



Hyphenators — language specifications

Languages:

Our hyphenators are not based on a hyphenated dictionary data base.

New hyphenator languages: **Hebrew, Irish/Gaelic** (see Windows Unicode Demo).

Recent updated hyphenator languages: **Slovak, Russian, Bulgarian, Romanian**.

66 language modules

Dutch (Update March 2007, ε 0.0044 ‰) (3x)

supports the generally accepted spelling (the Netherlands), progressive spelling (Belgium), and the 1996 & October 2005 spelling reforms — four principles have been integrated in one hyphenator. Support the **Belgium** (Flemish), **Surinam** and **Dutch** idiom. The hyphenator recognizes compound boundaries and covers the Dutch idiom in the most extensive way. The Dutch module includes support for the **Dutch, Flemish** and **Surinam** varieties.

English (Update August 2007, ε 0.0098 ‰) (6x)

supports phonetical hyphenation according to the world's most trusted dictionaries: Webster's Third New International Dictionary, Webster's New Twentieth Century Unabridged Dictionary (2nd edition), and Longman's Dictionary of Contemporary English; based on an unabridged learning corpus, coming close in size to Webster's Unabridged Dictionaries; a common hyphenator is available for the British, Canadian, and American idiom. The hyphenator solves the irregularity of the alternation of English strong and weak syllables. The new double layer model enables the user to disregard certain secondary divisions (adapt~able instead of adapt~a~ble). Hyphenation agrees with The Oxford Colour Spelling Dictionary (1995). Compared to the other dictionaries, this last dictionary has fewer syllables.

The English module has separate entries for

British English (1),

American English (2),

Canadian English (3),

Australian English (4),*

New-Zealand English (5),*

South-African English (6).*

* Br. English varieties comes with the British English module (1).

German old and new (Update August 2007, ε 0.0045 ‰) (2x)

support every characteristic German hyphenation according to the most recent Duden Rechtschreibung August 2004, e.g., Stilleben -> Still-leben; however, these cases are no longer present in the new spelling reform, die neue Rechtschreibung 1998, but still have to be supported up to 2005. The German hyphenator recognizes compound boundaries independent of the spelling reform. The new feature for “der Verwendung von Großbuchstaben SS für ß” correctly hyphenates both “Schreibungsweisen”. A special effort has been made to support medical and other scientific domains.

Swiss German old and new (Update August, 2007) (2x)

respond accurately to the typical Swiss German deviations and local idiom (including the β to ss transcription).

French (two versions, Update November 2004) (2x)

accepts etymological syllabification according to Grevisse's "le bon usage." A second version accepts phonetical hyphenation rules recommended by the leading French linguist Nina Catach in Paris. Both French versions use the new double layer technique to enable or disable hyphenation of muettes.

Canadian French (two versions, Update November 2004) (2x)

accepts etymological syllabification according to Grevisse's "le bon usage." A second version accepts phonetical hyphenation rules recommended by the leading French linguist Nina Catach in Paris. Both Canadian French versions use the new double layer technique to enable or disable hyphenation of muettes.

Spanish (Update January 2006, ϵ 0.0035 ‰) (2x)

supports the official hyphenation rules as published by large dictionary publishers; completely covers the Spanish and Latin American idiom. Mexican Spanish is included.

Italian (Update June 2006, ϵ 0.0008 ‰)

supports phonetical hyphenation, in Italian: "la sillabazione: basata prevalentemente sul criterio di tenere uniti i gruppi consonantici attestati, anche una sola volta, come iniziale di parola". In addition the new hyphenator handles hiatuses accurately, elisions (al-l'I.ta-lia), conjugations, declensions, and words that came from English and other foreign languages (beat-nik and not be-at-nik).

Iberian Portuguese (Update January 2006, ϵ 0.007 ‰) (2x)

based on the vowel as the syllabic unit, but falling diphthongs and final diphthongs are kept together. Doubling of the hyphen is supported (repetir o hífen na linha sequinte).

Brazilian Portuguese (Update January 2006, ϵ 0.007 ‰) (2x)

based on the vowel as the syllabic unit, but falling diphthongs and final diphthongs are kept together. Doubling of the hyphen is supported (repetir o hífen na linha sequinte).

Czech (Update January 2005)

supports the reformed spelling. As is the case in every Slavic language, a number of additive vowels and consonants exists, which have a large impact on hyphenation. Syllables that solely consist of consonants are supported (ji-tr-nice).

Slovak (Update June 2007)

supports the standard Slovak orthography. As is the case in every Slavic language, a number of additive vowels and consonants exists, which have a large impact on hyphenation. Syllables that solely consist of consonants are supported (ji-tr-nice).

Swedish (Update November 2005) (2x)

accepts the mekaniska principen, but compounded words are divided into their morphological roots. An overwhelming occurrence of compounds, and newly created forms, makes it a challenge worth accepting. You can switch between c-k or ck- hyphenation, and between within-word vowel-vowel hyphenation.

Finnish (Update January 2006)

is tuned to the peculiarities of the Finnish language and shares attributes with all Finno-Ugric languages. It has a rich structure, including a large number of falling and rising diphthongs. The phonetical base of the syllable is accepted, here, fully hyphenated despite it's overwhelming inflection structure. You may find its resemblance to

the neighboring Estonian remarkable.

Catalan (Update August 2007)

supports the mixed French and Spanish origins of the Catalan language. A peculiarity of Catalan, needing special care, is the l geminada (l·l).

Danish (Update August 2000)

accepts the hyphenation rules of the Dansk Sprognævns Retskrivningsordbog. Compounds and newly created forms are supported; please note that it even hyphenates Norwegian according to consonant rules.

Norwegian (Update April 2007, ε 0.0067 ‰) (4x)

accepts consonant rules (20) or the morphological rules of the Nordisk institutt of the University of Bergen(21).

Nynorsk (Update April 2007, ε 0.0067 ‰) (4x)

accepts consonant rules (20) or the morphological rules of the Nordisk institutt of the University of Bergen(21).

Icelandic (Update Januari 2006, ε 0.021 ‰)

accepts morphological rules which separate the attached article and nominative, dative, accusative, and genitive cases and is capable of dividing a pileup of compounds.

Estonian (January 1999)

behaves like the Finnish hyphenator and is capable of correctly hyphenating Estonian compounds and diphthongs. However, there are more diphthongs in the Estonian language than in the Finnish language which increases complexity.

New Greek (Update January 2005)

is tuned in to the Greek script, the Elot codepage or Unicode. It hyphenates more than between alpha and omega — not just the beginning and the end (Classical Greek), but a new era in progress (Modern Greek). Present-day Greek has evolved and is flourishing with diacritics.

Polish (February 1999)

hyphenation of the Polish language is hindered by an immense number of consonants, quite often unpronounceable for non-Polish speakers. However, the hyphenator has been fully adapted to these difficult syllables.

Latvian (August 1999)

is tuned to the properties of Baltic languages. Words are richly declined. Latvian uses additional consonants and vowels, which are recognized by the hyphenator.

Azerbaijani (August 1999)

is one of the new Transcaucasian republics that are now independent from the former USSR. Azerbaijani is related to Turkish. The Azerbaijani now use a Latin script. There is no standard script yet, but it does not violate *TALÖ's hyphenator principles.

Turkish (Update April 2006)

Present-day Turkish is spoken in SW Asia, but in earlier times the Turkish region reached into the north of China. In Chinese history, the name Tu-kiu was mentioned 600 years ago. Turkish is characterized by a lot of additive particles that change the meaning of a word. A word can take numerous forms and different parallel hyphenations.

Lithuanian (September 1999)

is one of the Baltic languages which is richly declined. The (semi-)diphthongs, palatals, and affricates have been taken into consideration for hyphenation.

Afrikaans (Update July 2007)

the Afrikaans language evolved from 17th-century Dutch and is an official language of South Africa. Its hyphenation has much in common with the Dutch language. Afrikanization of spelling has given the Afrikaans language its own identity. The Afrikaans hyphenator takes all Afrikaans peculiarities into consideration, including diaeresis hyphenation.

Russian (Update July 2007)

accepts Cyrillic characters, but does not complicate hyphenation. It is the nature of the Russian language: an abundance of prefixes and suffixes, modifying different moods in a fine gradation.

Basque (Update October 2001)

the Basque language is one of Europe's most exotic minority languages, probably unrelated to any other language in the world. The Basque hyphenator is tuned in to all those peculiarities of real-life language.

visit [download page](#) | view a Basque/Euskara example (PDF)

Hungarian (Update April 2006)

the Hungarian language has lost many of its Uralic characteristics and many words have been borrowed from the Turkish and European languages. The language is flavoured with compounds and special hyphenations (briddzsel -> bridsz-dszel).

Bahasa Indonesia (Update June 2005)

the Bahasa Indonesia (Standard Indonesian) is an Austronesian language full of prefixes, suffixes, infixes, in general terms affixes including large classes of sound changes. Hyphenation is inextricably tied to meaning, even when the boundaries are masked by sound changes (mengarang from meng + karang) hyphenation is affected.

Bahasa Melayu (Update June 2005)

what counts for Bahasa Indonesia applies as well to Bahasa Melayu.

Byelorussian (Update July 2007)

is the language of the new nation of Belarus. It was proclaimed the country's sole official language, but Russian remains dominant. Byelorussian is written in the Cyrillic alphabet.

Bulgarian (Update July 2007)

is spoken by 90 % of the population of Bulgaria, 7 million people. Modern Bulgarian alphabet is the same as the Russian alphabet.

Serbian (July 2004)

or srpski jezik is written in the Cyrillic alphabet. Serbian is closely related to Croatian, however, Serbian characters are written with single symbols Ы, ЈБ, and Ѓ. (Dž, Lj, Nj). Like words in any Slavic language Serbian words can have many prefixes to be hyphenated.

Galician (January 2002)

is now spoken in Spanish Galicia, situated north of Portugal. It is a Romance language related to Portuguese. The orthography differs slightly from Spanish.

Rhaeto-Romance (Februari 2002)

is the collective for three Romance dialects spoken in the northeastern Italy and southeastern Switzerland.

Greenlandic (April 2002)

is an Eskimo language spoken in Greenland. Greenlandic is written in the Latin alphabet. Words can be very long and one word can be a complete sentence.

Ukrainian (Update July 2007)

is the national language of Ukraine. It is spoken by a population of 35 million people. Ukrainian has many Polish loan words, but the influences of Russian can be found in the east of Ukraine too.

Romanian (Update August 2007)

is the national language of Romania. It is a Romance language written in the latin script. One third of all Romanian words are of French origin.

Croatian (July 2004)

or hrvatski jezik is written in the Latin alphabet. Croatian is closely related to Serbian. Croatian includes a few digraphs which sound like a single consonant (Dž, Lj, Nj). Like words in any Slavic language Croatian words can have many prefixes to be hyphenated.

Bosnian (July 2004)

or Bosanski Jezik exists since Bosnia & Herzegovina became independent. Bosnian has developed its own identity, written in Latin and closely related to Croatian.

Frisian (April 2002)

or Frysk is spoken in Friesland the northernmost province of The Netherlands. Frisian is closer related to English than Dutch.

Tagalog/Pilipino (April 2002)

is the national language of the Philippines. Three centuries of Spanish rule left a strong imprint on the vocabulary. The pre-, in- and suffixes to modify word meaning make hyphenation irregular.

Slovene (Update April 2006)

or Slovenski jezik is written in the Latin alphabet. Slovene includes a few digraphs (Dž, Lj, Nj). Slovene has many prefixes and inflections. Some syllables divide consonants only: hm-kniti, kr-tina, tr-den.

Thai (Update December 2004)

The Thai people build sentences in a different way. Therefore, the Thai module is not a hyphenator in the traditional sense, but it is a word segmentation tool, that takes context into consideration.

Macedonian (February 2004)

is the principal language of the new nation of Macedonia, it is closely related to Bulgarian and written in the Cyrillic alphabet.

Maltese (Update January 2006)

is one of the official languages of the islands of Malta, it is a Semitic language written in the Latin alphabet, including <ċ> <ħ> <ġ> and <ż>, the variety of root words has a great impact on hyphenation.

Saami (Update April 2006)

hyphenation agrees with the Nord Saami language as spoken in Finnmark county in the north of Norway.

Hebrew (December 2006)

is written in Hebrew consonants only and therefore hyphenation is partially uncertain. Within this uncertainty the hyphenator accepts graphical hyphenations.

Irish/Gaelic (December 2006)

is a Celtic language mainly spoken in Ireland.

Under development

Faeroese, and Esperanto.

September, 2007

*TALO by, Bussum, The Netherlands