



Spelling Checkers — language specifications

Languages and sizes of dictionaries

New languages: **Arab, Azerbaijani, Hebrew, Farsi, Urdu, Breton**; under development Hindi and Marathi

83 languages varieties

English (lexicon size between 310,000 and 315,000, selection September 2007) (5x)

The English language is divided in the following sublanguages

American English (1),

British English (2),

Canadian English (3),

South-African English (4),*

Australian/New-Zealand English (5)*

* Br. English varieties come with the British English module (2).

versions include a set of collocations and automatic respelling functions between American English, Canadian English, and British English orthographical varieties. The supplied idiom includes an extensive medical, chemical, social and geographical lexicon. Finally the idiom includes an extensive orthographical variety of building compounds.

French (lexicon size over 550,000, selection June 2005) (2x)

Includes the most extensive geographical lexicon. Two lexicons are available, one according to the spelling of the Le Nouveau Petit Robert (2003) and one according to the most recent Rectifications de l'orthographe of the Conseil supérieur de la langue française first published 6 December 1990 (see also <http://www.orthographe-recommandee.info>). Includes extensive re-spelling tools between previous and new spelling forms.

Canadian French (lexicon size over 550,000, selection June 2005) (2x)

Includes French Canadianisms and the most extensive geographical lexicon, see French.

German (lexicon size 1070,000, selection October 2007) (3x)

The German spelling has again been reformed in 2006. Previous versions are kept available for a while, but the regular German spelling is distributed in **three** versions, “alt, neu, dpa (2007)”, including automatic respelling from old to new spelling forms (e.g., Prozeß → Prozess) and spelling of “feste grammatische und lexikale Wendungen”. If you prefer “die alte Rechtschreibung” and wish to purify your texts, a full re-spelling system from new to old will surprise you (e.g., Prozess → Prozeß). A version for the Nachrichtenagenturen (dpa) as proposed by the German-speaking news agencies is also available. (<http://www.die-nachrichtenagenturen.de>). Spelling neue Rechtschreibung in agreement with the Duden 24 2006, und IDS Sprach Report, July 2006.

Swiss German (lexicon size 1085,000, Swiss additions to German) (3x)

There are **three** versions “alt, neu, dpa/SDA (2007/8)” see German.

Austrian German (lexicon size 1090,000, Austrian additions to German) (3x)

There are **three** versions “alt, neu, dpa (2007)” see German.

Spanish (lexicon size 855,000, selection July 2006)

The spelling is according to “Gran Diccionario de la lengua española” and “Diccionario Real Academia Española”, 2001.

Includes respelling of a set of common errors, e.g. Adam y Eva → Adán y Eva, Edinburgo → Edimburgo.

Italian (lexicon size 575,000, selection April 2007)

The spelling is according Lo Zingarelli 2006. Includes an extensive geographical lexicon (comuni e luoghi italiani).

Swedish (lexicon size over 1.75 million words, selection January 2005)

Includes geographical and proper names orthography according to Svenska Akademiens ordlista över svenska språket.

Portuguese (lexicon size 460,000, selection January 2006) (2x)

There are two versions, which include respelling between Iberian Portuguese and Brazilian Portuguese, e.g. equipolente versus equípolente or boleia versus boléia.

Dutch (Nederlands, lexicon size 655,000, selection January 2007, 120,000 new words added) (3x)

The spelling according to the governmental rules (Groene Boekje, Workgroup Spelling, 2005, Taalunie) and in agreement with Van Dale Groot Woordenboek van de Nederlandse Taal (XIV ed.).

The lexicon’s idiom covers national and mondial geographic information, medical, administrative, social and many other special terms. A set of collocations and respelling from old to new orthography is included.

Two varieties of the lexicons *1995 orthography* are included.

Flemish (Vlaams, lexicon size 660,000, selection January 2007, 120,000 new words added) (3x)

The spelling according to the governmental rules (Groene Boekje, Workgroup Spelling, 2005, Taalunie) and agrees with Van Dale Groot Woordenboek van de Nederlandse Taal (XIV ed.).

The lexicon’s idiom covers national and mondial geographic information, medical, administrative, social and many other special terms. A set of collocations and respelling from old to new orthography is included.

Two varieties of the lexicons *1995 orthography* are included.

Surinam Dutch (Surinaams-Nederlands, lexicon size 655,000, selection January 2007, with 120,000 new words added)

The Republic of Surinam has entered the Dutch Taalunie (January 2005) to unify their language with the Dutch language. The peculiarities of Surinam Dutch call for a separate lexicon. The spelling agrees with the governmental rules (Groene Boekje, Workgroup Spelling, 2005, Taalunie).

The lexicon’s idiom covers national and mondial geographic information, medical, administrative, social and many other special terms. A set of collocations and respelling from old to new orthography is included.

Catalan (lexicon size 470,000, selection December 2003)

The spelling agrees with Diccionari ortogràfic i de pronúncia, Enciclopèdia Catalana.

Danish (lexicon size 800,000, selection October 2007)

The spelling agrees with the Contemporary Danish spelling according to Dansk Retskrivningsordbogen, 1996.

Norwegian (lexicon size Bokmål 1,005,000 selection August 2006)

The spelling agrees with the Contemporary Norwegian spelling according to Tanums Store Rettskrivningsordbok.

Nynorsk (lexicon size Nynorsk 480,000, selection August 2006) The spelling agrees with the Contemporary Nynorsk spelling according to Det Norske Samlaget.

Saami (lexicon size 100,000, selection April 2006)

The spelling agrees with the Nord Saami language as spoken in Finnmark county in the north of Norway. Inhabitants of Finnmark can request a free version.

Finnish (lexicon size over 3.7 million words, selection April 2007)

The spelling agrees with the Contemporary Finnish, spelling according to Uusi Suomi-Englanti Suur-Sanakirja, 1984.

Afrikaans (lexicon size 265,000, selection September 2006)

The lexicon agrees with the spelling rules of the Suid-Afrikaanse Taalkommissie, 2002.

Latin (lexicon size 450,000, selection June, 2007)

The Latin lexicon has been compiled from classical, medieval, clerical, vulgate, and scientific texts. Names from the classical period and from the clerical (and Biblical) world have been included in the lexicon.

Basque (lexicon size 495,000, selection April 2007)

The Basque language is highly inflected, and so is the Basque lexicon. Geographical and proper names are included in the lexicon: Euskadi, Euskadik, Euskadiko, Euskadikoa, Euskadin, Euskadira, Euskadiren, Euskadirentzat, Euskaditik, Euskadiz etc.

Russian (lexicon size 1,000,000, selection January 2004)

The Russian language goes back to Old Church Slavic, but a literacy tradition less tied to the church and Old Church Slavic exists too. The last extensive spelling reform occurred in 1917.

Estonian (lexicon size 1,325,000, selection September 2007)

The Estonian language belongs to the Finno-Ugric family of languages. It is closely related to Finnish, and similar to Finnish prepositions are attached to the end of the word.

Icelandic (lexicon size 670,000, selection June 2005)

The Icelandic language is a North Germanic (Scandinavian) language, since 1935 the official language of Iceland. The historical morphological characteristics have been preserved.

Lithuanian (lexicon size 290,000, selection August 2006)

The Lithuanian language, like Latvian, belongs to the Baltic family of languages. Lithuanian uses the Latin alphabet with diacritics, including <ė>, <į>, <ų>. Lithuanian is highly inflected.

Latvian (lexicon size 410,000, selection April 2007)

The Latvian language is one of the Baltic languages (see Lithuanian). The orthography is based on the Latin alphabet with diacritic marks, including <ņ>, <ķ>, <ģ>, <ļ>.

Polish (lexicon size 1.6 million, selection August 2007)

The Polish language is a West Slavic language spoken by approximately 42 million speakers. It is written in the Latin alphabet with diacritic marks and special characters: ł, Ł, ź, Ź.

Frisian (lexicon size 95,000, selection January 2005)

The Frisian language is spoken by approximately 300,000 speakers in the Dutch province of Friesland. It has been standardized thanks to the efforts of the Fryske Akademy. It is distinct from East and North Frisian dialects in Northern Germany.

Galician (lexicon size 245,000, selection December 2001)

The Galician language is now spoken in Spanish Galicia, situated north of Portugal. It is a Romance language related to Portuguese. Spelling according “Dicionário da língua galega, Sotelo Blanco”.

Hungarian (lexicon size over 5 million words, selection May 2007)

The Hungarian language belongs to the Uralic family of languages. It is the official language of Hungary. There is a weak relation to the Finno-Ugric languages. The orthography includes characters with the Hungarumlaut: <ő>, <ű>.

Czech (lexicon size 1150,000, selection August 2006)

The Czech language is a West Slavic language. The orthography is based on the Latin alphabet, including diacritics: <č>, <ď>, <ě>, <ů>, <ž>.

Upper Sorbian (lexicon size 590,000, selection August 2007)

The Upper-Sorbian language is a West Slavic language. The orthography is based on the Latin alphabet. Upper and Lower Sorbian is spoken in the South Eastern section of the former German Democratic Republic. Spelling agrees with Hornjoserbskeje rěčneje komisje hač do junija 2005.

visit [download page](#)

Maltese (lexicon size 845,000, selection January 2006)

The Maltese language is a Semitic language written in the Latin alphabet, including <ċ> <ħ> <ġ> and <ż>, orthography according to Joseph Aquilina (1987/1990). The speller includes checks for proper use of assimilations of the article.

New Greek (lexicon size 750,000, selection September 2004)

The Greek characters α , β , γ , ... to ω have been used for millenniums. We do not know how Ancient Greek was pronounced, but modern Greek certainly is different. It now uses only a limited number of accents and diaereses.

Occitan (lexicon size 250,000, Selection August 2007)

Also known as Languedoc, is the original language spoken by the troubadours and Cathars in the South of France. The reconstruction of the language is based on the work of Loís Alibèrt (2000).

Esperanto (lexicon size 300,000, selection August 2003)

Esperanto is an artificial language, introduced by Dr. Lazaro Ludoviko Zamenhof. The language is based on several Indo-European languages. Typical for Esperanto are the characters <ĉ>, <ĝ>, <ĥ>, <ĵ>, <ŝ> and <ŭ>.

Turkish (lexicon size 505,000, selection July 2007)

The Turkish language is written in the Latin alphabet, but a few characters were added, such as the dotless-i which is very different from the dotted-i. Therefore the letter i is not a lower case of the majuscule letter I, a major problem to many systems.

Romanian (lexicon size 315,000, selection July 2007)

The Romanian language belongs to the Roman languages. It includes a few additional characters such as the a-breve <ă>, i-circumflex <î>, the s-cedille <ș>, the t-sedille <ț>.

Bulgarian (lexicon size 840,000, selection April 2004)

The Bulgarian language is written in the Cyrillic alphabet.

Faeroese (lexicon size 175,000, selection February 2005)

The Faeroese language is spoken by 50,000 inhabitants of the Faeroe Islands. It is based on the old Norse as is the Icelandic language.

Bahasa Indonesia (lexicon size 61,000, selection May 2005)

The Bahasa Indonesian language is the standard language written and spoken in the Republic of Indonesia. Many Austronesian languages are spoken in the Indonesian Archipelago, but Bahasa Indonesia is the lingua franca.

Slovene (lexicon size 365,000, selection October 2006)

The Slovene language is spoken in the Republic of Slovenia, situated between Austria, Hungary, Croatia, and Italy. It is a south Slavic language written in the Latin alphabet, including a few Slavic characters such as <č>, <š>, <ž> and the digraphs Lj and Nj. Slovene is highly inflected and nearly every noun has an adjective form too.

Croatian (lexicon size 520,000, selection April 2007)

The Croatian language, formerly named Serbo-Croatian, is closely related to Serbian. The Croatian language is written in the Latin alphabet, including a few typical Slavic characters such as <č>, <ć>, <š>, <ž>, and digraphs Lj and Nj.

Bosnian (lexicon size 525,000, selection April 2007)

The Bosnian language, formerly named Serbo-Croatian, is closely related to Serbian and Croatian.
visit [download page](#)

Serbian Cyrillic (lexicon size 520,000, selection April 2007)

The Serbian language is written in the Cyrillic alphabet, including typical Serbian characters Dž, Lj, Nj (Џ, Љ, Њ, Ђ).

Byelorussian (lexicon size 310,000, selection January 2004)

The Byelorussian language is written in the Cyrillic alphabet, like the Russian language, but the language was heavily influenced by Polish for centuries. Today, in the Byelorussian Republic, Byelorussian plays a lesser role compared to the Russian language.

Slovak (lexicon size 255,000, selection May 2007)

The Slovak language is closely related to Czech, but a few characters differ.

Ukrainian (lexicon size over 1 million words, selection Januari 2004)

The Ukrainian language is written in the Cyrillic alphabet, but for centuries the language was heavily influenced by Polish.

Swahili (lexicon size 75,000, selection February 2005)

The Swahili language is spoken along the East Coast of Africa. It is the lingua franca of many coastal nations. The standardized language is called Kiswahili Sanifu. It shares the word kamusi (dictionary) with the Melayu word kamus. Swahili is written in the Latin alphabet.

Bahasa Melayu (lexicon size 55,000, selection February 2005)

Bahasa Melayu is the standard language of the Republic of Malaysia. It has a common root with Bahasa Indonesia. However, Bahasa Melayu was heavily influenced by the English language while Bahasa Indonesia was influenced by Dutch during the colonial age.

Irish (Gaelic) (lexicon size 325,000, selection January 2007)

The Gaelic language is a Celtic language spoken in Western Ireland. A class of words is lenited, pronounced with palatalization. A slightly different variety is spoken in the Highlands of Scotland.

Welsh (lexicon size 360,000, selection February 2007)

The Welsh language is the Celtic language of Wales, spoken by about 500,000 people (mainly bilingual in English).

Greenlandic (lexicon size 60,000)

is an East Inuit language spoken by 50,000 Greenlanders.

The Greenlandic language adds particle to particle to words and leading to a single word sentence. The Latin alphabet is used whereas the Canadian Inuit make use of their own script.

Macedonian (lexicon size 310,000, selection February 2006)

The Macedonian language is written in the Cyrillic alphabet.

Albanian (lexicon size 110,000, selection February 2006)

The Albanian language is written in the Latin alphabet. The Albanians call their language shqip and their country Shqipëria.

Maori (lexicon selection March 2004)

The Maori language is spoken in New Zealand and is written in the Latin alphabet. A macron is placed above the vowels to differentiate between long and short vowels.

Xhosa (lexicon size 165,000, selection September 2005)

The Xhosa language is spoken in the Republic of South Africa and is written in the Latin alphabet.

Zulu (lexicon size 325,000, selection September 2005)

The Zulu language is spoken in the Republic of South Africa and is written in the Latin alphabet.

Arab (lexicon size ca. 5 million, selection May 2007)

The Arab languages have its own script and the orthography is mainly based on consonantal roots. These roots are unfolded to millions of words.

Azerbaijani (lexicon size 65,000, selection August 2007)

Azerbaijani is written in the Latin alphabet. It has much in common with Turkish.

Hebrew (lexicon size ca. 5 million, selection May 2007)

The Hebrew language is written in Hebrew characters, mainly consonants.

The orthography is based on roots of 3 radicals, which unfolded to millions of words.

Farsi (lexicon size 400,000, selection May 2007)

The Farsi or Persian language is written in the Arab script, but being an Indo-European language vowels are important.

Urdu (lexicon size 37,500, selection May 2007)

The Urdu language is closely related to Hindi, but written in the Arab script. Urdu and Hindi are Indo-European languages.

Breton (lexicon size 210,000, selection July 2007)

The Breton language is spoken in French Bretagne. It is a Celtic language once related to extincted Cornish in the UK.

September, 2007

*TALO by, Bussum, The Netherlands