繁体2	誤り	備考
Telephone	电话	電話			一意的
We	我们	我們			一意的
Start-off	出发	出發		出髮 齣髮 齣發	一対多
Dry	干燥	乾燥		干燥 幹燥 榦燥	一対多
	阴干	陰乾	陰干		文脈による

このように、表記対応表を使用することでコード変換方式に付随する曖昧性が解決され、「**誤り**」の欄に示されているような誤変換を回避できる。単語分節の曖昧性があるので、このような変換を行う際には、生のテキストを意味単位に分解することができる形態素解析ツールを使用する必要がある(Emerson 2000)。

2.2.3.語彙素変換方式

簡繁間変換に於て更に精緻化され、且つ最大の難題となっているのは**語彙素変換方式**と呼ばれる。これは簡体中文と繁体中文間で表記的にではなく**意味的に**等しい語彙素どうしを変換するものである。例えば簡体の「信息」(xìnxī)「情報」は意味的に等しい繁体の「資訊」(zīxùn)に変換される。これは英語の lorry と米語の truck の相違に似ている。

簡体中文と繁体中文の間には語彙的な相違が多数あり、特に Tsou(2000)で論証されたように、専門用語や固有名詞では顕著である。例えば、'Osama bin Laden'には 10 以上の異表記がある。問題を更に複雑にしているのは、時折、正しい繁体が地域によって異なることである。語彙素変換は簡繁間変換に於ける、難易度最高のものであり、対応表の使用によってのみ成し得るものである。表３で、地域間の語彙素異表記の諸パターンを示す。

表３ 語彙素変換

英語	簡体	台湾繁体	香港 繁体	他の繁体	誤りの 繁体 (表記レベル)
Software	软件	軟體	軟件		軟件
Taxi	出租汽车	計程車	的士	德士	出租汽車
Osama bin Laden	奥萨马本拉登	奧薩瑪賓拉登	奧薩瑪賓拉丹		奧薩馬本拉登
Oahu	瓦胡岛	歐胡島			瓦胡島

2.3.繁体中文の異表記

繁体字には一定の正書法が存在しない。繁体字には多数の異体字があり、一般にそれらは混同して用いられる。繁体中文(及びある程度の簡体中文)を処理するには、対応表を使用してこれら表記のゆれの曖昧性を除去する必要がある(Halpern 2001)。

2.3.1.台湾と香港に於ける繁体中文の表記のゆれ

繁体中文は、辞書によって標準表記が異なっている場合がある。。繁体中文の表記のゆれは表４に示したように、種々のタイプに類別できる。

表４ 繁体中文の異表記

異表記 1	異表記 2	英語	備考
裏	裡	Inside	100% 互換性
教	教	Teach	100% 互換性
著	着	Particle	異表記2 Big5に無し
為	爲	For	異表記2 Big5に無し
沉	沈	sink; surname	部分的互換性
泄	洩	leak; divulge	部分的互換性

繁体中文に異体字が存在する理由は様々であり、例えば、繁体中文の一部は Big Five 文字セットにないものがある、時によっては繁体中文の中に簡体字を用いるものがある、等である。

2.3.2.中国本土に於ける表記のゆれ 対 台湾に於ける表記のゆれ

限られてはいるが、繁体字は中華人民共和国でも、古典文学や華僑向けの新聞等で使用されており、これは簡体字(GB　2312-80)を繁体字(GB/T 12345-90)に対応させる規格に基づいて行う。しかし、対応する繁体字が、台湾で一般に使用されている繁体字と必ずしも一致しているわけではない。ここでは前者を「本土型繁体字」(STC)、そして後者を「台香型繁体字」(TTC)と呼ぶことにする。

表５　 STC 対 TTC 異体字

ﾋﾟﾝｲﾝ	SC	STC	TTC
xiàn	线	綫	線
bēng	绷	綳	繃
cè	厕	廁	廁

3. 日本語に於ける表記のゆれ

3.1. １つの言語、４つの文字種

日本語の表記法は非常に不規則的である。異体字や間違えやすい同訓異字語が数多くあるため、日本語の書記体系は、中国語を含む他のどの主要言語よりもはるかに複雑である。その主な要因として、日本語を書き表すために用いられる４つの文字種の複雑な相互作用があり、結果としてしばしば予測不可能な様々な表記で、書き表すことが可能な無数の単語を生み出している(Halpern 1990, 2000)。表６は「取り扱い」の異表記の様々なパターンを示している。

表６「取り扱い」の異表記

トリアツカイ	異表記の類型
取り扱い	**「標準」表記**
取扱い	**送り仮名異表記**
取扱	**漢字のみ**
とり扱い	**漢字を**平仮名で代用
取りあつかい	**漢字を**平仮名で代用
とりあつかい	平仮名のみ

日本語の情報検索がいかに困難であるかという例に、有名な「きんのたまごをうむにわとり」がある。「標準」表記は「金の卵を産む鶏」であろうが、実際、「たまご」には４つの異表記(卵、玉子、たまご、タマゴ)があり、「にわとり」には３つ(鶏、にわとり、ニワトリ)、「うむ」には２つ(産む、生む)がある。「金の卵を生むニワトリ」「金の玉子を産む鶏」等並べ替えると異表記が 24 通りにもなる。ウェブの検索で容易に確認できるように、これらの表記のゆれはウェブページで頻繁に見られる。アプリケーションが異表記処理機能を備えていなければ、ユーザに表記のゆれを見つける望みがないのは明らかである。

3.2. 送り仮名のゆれ

日本語の表記のゆれで最もよく見られる類型の一つは、漢字の語幹につく「送り仮名」と呼ばれる仮名文字の語尾で起こる。動詞(「飛出す」)から派生した名詞(「飛出し」)のような一部の送り仮名の異表記をアルゴリズムで生成することは可能ではあるが、一般にハードコードされたデータ表が必要である。送り仮名の使用法がしばしば予測不可能であり、異表記が数多くあることから、日本語の異表記処理に於て重要な役割を果たすべきである。

表 7 送り仮名の異表記

英語	読み	標準表記	異表記
publish	*kakiarawasu*	書き表す	書き表わす 書表わす 書表す
Perform	*Okonau*	行う	行なう
Handling	*toriatsukai*	取り扱い	取扱い 取扱

3.3. 文字種間の表記のゆれ

日本語は４つの文字種を混在させて表記される(Halpern 1990):「**漢字**」(中国語由来の文字)、「**平仮名**」と「**片仮名**」と呼ばれる２種の音節文字、それに「**ローマ字**」(ラテンアルファベット)である。日本語情報検索に於て大きな役割を果たす文字種間の表記のゆれは、極めて盛んであり、殆ど予測不可能な程に様々な書き方をされる。そのため同一の単語が平仮名、片仮名もしくは漢字で書かれ、更には２種の文字の混ぜ書きの可能性もある。表８は日本語の文字種間の表記のゆれの主なパターンを示している。

表 8 文字種間の表記のゆれ

漢字 対平仮名	大勢　おおぜい
漢字 対 片仮名	硫黄　イオウ
漢字 対平仮名 対 片仮名	猫　ねこ　ネコ
片仮名 対 混ぜ書き	ワイシャツ Yシャツ
漢字 対片仮名 対混ぜ書き	皮膚　ヒフ　皮フ
漢字 対混ぜ書き	彗星　すい星
平仮名 対 片仮名	ぴかぴか　ピカピカ

3.4. 仮名表記のゆれ

近年、主として借用語を書くための文字として、音節文字である片仮名を使用する傾向が急激に高まっている。日本語情報検索に於ける主要な厄介ごとの一つは、この片仮名表記がしばしば不規則なことである。つまり、同一単語に対して、アルゴリズムでは生成不可能な、複数の予測し難い書き方をする場合が、極めて普通に見られる。一方、平仮名は主として文法要素及び和語の表記に用いられる。平仮名の表記法は概して安定しているが、少数ながら不規則な異表記が存在する。仮名表記のゆれの主な例をいくつか表９に挙げる。

表 9 片仮名と平仮名の異表記

類型	英語	読み	「標準」表記	異表記
長音符号	Computer	*konpyuuta* *konpyuutaa*	コンピュータ	コンピューター
長母音	Maid	*meedo*	メード	メイド
複数仮名	Team	*chiimu* *tiimu*	チーム	ティーム
旧仮名遣いの継承	Big	*Ookii*	おおきい	おうきい
づ 対 ず	Continue	*tsuzuku*	つづく	つずく

上記の表は、仮名表記のゆれの主要な類型を簡単に紹介したに過ぎない。他にも数多くあり、例えば、任意で中に使用する中黒、小文字の片仮名の異表記「クォ)対「クオ」、また伝統的仮名遣い(じ対ぢ)や旧仮名遣い(い対ゐ)の使用等がある。

3.5. 雑多な表記のゆれ

日本語には、他にも様々なタイプの表記のゆれがあるが、それらは本稿の範囲を超えており、ここでは 2･3 の主なものに触れておくに留める。詳細な記述が Halpern(2000)にある。

3.5.1. 漢字表記のゆれ

日本語の書記体系は戦後の時期に大きな改革を経て、漢字の字形は既に標準化されてはいるが、まだかなりの数の異体字が一般に使用されている。例えば、現代日本語の省略形(「歳」に対する「才」や「幅」に対する「巾」)、固有名詞や古典作品に残る伝統的な形(「島」に対する「嶋」や「発」に対する「發」)等である。

3.5.2. 同訓異字語

日本語の書記体系が複雑である理由として、多数の同訓異字語(発音は同じだが表記が異なる語)とその様々な表記のゆれの存在がある(Halpern 2000)。ひとつひとつの漢字に多くの訓読みがあるだけでなく、多くの訓読みの単語が驚くほど様々な書き方をされる。同訓異字語の中には、各々が近似又は同一の意味を持つために混同され易いものも多い。例えば、「のぼる」は「上る」と書くと go up「上に行く」という意味だが、「登る」と書くと climb「(手足を使って)登る」を意味する。また、「やわらかい」は「柔らかい」もしくは「軟らかい」と書くが、意味は同じである。

4. 韓国語に於ける表記のゆれ

4.1. 不規則な表記法

韓国語の表記は多くの人が思うほど規則的ではない。ハングルはしばしば「論理的」と言われるが、実は現代韓国語にはかなりの表記のゆれがある。このことは、韓国語の形態的な複雑さとあいまって、情報検索ツールの開発者にとって大きな課題である。韓国語に於ける主な表記のゆれは以下の通りである。

4.2 ハングル表記のゆれ

韓国語に於ける表記のゆれで最も重要なものに借用語を記述する際のハングル表記のゆれがある。もう一つは表 10 で示したように韓国人以外の人名を書く場合の表記のゆれである。

表10 ハングルの異表記

cake	케이크 (*keikeu*)	케잌 (*keik*)
yellow	옐로우 (*yelrou*)	옐로 (*yelro*)
Mao Zedong	마오쩌뚱 (*maojjeottung*)	모택동 (*motaekdong*)
Clinton	클린턴 (*keulrinteon*)	클린톤 (*keulrinton*)

4.3 文字種間の表記のゆれ

韓国語書記体系の複雑さの一因は、複数の文字種を用いることにある。韓国語は３つの文字種を混ぜて書き、一つはハングルと呼ばれる字母の音節文字、又「ハンジャ」と呼ばれる漢字(使用頻度は減少している)、それにローマ字と呼ばれるラテンアルファベットである。文字種間の表記のゆれも珍しいことではなく、その主なパターンを表 11 に示した。

表11 文字種間異表記

異表記の類型	英語	異表記 1	異表記 2	異表記3
漢字対ハングル	Many people	大勢 (*daese*)	대세 (*daese*)	
ハングル対混ぜ書き	shirt	와이셔츠(*wai-syeacheu*)	Y셔츠 (*wai-syeacheu*)	
ハングル対 数詞 対 漢字	One o'clock	한시 (*hansi*)	1시 (*hansi*)	一時 (*hansi*)
英語対ハングル	sex	sex	섹스 (*sekseu*)	

4.4. 雑多な表記のゆれ

4.4.1 北朝鮮 vs.南朝鮮

ハングル表記法が不規則であるもう一つの原因は、韓国(S.K.)と北朝鮮(N.K.)の綴りの違いである。主な違いは、借用語や朝鮮以外の固有名を記述する際に本来の韓国語の単語が優先されることである。主な類型を下に示す。

1. **地名:**「大阪」 北朝鮮 오사까 (*osakka*) 対 韓国 오사카 (*osaka*)
2. **人名:**「ブッシュ」 北朝鮮 부슈 (*busyu*) 対 韓国부시 (*busi*)
3. **借用語:**「ミサイル」北朝鮮 미싸일 (*missail*) 対 韓国 미사일 (*misail*)
4. **ロシア語対英語:** 北朝鮮 그루빠 (*guruppa*) 対 韓国 그룹 (*geurup*)
5. **形態音素:** 北朝鮮 람용 (*ramyong*) 対 韓国 남용 (*namyong*)

4.4.2 新表記法.対 旧表記法

ハングル文字は歴史上数回の改革を経ており、最近では 1988 年に行われた。新表記法は現在確立しているが、旧表記法に影響を受けた語の使用頻度が高くまたその数も多いので、まだ旧正書法も重要である。例えば、現在の 일군 「労働者」 (*ilgun*) は 1988 年以前は일꾼 (*ilkkun*)、また、빛깔 (*bitkkal*) 「色」は빛갈(*bitgal*)と 書かれていた。

4.4.3 漢字表記のゆれ

韓国に於ける言語改革では漢字の字形の簡略化は行われなかったが、日本の韓国占領の結果、簡略化した日本語の文字(例えば「發」 (*bal*)の代わりに「発」)が多く使用されることとなった。

4.4.3 雑多な表記のゆれ

表記のゆれには他に種々の類型があるが、それは本稿の範囲を超えている。ここでは複数の語から成る合成語を分けて書く場合の略語と頭字語の使用法について述べる。例えば 「カリブ海」 (*karibeuhae*)は空白なしで (카리브해) と表記されることもあるし、空白を入れて (카리브 해)と表記されることもある。

5 語彙データベースの役割

日中韓各語では表記法が不規則なので、異表記処理のような語彙素レベルの処理を、例えばバイグラミングのような確率的手法だけに基づいて行うことはできない。Brill(2001)及び Goto *et al.*(2001)等の多くの試みがこの方針に沿って行われ、ある研究者は辞書を用いた手法と同等の成果があったと主張している。一方 、Kwok(1997)は非常に小さな辞書と単純な分節ツールで良い成果を上げたと報告している。
このような方法は単なる情報検索(文書検索と同義)には十分かもしれないが、異表記処理や簡繁間変換を行うには不十分である。Emerson(2000)や他の研究

者は、語彙素を処理できる強力な形態素解析ツールは、バイグラムやNグラムよりむしろ大規模な計算機辞書(10 万語でもまだ非常に小さすぎる)を備えていなければならないことを示している。

日中韓辭典研究所(CJKI)は日中韓各語のコンピュータによる辞書編纂を専門にしており、包括的な日中韓各語語彙データベース(現在 550 万語)を編纂するために、特に異表記処理と固有名詞に重点を置いて、たゆまぬ研究と開発を展開している。知的情報検索用ツールと異表記処理に役立つ主要なデータベースを下記に示す。

1　中国語対中国語変換

1996 年、日中韓辭典研究所(CJKI)は簡繁間変換問題を徹底的に調査するプロジェクトに着手し、100 パーセントに近い変換精度を持つ包括的な対応表(現在簡体 130 万、繁体 120 万項目)を構築することを目標としている。

a. 簡繁間コードレベル対応表
b. 一般語彙の簡繁間表記レベル及び語彙素レベル対応表
c. 固有名詞の簡繁間表記レベル対応表
d. 専門用語(特に情報技術用語)の包括的簡繁間表記レベル及び語彙素レベル対応表

2　繁体表記法正規化テーブル

a. 繁体正規化対応表
b. 繁体大陸型・台香型間文字対応表

3　日本語異表記データベース

a. 包括的日本語表記法データベース
b. 同訓異字語意味分類データベース
c. 同義語拡張処理用同義語グループ（意味分類済み）(日本語シソーラス)
d. 言語間情報検索用英日辞書
e. 未登録異表記の識別ルール集

結論

日中韓各語の情報検索用ツールは、情報検索に於ては特に、また情報技術一般に於ても、益々重要になりつつある。これまで述べてきたように、 日中韓各語の表記法が不規則であるために、知的情報検索には高度な形態素解析ツールだけでなく異表記処理のために高度に精製された語彙データベースが必要である。表記の曖昧性を除去する日中韓語情報検索ツールがあるとしても、その数は非常に少ない。なぜならば、真に「知的」情報検索を実現するには、辞書を用いた異表記処理機能を備えているだけでなく、言語間情報検索同義語拡張処理、同音異義語間検索といった新しい技術も備えていなければならないからである。

現在、当研究所は知的 日中韓各語 情報検索ツールの構築や、正確な分節をする技術のサポートに必要とされる語彙資源の、更なる拡張を図っている。

参考文献

Brill, E. and Kacmarick, G. and Brocket, C. (2001) *Automatically Harvesting Katakana-English Term Pairs from Search Engine Query Logs*. Microsoft Research, Proc. of the Sixth Natural Language Processing Pacific Rim Symposium, Tokyo, Japan.

Emerson, T. (2000) *Segmenting Chinese in Unicode.* Proc. of the 16th International Unicode Conference, Amsterdam

Goto, I., Uratani, N. and Ehara T. (2001) *Cross-Language Information Retrieval of Proper Nouns using Context Information*. NHK Science and Technical Research Laboratories. Proc. of the Sixth Natural Language Processing Pacific Rim Symposium, Tokyo, Japan

Jacquemin, C. (2001) *Spotting and Discovering Terms through Natural Language Processing*. The MIT Press, Cambridge, MA

Halpern, J. (1990) *Outline Of Japanese Writing System.* In "New Japanese-English Character Dictionary*",* 6th printing, Kenkyusha Ltd., Tokyo, Japan (www.kanji.org/kanji/japanese/writing/outline.htm)

Halpern, J. and Kerman J. (1999) *The Pitfalls and Complexities of Chinese to Chinese Conversion*. Proc. of the Fourteenth International Unicode Conference in Cambridge, MA.

Halpern, J. (2000) *The Challenges of Intelligent Japanese Searching.* Working paper (www.cjk.org/ cjk/joa/joapaper.htm), The CJK Dictionary Institute, Saitama, Japan.

Halpern, J. (2001) *Variation in Traditional Chinese Orthography.* Working paper (www.cjk.org/cjk/ cjk/reference/chinvar.htm), The CJK Dictionary Institute, Saitama, Japan.

Kwok, K.L. (1997) *Lexicon Effects on Chinese Information Retrieval*. Proc. of 2nd Conf. on Empirical Methods in NLP. ACL. pp.141-8.

Lunde, Ken (1999) *CJKV Information Processing.* O'Reilly & Associates, Sebastopol, CA.

Yu, Shiwen, Zhu, Xue-feng and Wang, Hui (2000) *New Progress of the Grammatical Knowledge-base of Contemporary Chinese*. Journal of Chinese Information Processing, Institute of Computational Linguistics, Peking University, Vol.15 No.1.

Tsou, B.K., Tsoi, W.F., Lai, T.B.Y. Hu, J., and Chan S.W.K. (2000) *LIVAC,* a *Chinese synchronous corpus, and some applications.* In "2000 International Conference on Chinese Language ComputingICCLC2000", Chicago .

Zongbiao (1986) 简化字总表 (*Jianhuazi zongbiao)* (Second Edition). 国家语言文字工作委员会, 语文出版社, China.