CIYA SOFT

Product Description

CiyaSoft MT Team
Initial Version 11/2002
Latest update: 9/2018

9102: Most-frequently used words and phrases. This database contains Farsi and Dari entries with the frequency of usage and include meanings, contexts, parts of speech and other elements. (Note: 9102 is a subset of another CiyaSoft product, namely, 9104 which contains over 1,740,000 Farsi words and phrases with meanings in English along with other attributes optimized for CiyaSoft's Machine Translation).
The size of this database is about 2,850,000 and was created by analysis of millions of electronic Farsi contemporary documents that have been collected during the past 20 years from newspapers, magazines, books, emails and the WEB. They are ordered by rank (the most frequent appearing first) and include frequency.
The database consists of three modules, each module being a subset of the other:

A. (9102A) Low-frequency words and phrases: 650,000 entries
B. (9102B) Medium-frequency words and phrases: 1,200,000 entries
C. (9102C) High-frequency words and phrases: 650,000 entries

9108: Proper nouns, Farsi to English. This database contains over 1,000,000 proper names from Western origins, Farsi- and Pashtu- and Arabic-speaking countries. *Pinglish* is provided in the Farsi fields when the source is not Farsi or Arabic, and *Finglish* is provided in the English fields when the source is either Farsi or Arabic. The Pinglish and Finglish have been manually entered and reviewed and there may be more than one equivalent for each entry, as in examples bellow:

Mohammad, Muhammd, Mohammed, … محمد	
Jack	ژاک، جک

The database consists of several modules based on the source (Farsi, Arabic, …), and nature (proper name, geographical locations, companies, names of buildings, ports, cities, airlines, etc.).

A. Proper names of Western origin with Finglish

B. All proper nouns of Western origin with Finglish
C. All proper nouns with Finglish, or Pinglish (whichever appropriate)

9302: Domain-Specific Databases. This database used to be in 30 separate modules covering about 80 different domains. It is now consolidated into 18 databases covering the following domains:

Accounting	Aeronautics	Agriculture	Air Force	Anatomy	Architecture
Atmospherics	Biochemistry	Genealogy	Metallurgy	Sociology	Insurance
Civil Engineering	Biology	Geography	Navy	Surgery	Management
Communications	Botany	Economics	Optometry	Textile	Mathematics
Computer Science	Chemistry	Dentistry	Painting	Theater	Oil Engineering
Industrial Engineering	Cinema	Criminology	Paleontology	Trade	Pharmacology
International Law	Astronomy	Medicine	Physiology	Trigonometry	Space Sciences
Mechanical Engineering	Electronics	Meteorology	Science	Veterinary	Logic
Electrical Engineering	Gardening	Microbiology	Psychiatry	Zoology	Humanities
Genetic Engineering	Mining	Mythology	Psychology	Physics	Law
Military Sciences	Music	Navigation	Religions	Philosophy	Literature
Telecommunications	Poetry	Political	Sculpture	Petrology	
Grammar	Geometry	Geology	Arts	History	

More than 60% of this database was created by review of over 200 domain-specific dictionaries, selecting MT-useable words, and phrases and acronyms and optimizing for MT. The remaining 40% was compiled from technical and scientific sources. The database includes the parts of speech, frequency, specific (coded) domain, as well as one or more equivalents. It contains a total of about 1,200,000 entries. Each entry has been manually reviewed at several revisions by linguistic engineers for each given domain and then reviewed for optimization for machine translation. The database is not merely the contents of the dictionaries used. Each equivalent has been changed, if necessary, to contain the needed information used by MT engines during various phases of processing. For example, in the domain of Botany, there are several plant names, which may not have an equivalent in the target language. Such words are defined in the dictionaries in a way that cannot be used by MT if they appear in a sentence as a noun and therefore have been modified accordingly. Other examples of optimization for MT use include changes in defining compound verbs and forward adjectives which are transformed in the early stages of preprocessing into forms that can be looked up in the databases. In short, these databases are not dictionaries, or encyclopedias (as some dictionaries tend to be), they do not just contain the meanings; rather, they include additional information to help MT engines.

A. All 18 databases including all the fields
B. Same as A, but with meaning fields only
C. Same as B, but only including Physics (which includes Nuclear Science), all Engineering domains and all military-related domains

The ratio of word count to phrase count (proper nouns excluded) is approximately 1, i.e. the number of phrases equals the number of words. The definition of a phrase in MT databases is any entry in which there are two or more words separated by spaces or hyphens. Here are some samples from 9102:

ابطال
فسخ‌شده
مشبك
سرطان
استعداد رشد سرطان
خسارت ابطال
مدار فسخ‌کننده
قضیه ابطال
ابطال وصیت‌نامه
اقاله کردن معامله
ابطال بدهی
الغاء توقیف مال

Samples from Physics database of 9302:

cross-lap joint	اتصال نیم‌نیم
crossbar switch	کلید تقاطعی
crosstalk	تداخل سیگنال‌ها
cryogenic memory	حافظه سرمایشی
cryogenics	فیزیك سرما
cryoscope	کریوسکپ
cryostat	کریوستات
cryotron	کریوترون
cryptogram	کریپتوگرام
cryptographic	کریپتوگرافیك
cryptographic algorithm	الگوریتم کریپتوگرافیکی
crystallite	بلورك
current drain	کشش جریان
current instruction register	ثبات دستور جاری
current pulse	ضربه جریان
cursor	مکان‌نما
curve fitting	برازاندن خم