The CJK Dictionary Institute, Inc.
日中韓辭典研究所

ID	Code	Resource	Type of database	Headwords	General description	Resource purpose
	Chinese					
C01	CETERM	Simplified Chinese↔English Technical Terms	Technical Terms	3,319,160	A large comprehensive dictionary of Chinese-English technical terms, covering over 3.3 million terms from 68 domains, including chemical, computer/IT, medical, civil engineering, economy/finance, and mechanical engineering.	MT dictionaries; information retrieval applications for accurate term recognition and indexing; NLP tools like morphological analyzers and tokenizers; handheld electronic dictionaries and smartphone applications
C02	CED	Simplified Chinese-to-English Dictionary	General Vocabulary	500,000	Compiled in collaboration with lexicographers from a leading Chinese university, and based on the world's most authoritative and comprehensive dictionaries that have been published in China. It has undergone extensive proofreading and validation by a team of native Chinese editors. Covers general vocabulary, technical terms, and important proper nouns. Includes pinyin and part-of-speech codes.	MT dictionaries; CLIR applications for accurate term recognition and indexing; handheld electronic dictionaries and mobile device applications, and language learning applications
C03	ECD	English-to-Simplified Chinese Dictionary	General Vocabulary	80,000	80,000 headwords, expandable to 100,000, of general vocabulary and important proper names	MT dictionaries; CLIR applications for accurate term recognition and indexing; handheld electronic dictionaries and mobile device applications
C04	CEC	Chinese-English Database of Proverbs and Idioms (Chengyu)	General Vocabulary	50,000	This database is important for translating 成語 *chengyu* (Chinese proverbs and idioms), which cannot be translated literally since they are often based on classical Chinese. For example, 臨陣磨槍, literally 'face battle sharpen spear', which means "do something at the last moment," cannot be correctly translated by MT or NMT systems based on the characters alone. The database includes a fairly large variety of English translations, as well as ranking information, literal meanings, and syntactic codes.	For enhancing the accuracy of MT and NMT systems.
C05	YPD	Cantonese Readings Database	General Vocabulary	300,000	This database is not only comprehensive but also linguistically accurate. It is based on solid principles of Cantonese phonology and semantics, and takes into account the phenomena of polyphony as well as tone change, which is unpredictable and requires manual proofreading. It covers 300,000 entries, including 80,000 readings and romanized variants for the 13,000 Big Five single characters.	Chinese IME systems, pedagogical applications, transcription systems, speech synthesis.
C06	CJTERM	Chinese-Japanese Technical Terms Dictionary	Technical Terms	820,000	Covers over 820,000 terms from over 20 science and technology domains, including computers/IT, mechanical engineering, biotechnology, chemistry, and medicine.	MT dictionaries; information retrieval applications for accurate term recognition and indexing; NLP tools like morphological analyzers and tokenizers; handheld electronic dictionaries and mobile device applications
C07	CPD	Chinese Phonological Database	NLP Lexicons	2,500,000	A large-scale database of Chinese pinyin readings. Especially noteworthy are the differences in pronunciation between the PRC and Taiwan, for example 期待 *qī dài* (PRC) and *qí dài* (Taiwan).	Chinese IME systems, pedagogical applications, transcription systems, speech technology.

ID	Code	Resource	Type of database	Headwords	General description	Resource purpose
C08	C2C	Simplified to Traditional Chinese Conversion	NLP Lexicons	700,000	SC<>TC mapping tables for Orthographic and Lexemic conversion levels together with a conversion engine. The mapping tables are comprehensive, and include approximately 700,000 items covering general vocabulary and some technical terms and proper nouns. They also include various other attributes, such as pinyin readings, part of speech, and semantic classification codes.	Conversion between Simplified and Traditional Chinese.
C09	CHD	Hanzi Pinyin Database for Simplified Chinese	General Vocabulary, Proper Nouns, Technical Terms	600,000	Covers entries of general vocabulary, along with high-frequency technical terms and proper nouns. In addition to large coverage and high level of accuracy, the database has several special features including explicit codes to indicate headword type and part-of speech, coverage of all polyphones, and correct pinyin for the neutral tone based on actual usage.	Chinese IME systems, pedagogical applications, transcription systems, speech synthesis.
		CJKI Comprehensive Database of Chinese Personal Names	See M02 in Multilingual section			
C10	CEN CEP	Chinese-English Database of Proper Nouns	Proper Nouns	2,245,000	A large comprehensive database of Chinese-English personal and place names, with coverage of not only native Chinese proper nouns, but also Japanese, Korean, and Western proper nouns as well.	Used for a wide variety of applications, including MT, information retrieval, morphological analysis, electronic and mobile platform dictionaries, IME systems, NER, data cleansing, and mapping and geodata.
C11	CJN CJP	Chinese-Japanese Database of Proper Nouns	Proper Nouns	2,096,200	A large comprehensive database of Chinese-Japanese personal and place names, with coverage of not only native Chinese proper nouns, but also Japanese, Korean, and Western proper nouns as well.	Used for a wide variety of applications, including MT, information retrieval, morphological analysis, electronic and mobile platform dictionaries, IME systems, NER, data cleansing, and mapping and geodata.
		Korean-Chinese Database of Proper Nouns	see K04 in Korean section			
C12		Database of Chinese Names	Proper Nouns	1,650,000	Chinese name components, accompanied by accurate pinyin readings, gender codes, and flags denoting whether name is a given name, surname, or both.	Used for a wide variety of applications, including MT, information retrieval, morphological analysis, electronic and mobile platform dictionaries, IME systems, NER, data cleansing, and mapping and geodata.
C13	CNV	Database of Chinese Name Variants	Proper Nouns	7,600,000	Provides comprehensive coverage for the major Chinese romanization systems and their variants, and if needed can be expanded considerably with dialectal variants (Cantonese, Hakka, Hokkien, etc.).	Used for a wide variety of applications, including MT, information retrieval, morphological analysis, electronic and mobile platform dictionaries, IME systems, NER, data cleansing, and mapping and geodata.

ID	Code	Resource	Type of database	Headwords	General description	Resource purpose
C14	CFN	Database of Chinese Full Names	Proper Nouns	4,000,000	Covers Chinese full names of real people, including celebrities. Includes pinyin readings.	Used for a wide variety of applications, including MT, information retrieval, morphological analysis, electronic and mobile platform dictionaries, IME systems, NER, data cleansing, and mapping and geodata.
C15	CLD	Chinese Lexical Database	NLP Lexicons	500,000	A comprehensive monolingual lexical database of Chinese consisting of Simplified and Traditional Chinese modules, covering general vocabulary and important technical terms. Each entry is accompanied by various attributes, such as phonological, grammatical, and morphological information, as well as semantic classification codes.	Fine-tuned for NLP applications such as MT, information retrieval and morphological analysis.
C17	CWL	Comprehensive Wordlist of Simplified Chinese	NLP Lexicons	5,261,017	Comprehensive monolingual wordlist for Simplified Chinese. Pinyin is provided, making this database ideal for speech-related applications such as speech synthesis.	Suitable for a variety of natural language processing applications, including information retrieval, morphological analysis and word segmentation, as well as speech-related applications.
C18	CWL	Comprehensive Wordlist of Traditional Chinese	NLP Lexicons	5,465,068	Comprehensive monolingual wordlist for Traditional Chinese. Zhuyin is provided, making this database ideal for speech-related applications such as speech synthesis.	Suitable for a variety of natural language processing applications, including information retrieval, morphological analysis and word segmentation, as well as speech-related applications.
	Japanese					
J01	JED	Japanese – English Dictionary	General Vocabulary	110,000	This database covers general vocabulary, and includes part-of-speech codes and readings. This up-to-date dictionary is optimized for the convenience of users of electronic dictionaries and online translation tools. It has just the right amount of detail: enough equivalents to give an in-depth understanding, yet short enough not to clutter up the screen.	MT dictionaries; CLIR applications for accurate term recognition and indexing; handheld electronic dictionaries and mobile device applications, and language learning applications
J02	EJD	English – Japanese Dictionary	General Vocabulary	82,000	This database covers about 82,000 headwords, and includes part-of-speech codes as well as other grammatical/phonological data upon request.	MT dictionaries; CLIR applications for accurate term recognition and indexing; handheld electronic dictionaries and mobile device applications, and language learning applications
J03	JMP	Multilingual Database of Japanese Points-of-Interest (CJKE)	Proper Nouns	1,172,083	A large-scale database of Japanese place names and POIs in CJK and English languages.	Used for a wide variety of applications, including MT, information retrieval, morphological analysis, electronic dictionaries and mobile device applications, IME systems, and NER.

ID	Code	Resource	Type of database	Headwords	General description	Resource purpose
J04	JPD	Japanese Phonological Database	NLP Lexicons	70,000	Compiled by experienced editors with in-depth knowledge of Japanese phonology and phonetics. Provides IPA phonetic transcriptions (SAMPA on request) that accurately indicate how Japanese words are pronounced in actual speech, as well as accent codes, for each entry. Includes accent information for personal names and place names, making the resource unique. Includes accent information for place names and personal names.	Speech synthesis systems; pedagogical research to help in the acquisition of Japanese as a foreign language; research and development of Japanese speech technology in general.
J05	JLD	Japanese Lexical Database	NLP Lexicons	290,000	Monolingual lexical database with a rich set of grammatical attributes such as derivational attributes, suffixes and prefixes and bound morphemes.	Fine-tuned for NLP applications such as MT, information retrieval and morphological analysis.
J07	JOD	Japanese Orthographical Database	NLP Lexicons	127,600	Our Japanese Orthographical Database (JOD) plays a critical role in enhancing the accuracy of information retrieval, machine translation and morphological analysis applications as it helps identify and disambiguate the numerous Japanese orthographic variants that have identical meanings.	Fine-tuned for NLP applications such as MT, information retrieval and morphological analysis.
J08	JMP	Multilingual Database of Japanese Points-of-Interest (non-CJKE)	Proper Nouns	1,951,518	A large-scale database of Japanese place names and POIs in European languages such as German and French, and other Asian languages like Vietnamese and Indonesian.	Used for a wide variety of applications, including MT, information retrieval, morphological analysis, electronic dictionaries and mobile device applications, IME systems, and NER.
J09	JCD	Japanese Companies and Organizations	Proper Nouns	580,000	Japanese company and organization names with English equivalents when available.	Used for information retrieval and morphological analysis in business intelligence software and machine translation.
J10	JNV	Database of Japanese Name Variants	Proper Nouns	4,000,000	This resource covers four million Japanese names and their romanized variants, and includes gender codes, classification codes, and frequency rankings.	Business intelligence software and machine translation.
		Korean-Japanese Database of Proper Nouns	see K03 in Korean section			
J11	JEN	Japanese – English Database of Proper Nouns	Proper Nouns	660,000	Covers over 660,000 entries and includes various data fields such as hiragana and romanized readings, classification codes and locale codes, English equivalents, and more. Included are a large variety of both Japanese and non-Japanese personal and place names.	Used for a wide variety of applications, including MT, information retrieval, morphological analysis, electronic dictionaries and mobile device applications, IME systems, and NER.
J12	JETERM	Japanese - English Dictionary of Technical Terms	Technical Terms	920,390	Comprehensive Japanese-English bilingual, bidirectional database of technical terms covering a broad spectrum of fields ranging from computer science to business and finance to biotechnology.	MT dictionaries; information retrieval applications for accurate term recognition and indexing; NLP tools like morphological analyzers and tokenizers; handheld electronic dictionaries and mobile device applications.

ID	Code	Resource	Type of database	Headwords	General description	Resource purpose
J13	JWL	Comprehensive Wordlist of Japanese	NLP Lexicons	2,646,853	Comprehensive monolingual wordlist for Japanese. Readings are provided, making this database ideal for speech-related applications such as speech synthesis.	Suitable for a variety of natural language processing applications, including information retrieval, morphological analysis and word segmentation, as well as speech-related applications.
J14	JFULEX	Japanese Full Form Lexicon	NLP Lexicons	120,000,000	This is a comprehensive, full-form lexicon for Japanese general vocabulary in which all inflected forms are included.	Enhanced translation quality for MT and other NLP applications; morphological analysis; named entity recognition and entity extraction; and query processing for information retrieval applications.
		Korean-Japanese Dictionary of Technical Terms	see K01 in Korean section			
		Chinese-Japanese Technical Terms Dictionary	see C06 in Chinese section			
	Korean					
K01	KJTERM	Korean-Japanese Dictionary of Technical Terms	Technical Terms	988,347	Bilingual, bidirectional database of technical terms covering fields including civil engineering, business and finance, mechanical engineering, IT/computer, and more.	MT dictionaries; information retrieval applications for accurate term recognition and indexing; NLP tools like morphological analyzers and tokenizers; handheld electronic dictionaries and mobile device applications
K02	KEN KEP	Korean-English Database of Proper Nouns	Proper Nouns	1,820,200	A large comprehensive database of Korean-English personal and place names, with coverage of not only native Korean proper nouns, but also Chinese, Japanese, and Western proper nouns as well.	Used for a wide variety of applications, including MT, information retrieval, morphological analysis, electronic and mobile platform dictionaries, IME systems, NER, data cleansing, and mapping and geodata.
K03	KJN KJP	Korean-Japanese Database of Proper Nouns	Proper Nouns	2,250,700	A large comprehensive database of Korean-Japanese personal and place names, with coverage of not only native Korean proper nouns, but also Chinese, Japanese, and Western proper nouns as well.	Used for a wide variety of applications, including MT, information retrieval, morphological analysis, electronic and mobile platform dictionaries, IME systems, NER, data cleansing, and mapping and geodata.
K04	KCN KCP	Korean-Chinese Database of Proper Nouns	Proper Nouns	2,483,600	A large comprehensive database of Korean-Chinese personal and place names, with coverage of not only native Korean proper nouns, but also Japanese, Chinese and Western proper nouns as well.	Used for a wide variety of applications, including MT, information retrieval, morphological analysis, electronic and mobile platform dictionaries, IME systems, NER, data cleansing, and mapping and geodata.
K05	KLD	Korean Lexical Database	NLP Lexicons	97,000	Monolingual lexical database which includes a significant number of affixes, particles, auxiliaries and conjugation patterns to account for all the inflectional and derivational morphology in Korean so as to enable recognition of inflected forms.	Fine-tuned for NLP applications such as MT, information retrieval and morphological analysis.

ID	Code	Resource	Type of database	Headwords	General description	Resource purpose
K06	KWL	Comprehensive Wordlist of Korean	NLP Lexicons	1,040,887	Comprehensive monolingual wordlist for Korean. Romanization (IPA optional) is provided, making this database suitable for speech-related applications such as speech synthesis.	Suitable for a variety of natural language processing applications, including information retrieval, morphological analysis and word segmentation, as well as speech-related applications.
K07	KNV	Database of Korean Name Variants	Proper Nouns	183,000	Provides coverage for the major Korean romanization systems, including the latest standard published by the Korean government in 2000.	Used for a wide variety of applications, including MT, information retrieval, morphological analysis, electronic and mobile platform dictionaries, IME systems, NER, data cleansing, and mapping and geodata.
	Arabic					
A01	ArabLEX	Arabic Full-Form Lexicon	NLP Lexicons	530,000,000	This comprehensive database covers inflected, conjugated and cliticized wordforms, and is rich in morphological, grammatical, phonological, and orthographical attributes. In addition, it maps all unvocalized forms to their vocalized counterparts and to the lemma, and provides precise phonemic and phonetic transcriptions.	This database can significantly contribute to the training of language models for speech technology (both synthesis and recognition) and machine translation.
A02	DAP	Database of Arabic Plurals	NLP Lexicons	3,137	This database covers both regular and irregular Arabic plurals, and was developed by experts over a period of several years. The data includes various grammatical attributes such as part-of-speech, collectivity codes, gender codes, and full vocalization.	Used in software development, machine translation, and Arabic language education.
A03	DAN	Database of Arab Names	Proper Nouns	6,500,000	Comprehensive database of Arabic personal names and name variants mapped to the original Arabic script with a large variety of supplementary information.	Suitable for NER, MT, variant normalization, information retrieval of Arabic names, risk compliance systems, and transcription and transliteration.
A04	DANA	Database of Arab Names in Arabic	Proper Nouns	222,000	A resource of Arab personal names and variants, in the original Arabic script, this database covers several hundred thousand Arabic script variants, along with common spelling mistakes. Every Arabic name is normalized and vocalized.	Suitable for NER, MT, variant normalization, and information retrieval of Arabic names, and transcription and transliteration.
A05	DAFNA	Database of Foreign Names in Arabic	Proper Nouns	37,000	This database covers non-Arabic names, their Arabic equivalents, and Arabic script variants for each name.	Suitable for NER, MT, variant normalization, information retrieval of Arabic names, risk compliance systems, and transcription and transliteration.
A06	DAPNA	Database of Arabic Place Names	Proper Nouns	10,000	This is bilingual bidirectional English-Arab dictionary provides worldwide coverage of common place names. It includes both the standard MSA spellings as well as Arabic spelling variants for many place names.	Suitable for NER, MT, variant normalization, and information retrieval of Arabic names, and transcription and transliteration.
A07	AWL	Comprehensive Wordlist of Arabic	NLP Lexicons	210,000	Comprehensive monolingual wordlist for Arabic. Phonemic transcriptions are provided, making this database ideal for speech-related applications such as speech synthesis.	Suitable for a variety of natural language processing applications, including information retrieval, morphological analysis and word segmentation, as well as speech-related applications.

ID	Code	Resource	Type of database	Headwords	General description	Resource purpose
A08	DiaLEX	Arabic Dialects Full-Form Lexicon	NLP Lexicons	100,000,000	Comprehensive computational lexicon covering several major Arabic dialects and subdialects, including Egyptian, Kuwaiti, Qatari, Emirati, Saudi Arabian Najdi, Saudi Arabian Hejazi, and Palestinian.	Suitable for a variety of natural language processing applications, including information retrieval, morphological analysis and word segmentation, as well as speech-related applications.
	Multilingual					
		Multilingual Database of Japanese Points-of-Interest (POIs)	see J03 in Japanese section			
M02		Multilingual Proper Noun Database	Proper Nouns	150,000	Brings together six languages -- Simplified Chinese, Traditional Chinese, Japanese, Korean, English (Arabic upon request) -- in a multidirectional format. The database includes various data fields, such as readings in pinyin and zhuyin, hiragana, romanization in all major and most minor romanization systems, semantic classification codes, locale codes, and other useful information.	Suitable for a wide variety of applications such as online multilingual maps, NER, MT, and information retrieval.
	Others					
X01	DPN	Database of Persian Names	Proper Nouns	450,000	A unique resource that has been developed in cooperation with a team of native-speaker experts in Persian phonology. The data includes a confidence rank to indicate the relative likelihood that a variant will be encountered in the real world.	Suitable for NER, MT, variant normalization, information retrieval of Persian names, risk compliance systems, and transcription and transliteration.
X02	SFULEX	Spanish Full-form Lexicon (Monolingual)	NLP Lexicons	1,000,000	This is an extremely comprehensive Spanish full-form lexicon for general vocabulary in which all forms, including inflected, plural, feminine and affixed forms, are included.	Enhanced translation quality for MT and other NLP applications; morphological analysis; named entity recognition and entity extraction; and query processing for information retrieval applications.
X03	SFULEX	Spanish Full-form Lexicon (Bilingual)	NLP Lexicons	26,000,000	This is an extremely comprehensive Spanish-English lexicon for general vocabulary in which not only are all forms, including inflected, plural, feminine and affixed forms included, but all English equivalents for each of these forms is given as well.	Enhanced translation quality for MT and other NLP applications; morphological analysis; named entity recognition and entity extraction; and query processing for information retrieval applications.
X04		Vietnamese - Japanese Dictionary	General Vocabulary	140,000	This database covers general vocabulary, and includes part-of-speech codes and readings. Additionally, Chinese characters are given for Sino-Vietnamese compounds.	MT dictionaries; handheld electronic dictionaries and mobile device applications, and language learning applications
Corpora						
M1		Korean - Chinese - Japanese	Corpora	85,000	Original sentences created in Korean by a native speaker, then translated into Chinese and Japanese by a Korean translator. Domain of the corpus is travel, a dialog set between two people.	Suitable for a variety of natural language processing applications, but especially for morphological analysis and the training of machine translation systems.

ID	Code	Resource	Type of database	Headwords	General description	Resource purpose
M2		Korean - Chinese - English	Corpora	510,000	Original sentences created in Korean by a native speaker, then translated into Chinese and English by a Korean translator.	Suitable for a variety of natural language processing applications, but especially for morphological analysis and the training of machine translation systems.
M3		Korean - Chinese	Corpora	1,620,000	Original sentence pairs were created as follows: 600,000 pairs were created in Korean by a native speaker and translated to Chinese by a Korean translator. 900,000 pairs were collected from textbooks and websites. Note that the KC pairs in the KCE and KCJ corpora are not included in this KC corpus.	Suitable for a variety of natural language processing applications, but especially for morphological analysis and the training of machine translation systems.
M4		English - Chinese	Corpora	4,000,000	Sentence pairs were collected from online and offline English textbooks and English learning websites. Original sentences were created in English by Chinese native speakers and translated into Chinese by Chinese translators.	Suitable for a variety of natural language processing applications, but especially for morphological analysis and the training of machine translation systems.
M5		Korean	Corpora	9,000,000	Sentences were collected from online and offline sources such as textbooks, newspapers. Original sentences were created in Korean by Korean native speakers.	Suitable for a variety of natural language processing applications, but especially for morphological analysis and the training of machine translation systems.
M6-1		Korean - Vietnamese	Corpora	350,000	Sentences of the KV corpus created by Korean native speakers and translated to Vietnamese by Vietnamese translators.	Suitable for a variety of natural language processing applications, but especially for morphological analysis and the training of machine translation systems.
M6-2		Vietnamese - Korean	Corpora	240,000	Sentences of the VK corpus created by Vietnamese native speakers and translated to Korean by a Vietnamese translator.	Suitable for a variety of natural language processing applications, but especially for morphological analysis and the training of machine translation systems.
M7		Korean - English	Corpora	2,250,000	Sentence pairs were collected from online and offline English textbooks and English learning websites. Original sentences were created in Korean by Korean native speakers and translated into English by Korean translators.	Suitable for a variety of natural language processing applications, but especially for morphological analysis and the training of machine translation systems.
M8		English - Korean	Corpora	2,510,000	Sentence pairs were collected from online and offline English textbooks and English learning websites. Original sentences were created in English by Korean native speakers and translated into Korean by Korean translators.	Suitable for a variety of natural language processing applications, but especially for morphological analysis and the training of machine translation systems.
M9		Korean - Japanese	Corpora	800,000	The 800,000 sentence pairs were collected from online and offline Japanese textbooks, learning websites and news sites in Korea. The original sentences were created in Korean and translated into Japanese by Korean native speakers. The corpus covers various domains such as science and technology, English conversation, and miscellaneous subjects.	Suitable for a variety of natural language processing applications, but especially for morphological analysis and the training of machine translation systems.

ID	Code	Resource	Type of database	Headwords	General description	Resource purpose
VEC		Vietnamese- English	Corpora	27,500	Bilingual corpus of high quality created by our partner, a lexicographic institute in Hanoi. This can be used bidirectionally.	Suitable for a variety of natural language processing applications, but especially for morphological analysis and the training of machine translation systems.
EVC		English - Vietnamese	Corpora	73,000	Bilingual corpus of high quality created by our partner, a lexicographic institute in Hanoi. This can be used bidirectionally.	Suitable for a variety of natural language processing applications, but especially for morphological analysis and the training of machine translation systems.
VMC		Vietnamese Monolingual Corpus	Corpora	1,000,000	Monolingual corpus based on contemporary Vietnamese sources.	Suitable for a variety of natural language processing applications, but especially for morphological analysis and the training of machine translation systems.