CJKI多言語データベース一覧

No.	Code	名称	タイプ	語数	特徴	使用分野
中国語データベース						
C01	CETERM	中英専門用語データベース	専門用語	3,319,160	化学、コンピュータ、医学、土木、経済、機械等に特化した68分野の約330万語を収録した大型中英英中専門用語データベース。	機械翻訳、用語抽出及び索引付け等の情報検索、形態素解析と単語分割等の自然言語処理、電子辞書とモバイルアプリの開発
C02	CED	中英一般語彙データベース	一般語彙	500,000	有名大学の言語学者たちと協力し開発したデータベース。権威ある中国語辞書を参考に、ネイティブ編集チームによる緻密な校正を経て完成。一般語彙、専門用語と重要な固有名詞をカバーし、すべての見出し語にピンインと品詞コードを付記。	機械翻訳、用語抽出及び索引付け等の情報検索、電子辞書・モバイルアプリ・言語学習アプリの開発
C03	ECD	英中一般語彙データベース	一般語彙	80,000	一般語彙と重要な固有名詞を8万語収録。見出し語はご要望に応じ10万語まで拡張可能。	機械翻訳、用語抽出及び索引付け等の情報検索、電子辞書・モバイルアプリ・言語学習アプリの開発
C04	CEC	中英成語（諺と慣用句）データベース	一般語彙	50,000	成語は中国古典の故事に由来するものが多いので、字義通りに訳すと意味が正しく伝わらなかったりする。例えば、「事態が切迫してからやっと準備する」を喩える「臨陣磨槍」は、MTやNMTシステムでは字義に基づいた直訳「戦に臨んで槍を磨く」以上は訳出されない。本データベースはこのような中国語成語（諺と慣用句を含む）に絡む機械翻訳の精度を向上させることができる。頻度情報、字義通りの翻訳、構文コード、類似する英訳等も採録。	機械翻訳(MT)とニューラル機械翻訳(NMT)システムの品質向上

C05	YPD	広東語発音データベース	一般語彙	300,000	本データベースは包括的で且つ言語学的な正確さを誇るものである。広東語の音韻論と意味論知識に基づき作成し、入念な校正により、きわめて難解な多音字と変調を正確に表記した。Big5文字セットにある単漢字13,000字、その読みとローマ字異表記8万項目、熟語30万語を収録。	中国語入力システム、教育アプリ、音声認識システム、音声合成システム等の開発
C06	CJTERM	中日専門用語データベース	専門用語	820,000	コンピュータ、機械、生物、化学、医学等20分野の科学技術用語を82万語収録。	機械翻訳用辞書、用語抽出及び索引付け等の情報検索、形態素解析と単語分割等の自然言語処理、電子辞書とモバイルアプリの開発
C07	CPD	中国語(簡/繁)ピンインデータベース	自然言語処理資源	2,500,000	中国語ピンインの大型データベース。同じ単語が中国大陸と台湾で発音が異なるもの(期待 qī dài・中国大陸とqí dài・台湾)が少なからずあるが、本データベースはその違いを正確に捉え正しく表記している。	中国語入力システム、教育アプリ、音声認識システム、音声合成システム等の開発
C08	C2C	中国語(簡/繁)変換システム	自然言語処理資源	700,000	表記レベル変換と語彙レベル変換に対応する中国語簡体字と繁体字の変換テーブル。非常に包括的なマッピングテーブルで、一般語彙と専門用語及び固有名詞を合わせ約70万語を網羅。ピンイン、品詞コード、語彙分類コード等の属性を含む。	簡体字と繁体字の変換システム
C09	CHD	中国語ピンインデータベース	一般語彙、固有名詞、専門用語	600,000	一般語彙と出現頻度が高い専門用語と固有名詞を一部収録。網羅している見出し語は正確で収録範囲が幅広く、すべての項目に種類と品詞コード等の属性を付記。多音字を多数収録、実際の発音に基づき軽声を表記等の特長がある。	中国語入力システム、教育アプリ、音声認識システム、音声合成システム等の開発
		包括的な中国人名データベース	その他のデータベースのM02を参照。			

C10	CEN CEP	中英固有名詞データベース	固有名詞	2,245,000	包括的な中英固有名詞(人名・地名)データベース。中国のみならず、日本、韓国、西洋の人名と地名も幅広く収録。	機械翻訳、用語抽出及び形態素解析アプリ、電子辞書とモバイルアプリ、入力システム、固有表現認識、データクレンジング、地図と地理データの開発等
C11	CJN CJP	中日固有名詞データベース	固有名詞	2,096,200	包括的な中日固有名詞(人名・地名)データベース。中国のみならず、日本、韓国、西洋の人名と地名も幅広く収録。	機械翻訳、用語抽出及び形態素解析アプリ、電子辞書とモバイルアプリ、入力システム、固有表現認識、データクレンジング、地図と地理データの開発等
		中韓固有名詞データベース	韓国語データベースのK04を参照。			
C12		中国人名データベース	固有名詞	1,650,000	中国人の人名データベース。すべての見出し語にピンイン、性別コード、姓・名のフラグ等を付記。	機械翻訳、用語抽出及び形態素解析のためのアプリ、電子辞書とモバイルアプリ、入力システム、固有表現認識、データクレンジング、地図と地理データの開発等
C13	CNV	中国人名異表記データベース	固有名詞	7,600,000	中国人の人名と主なローマ字表記及びその異表記を網羅したデータベース。方言表記(広東語、客家語、福建語等)を追加した大幅な拡張も可能。	機械翻訳、用語抽出及び形態素解析のためのアプリ、電子辞書とモバイルアプリ、入力システム、固有表現認識、データクレンジング、地図と地理データの開発等

C14	CFN	中国人氏名データベース	固有名詞	4,000,000	著名人を含む実在する中国人のフルネームを収録したデータベース。見出し語すべてにピンインを付記。	機械翻訳、用語抽出及び形態素解析のためのアプリ、電子辞書とモバイルアプリ、入力システム、固有表現認識、データクレンジング、地図と地理データの開発等
C15	CLD	中国語語彙データベース	自然言語処理資源	500,000	包括的な中国語単言語語彙データベース。一般語彙と重要な専門用語を一部収録し、簡体字と繁体字の二つのモジュールがある。各見出し語には、音韻・文法・形態・意味に基づく多様な分類コードを併記。	機械翻訳、用語抽出及び形態素解析等の自然言語処理
C17	CWL	中国語簡体字用語集	自然言語処理資源	5,261,017	中国語簡体字の大型単言語用語集。ピンイン付きで、音声合成等の音声アプリ開発に最適。	情報検索、形態素解析、単語分割、音声アプリ等多様な自然言語処理
C18	CWL	中国語繁体字用語集	自然言語処理資源	5,465,068	中国語繁体字の大型単言語用語集。ピンイン付きで、音声合成等の音声アプリ開発に最適。	情報検索、形態素解析、単語分割、音声アプリ等多様な自然言語処理
日本語データベース						
J01	JED	日英一般語彙データベース	一般語彙	110,000	品詞コード、読みを含む一般語彙辞書。電子辞書やオンライン翻訳ツールを利用するユーザの利便性に即して考案された辞書。辞書内容は、言語知識を得るに十分な上、要領よく纏めてある。	機械翻訳、用語抽出及び索引付け等の情報検索、電子辞書・モバイルアプリ・言語学習アプリの開発
J02	EJD	英日一般語彙データベース	一般語彙	82,000	約8万2000語の見出し語を収録。品詞コードと文法・形態に基づく分類コードが提供可能。	機械翻訳、用語抽出及び索引付け等の情報検索、電子辞書・モバイルアプリ・言語学習アプリの開発

J03	JMP	日本POI多言語（日中韓英）データベース	固有名詞	1,172,083	日本地名とPOIの大型多言語（日中韓英）データベース。	機械翻訳、情報検索、形態素解析、電子辞書とモバイルアプリ、入力システム、固有表現認識等幅広い分野のアプリ開発
J04	JPD	日本語音韻データベース	自然言語処理資源	70,000	音韻・音声知識に精通した経験豊かな編集チームが編纂した日本語音韻データベース。実際の発音を転記したIPA（SAMPAも提供可能）と、固有名詞データベースでは大変ユニークなアクセント情報を提供可能。	音声合成システムの開発、外国語としての日本語教育、日本語音声技術の研究・開発
J05	JLD	日本語語彙データベース	自然言語処理資源	290,000	日本語単言語語彙データベース。派生語、接辞（接尾辞と接頭辞）、拘束形態素等の文法情報を詳しく収録。	機械翻訳、用語抽出及び形態素解析等の自然言語処理
J07	JOD	日本語異表記データベース	自然言語処理資源	127,600	日本語異表記データベース（JOD）は、日本語異表記の曖昧性解消に役立ち、情報検索、機械翻訳、形態素解析等の分野における品質向上に貢献できる。	機械翻訳、情報検索、形態素解析等自然言語処理
J08	JMP	日本POI多言語（日中韓英を除く）データベース	固有名詞	1,951,518	日本地名とPOIの大型データベース。ドイツ語やフランス語等のヨーロッパ言語と、ベトナム語やインドネシア語等のアジア言語により構成。	機械翻訳、情報検索、形態素解析、電子辞書とモバイルアプリ、入力システム、固有表現認識等幅広い分野におけるアプリ開発
J09	JCD	日本企業・団体名データベース	固有名詞	580,000	日本企業と団体名データベース。英訳を一部提供。	BIツールと機械翻訳における用語抽出及び形態素解析
J10	JNV	日本人名異表記データベース	固有名詞	4,000,000	日本人名とそのローマ字異表記を400万項目収録。性別、姓・名の分類コード、頻度情報を付記。	BIツールと機械翻訳
		韓日固有名詞データベース	韓国語データベースのK03を参照。			

J11	JEN	日英固有名詞データベース	固有名詞	660,000	66万語の見出し語を収録。平仮名、ローマ字読み、分類コード、ローカルコード、英訳等の情報も提供。日本と外国の人名と地名を広範囲に亘り収録。	機械翻訳、情報検索、形態素解析、電子辞書とモバイルアプリ、入力システム、固有表現認識等幅広い分野におけるアプリ開発
J12	JETERM	日英専門用語データベース	専門用語	920,390	日英双方向対訳専門用語データベース。コンピュータ、経済、生物等幅広い分野の専門用語を収録。	機械翻訳、用語抽出及び索引付け等の情報検索、形態素解析及び単語分割等の自然言語処理、電子辞書とモバイルアプリの開発
J13	JWL	日本語用語集	自然言語処理資源	2,646,853	日本語の大型単言語用語集。読み付きで、音声合成等の音声アプリ開発に最適。	情報検索、形態素解析、単語分割、音声アプリ等多様な自然言語処理
J14	JFULEX	日本語全活用形データベース	自然言語処理資源	120,000,000	包括的な日本語一般語彙の全活用形データベース。あらゆる活用形を含む。	機械翻訳と自然言語処理アプリの品質向上、形態素解析、固有表現認識、検索エンジン
		韓日専門用語データベース	韓国語データベースのK01を参照。			
		中日専門用語データベース	中国語データベースのC06を参照。			
韓国語データベース						
K01	KJTERM	韓日専門用語データベース	専門用語	988,347	韓日双方向対訳専門用語データベース。土木、経済、機械、コンピュータ等幅広い分野の専門用語を収録。	機械翻訳、用語抽出及び索引付け等情報検索、形態素解析及び単語分割等の自然言語処理、電子辞書とモバイルアプリの開発

K02	KEN KEP	韓英固有名詞データベース	固有名詞	1,820,200	包括的な韓英固有名詞(人名・地名)データベース。韓国のみならず、中国、日本、西洋の人名と地名も幅広く収録。	機械翻訳、用語抽出及び形態素解析アプリ、電子辞書とモバイルアプリ、入力システム、固有表現認識、データクレンジング、地図と地理データの開発等
K03	KJN KJP	韓日固有名詞データベース	固有名詞	2,250,700	包括的な韓日固有名詞(人名・地名)データベース。韓国のみならず、中国、日本、西洋の人名と地名も幅広く収録。	機械翻訳、用語抽出及び形態素解析アプリ、電子辞書とモバイルアプリ、入力システム、固有表現認識、データクレンジング、地図と地理データの開発等
K04	KCN KCP	韓中固有名詞データベース	固有名詞	2,483,600	包括的な韓中固有名詞(人名・地名)データベース。韓国のみならず、中国、日本、西洋の人名と地名も幅広く収録。	機械翻訳、用語抽出及び形態素解析アプリ、電子辞書とモバイルアプリ、入力システム、固有表現認識、データクレンジング、地図と地理データの開発等
K05	KLD	韓国語語彙データベース	自然言語処理資源	97,000	韓国語単言語語彙データベース。派生と屈折に関わる接辞、小詞、助動詞、活用型等活用形認識に必要な情報を幅広く収録したデータベースで、形態素解析に役立つ。	機械翻訳、用語抽出及び形態素解析等の自然言語処理
K06	KWL	韓国語用語集	自然言語処理資源	1,040,887	韓国語の大型単言語用語集。読み付きで、音声合成システム等の音声アプリ開発に最適。	情報検索、形態素解析、単語分割、音声アプリ等多様な自然言語処理
K07	KNV	韓国人名異表記データベース	固有名詞	183,000	韓国人の人名と主なローマ字表記を網羅したデータベース。文化観光部2000年式に従った表記も収録。	機械翻訳、用語抽出及び形態素解析のためのアプリ、電子辞書とモバイルアプリ、入力システム、固有表現認識、データクレンジング、地図と地理データの開発等
アラビア語データベース						

A01	ArabLEX	アラビア語全活用形データベース	自然言語処理資源	530,000,000	派生・屈折・接語からなる活用形を包括的に網羅した全活用形データベース。見出し語はすべて母音付きで、対応する母音無し表記と基準形を併記している。形態・文法・音韻・異表記に基づく各種属性、音声表記、音素表記等も付記	機械翻訳と音声（合成・認識）関係の言語モデルの最適化
A02	DAP	アラビア語複数形データベース	自然言語処理資源	3,137	アラビア語の規則・不規則複数形を収録したデータベース。専門家達が数年を費やして開発したもので、品詞コード、集合名詞コード、性別コード、完全な母音表記等多様な文法情報も提供。	ソフトウェア開発、機械翻訳、アラビア語学習
A03	DAN	アラブ人名データベース	固有名詞	6,500,000	包括的なアラブ人名データベース。アラブ人名と異表記をすべて基準形に対応付けたほか、付加情報も多数提供。	固有表現認識、機械翻訳、異表記の正規化、アラブ人名検索、リスクコンプライアンスシステム、音訳・字訳等の翻訳
A04	DANA	アラビア語アラブ人名データベース	固有名詞	222,000	アラブ人名と異表記データベースの基準形データ。よくある誤表記を意図的に含んだ数十万語の表記を収録。見出し語のアラブ人名はすべて正規化し母音記号を付記。	固有表現認識、機械翻訳、異表記の正規化、アラブ人名検索、音訳・字訳等の翻訳
A05	DAFNA	アラビア語外国人名データベース	固有名詞	37,000	アラブ人以外の外国人名のアラビア語訳語とその異表記。	固有表現認識、機械翻訳、異表記の正規化、アラブ人名検索、リスクコンプライアンスシステム、音訳・字訳等の翻訳
A06	DAPNA	アラビア語地名データベース	固有名詞	10,000	よく知られている世界地名を網羅した英語とアラビア語の双方向対訳データベース。地名は、現代標準アラビア語 (MSA)表記と幾つかの異表記で収録。	固有表現認識、機械翻訳、異表記の正規化、アラブ人名検索、音訳・字訳等の翻訳

A07	AWL	アラビア語用語集	自然言語処理資源	210,000	アラビア語の大型単言語用語集。音声表記付きで、音声合成システム等の音声アプリ開発に最適。	情報検索、形態素解析、単語分割、音声アプリ等多様な自然言語処理
A08	DiaLEX	アラビア語方言全活用形データベース	自然言語処理資源	100,000,000	エジプト方言、クウェート方言、カタール方言、エミラティ方言、サウジアラビアのナジディ方言、サウジアラビアのヘジャジ方言、パレスチナ方言等、アラビア語の主な方言と下位方言をカバーする包括的な計算学的辞書	情報検索、形態素解析、単語分割、音声アプリ等多様な自然言語処理
多言語データベース						
		日本POI多言語（日中韓英）データベース	日本語データベースのJ03を参照。			
M02		多言語固有名詞データベース	固有名詞	150,000	中国語簡体字、中国語繁体字、日本語、韓国語、英語、（ご要望に応じ）アラビア語の6言語をカバーする多言語固有名詞データベース。任意方向での提供が可能。ピンイン、注音、平仮名、主な各種ローマ字表記、意味分類コード、ローカル等多様な付加情報を提供。	オンラインの多言語地図サービス、固有表現認識、機械翻訳、用語抽出等
その他のデータベース						
X01	DPN	ペルシャ人名データベース	固有名詞	450,000	ペルシア語音韻論に詳しいネイティブ専門家チームと協力して開発したユニークなデータベース。現実世界における出現頻度情報も収録。	固有表現認識、機械翻訳、異表記の正規化、ペルシャ人名の情報検索、リスクコンプライアンスシステム、音訳と字訳翻訳等
X02	SFULEX	スペイン語全活用形単言語データベース	自然言語処理資源	1,000,000	非常に包括的なスペイン語の全活用形データベース。一般語彙の屈折、複数形、女性形、接辞等様々な語形をカバーする全活用形を収録。	機械翻訳及びその他の自然言語処理、形態素解析、固有表現認識、検索エンジン等

X03	SFULEX	スペイン語全活用形対訳データベース	自然言語処理資源	26,000,000	非常に包括的なスペイン語と英語の全活用形対訳データベース。一般語彙の屈折、複数形、女性形、接辞等様々な語形をカバーする全活用形、及びその英語対訳語を提供。	機械翻訳及びその他の自然言語処理、形態素解析、固有表現認識、検索エンジン等
X04		越日一般語彙データベース	一般語彙	140,000	越日一般語彙データベース。品詞コード、読み、漢越語の漢字等を付記。	機械翻訳、電子辞書・モバイルアプリ・言語学習アプリの開発