Moby Project

The **Moby Project** is a collection of public-domain lexical resources created by Grady Ward. The resources were dedicated to the public domain, and are now mirrored at Project Gutenberg. As of 2007, it contains the largest free phonetic database, with 177,267 words and corresponding pronunciations.[1]

Hyphenator

The **Moby Hyphenator II** contains hyphenations of 187,175 words and phrases (including 9,752 entries where no hyphenations are given, such as *through* and *avoir*). The character encoding appears to be MacRoman, and hyphenation is indicated by a bullet (⟨•⟩, character value 165 decimal, or A5 hexadecimal). Some entries, however, have a combination of actual hyphens and character 165, such as "bar•ber-sur•geon".

There is little to no documentation of the hyphenation choices made; the following examples might give some flavour of the style of hyphenation used: at•mos•phere; at•tend•ant; ca•pac•i•ty; un•col•or•a•ble.

Languages

Moby Language II contains wordlists of five languages: French, German, Italian, Japanese, and Spanish. Their statistics are:

Language	Words	Size (in bytes)
French	138,257	1,524,757
German	159,809	2,055,986
Italian	60,453	561,981
Japanese	115,523	934,783
Spanish	86,059	850,523
Total	560,101	**5,928,030**

However, some of the lists are contaminated: for example, the Japanese list contains English words such as *abnormal* and non-words such as *abcdefgh* and *m,./*. There are also unusual peculiarities in the sorting of these lists, as the French list contains a straight alphabetical listing, while the German list contains the alphabetical listing of traditionally capitalized words and then the alphabetical listing of traditionally lower-cased words. The list of Italian words, however, contains no capitalized words whatsoever.

The lists do not use accented characters, so "e^tre" is how a user would look up the French word *être* ("to be").

Part-of-Speech

Moby Part-of-Speech contains 233,356 words fully described by part(s) of speech, listed in priority order. The format of the file is *word\parts-of-speech*, with the following parts of speech being identified:

Part-of-speech	Code
Noun	N
Plural	p
Noun phrase	h
Verb (usually participle)	V
Transitive verb	t
Intransitive verb	i
Adjective	A
Adverb	v
Conjunction	C
Preposition	P
Interjection	!
Pronoun	r
Definite article	D
Indefinite article	I
Nominative	o

Pronunciator

The **Moby Pronunciator II** contains 177,267 entries with corresponding pronunciations. Most of the entries describe a single word, but approximately 79,000[2] contain hyphenated or multiple word phrases, names, or lexemes. The Project Gutenberg distribution also contains a copy of the cmudict v0.3. The file contains lines of the format *word[/part-of-speech] pronunciation*. Each line is ended with the ASCII carriage return character (CR, '\r', 0x0D, 13 in decimal).

The *word* field can include apostrophes (e.g. *isn't*), hyphens (e.g. *able-bodied*), and multiple words separated by underscores (e.g. *monkey_wrench*). Non-English words are generally rendered, as stated in the documentation, without accents or other diacritical marks. However, in 36 entries (e.g. *São_Miguel*), some non-ASCII accented characters remain, represented using Mac OS Roman encoding.

The part-of-speech field is used to disambiguate 770 of the words which have differing pronunciations depending on their part-of-speech. For example, for the words spelled *close*, the verb has the pronunciation /ˈkloʊz/, whereas the adjective is /ˈkloʊs/. The parts-of-speech have been assigned the following codes:

Part-of-speech	Code
Noun	n
Verb	v
Adjective	aj
Adverb	av
Interjection	interj

Following this is the pronunciation. Several special symbols are present:

Symbol	Meaning
_	Used to separate words
'	Primary stress on the following syllable
,	Secondary stress on the following syllable

The rest of the symbols are used to represent IPA characters. The pronunciations are generally consistent with a General American dialect of English, that exhibits father-bother merger, hurry-furry merger and lot-cloth split, but does not exhibit cot-caught merger or wine-whine merger. Each phoneme is represented by a sequence of one or more characters. Some of the sequences are delimited with a slash character "/", as shown in the following table, but note that the sequence for /ɔɪ/ is delimited by *two* slash characters at either end:

Symbol	IPA
/&/	æ
/-/	ə
/@/	ʌ, ə
/[@]/r	ɜr, ər
/A/	ɑ, ɑː
/aI/	aɪ
/AU/	aʊ
b	b
d	d
/D/	ð
/dZ/	dʒ
/E/	ɛ
/eI/	eɪ
f	f
g	g
h	h
hw	hw
/i/	iː
/I/	ɪ
/j/	j
/ju/	juː
k	k
l	l
m	m
n	n
/N/	ŋ
/O/	ɔ, ɔː
//Oi//	ɔɪ
/oU/	oʊ
p	p
r	r
s	s
/S/	ʃ
t	t

/T/	θ
/tS/	tʃ
/u/	uː
/U/	ʊ
v	v
w	w
z	z
/Z/	ʒ

To this collection are added a number of extra sequences representing phonemes found in several other languages. These are used to encode the non-English words, phrases and names that are included in the database. The following table contains these extra phonemes, but note that the extent to which some of these may exist due to encoding errors is not clear.

Symbol	IPA
A	a
e	e, ɛ
i	i, ɪ
N	Nasalisation of preceding vowel
o	o
O	[intent not clear]
R	ʁ
S	s
u	u
V	ʌ, β, ʊ
W	w
/x/	x
/y/	ø
Y	y
/z/	ts
Z	z

Shakespeare

Moby Shakespeare contains the complete unabridged works of Shakespeare. This specific resource is not available from Project Gutenberg, but it is available in a 1993 version on the web.[3]

Thesaurus

The **Moby Thesaurus II** contains 30,260 root words, with 2,520,264 synonyms and related terms – an average of 83.3 per root word. Each line consists of a list of comma-separated values, with the first term being the root word, and all following words being related terms.

Grady Ward placed this thesaurus in the public domain in 1996. It is also available as a Debian package although the package has been discontinued starting with Bullseye.[4]

Words

Moby Words II is the largest wordlist in the world.[1] The distribution consists of the following 16 files:

Filename	Words	Description
ACRONYMS.TXT	6,213	Common acronyms and abbreviations
COMMON.TXT	74,550	Common words present in two or more published dictionaries
COMPOUND.TXT	256,772	Phrases, proper nouns, and acronyms not included in the common words file
CROSSWD.TXT	113,809	Words included in the first edition of the Official Scrabble Players Dictionary
CRSWD-D.TXT	4,160	Additions to the Official Scrabble Players Dictionary in the second edition
FICTION.TXT	467	A list of the most commonly occurring substrings in the book *The Joy Luck Club*
FREQ.TXT	1,000	Most frequently occurring words in the English language, listed in descending order
FREQ-INT.TXT	1,000	Most frequently occurring words on Usenet in 1992, listed with corresponding percentage in decreasing order
KJVFREQ.TXT	1,185	Most frequently occurring substrings in the King James Version of the Bible, listed in descending order
NAMES.TXT	21,986	Most common names used in the United States and Great Britain
NAMES-F.TXT	4,946	Common English female names
NAMES-M.TXT	3,897	Common English male names
OFTENMIS.TXT	366	Most common misspelled English words
PLACES.TXT	10,196	Place names in the United States
SINGLE.TXT	354,984	Single words excluding proper nouns, acronyms, compound words and phrases, but including archaic words and significant variant spellings
USACONST.TXT	7,618	United States Constitution including all amendments current to 1993
Total	**863,149**	Not the total of unique words.
Total Uniq	**639,995**	Total of single, proper nouns, acronyms, and compound words and phrases (all of the files that contain unique words).

References

1. "ACL SIGLEX Resource Links" (https://web.archive.org/web/20181215174820/https://www.clres.com/dict.html). Special Interest Group on the Lexicon of the Association for Computational Linguistics. August 13, 2004. Archived from the original (https://www.clres.com/dict.html) on December 15, 2018. Retrieved May 9, 2022. "Moby Words: 610,000+ words and phrases. The largest word list in the world"
2. Obtained by running the UNIX command *grep '.*[-_].* .*' mobypron.unc* | *wc -l* after converting the line endings and correcting some encoding errors.
3. mobyshak.txt 1993 version (http://shakespearereadingsociety.co.uk/texts/1993originals/mobyshak.txt)
4. Tosi, Sandro (July 13, 2020). "RM: dict-moby-thesaurus -- RoQA; dead upstream (10+ years); python2-only; no extrenal [*sic*] deps; extremely low popcon" (https://bugs.debian.org/cgi-bin/bugreport.cgi?bug=964991). *Debian Bug report logs*. Retrieved May 10, 2022.

External links

- Former Moby Project site (icon.shef.ac.uk/Moby/) – No longer accessible. View a copy (https://web.archive.org/web/20170930060409/http://icon.shef.ac.uk/Moby/) made by the Wayback Machine, as it was on 30 September 2017. ("Last modified: October 24, 2000") working download site (http://ai1.ai.uga.edu/ftplib/natural-language/moby/).
- Project Gutenberg downloads (http://www.gutenberg.org/ebooks/3201)
- *Searching for Rhymes with Perl (http://www.foo.be/docs/tpj/issues/vol4_4/tpj0404-0003.html)*; corresponding code (http://interglacial.com/~sburke/mpron/)
- Wiktionary:Appendix:Moby Thesaurus II
- http://digital.library.upenn.edu/webbin/gutbook/lookup?num=3201