



## Spelling Checkers — language specifications version 6.2.4.6.

Languages and sizes of dictionaries

New languages: **Oriya (Odia India), Luxembourgish, Friulian (Italy), Latin, Kazakh (Cyrillic/Latin), Khmer (Cambodia), Kurdish (Northern), Nepalese, Marathi, Azerbaijani.**

Recent upgraded spell checker languages: **Greenlandic, New Greek, Modern Hebrew, Polish, Arabic, Spanish (4x), French/Canadian French, English (5x), Danish, Swedish, Dutch, Flemish, Surinam Dutch, Norwegian, Nynorsk, Finnish, Italian, German, Swiss German, Austrian German, Zulu, Afrikaans, Slovak, Czech, Galician, Frisian, Icelandic, Slovenian, Persian/Farsi, Portuguese (acordo ortográfico)**

96 languages (varieties)

**English** (lexicon size between 512,550 and 514,000 plus, selection February 2025) (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 659,600, February 2025) (2x)

Includes the most extensive geographical lexicon. Two lexicons are available, one according to the spelling of Le Larousse (2008), Le Nouveau Petit Robert (2025) & Le Robert illustré & Internet (2015) 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>) and has become more and more accepted at present time. *La nouvelle orthographe du français n'est pas imposée, mais elle est officiellement recommandée. Les modifications, modérées, touchent environ deux-mille mots. Exemples :*

- un compte-goutte, des compte-gouttes ;
- un après-midi, des après-midis; cout ;
- entrainer, nous entrainons ;
- paraître, il paraît; j'amoncèle, amoncèlement, tu époussèteras

and most important: Les rectifications de l'orthographe ont été approuvées initialement par :

- Le Conseil supérieur de la langue française (Paris);
- L'Académie française (France).

The system includes Canadian French versions and extensive re-spelling tools between previous and new spelling

forms , and common mistakes.

### **Orthographic recommendations 1990, now preferred by the French Academy**

**Canadian French** (lexicon size over 655,400, selection February 2025) (2x)

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

### **Orthographic recommendations 1990, now preferred by the French Academy**

**German** (lexicon size 1,334,300, selection October 2024) (4x)

The German orthography has been updated with the acceptance of the uppercase Eszett. Previous versions will be kept in the meantime.

The German spelling is distributed in four versions, “alt (pre 1996), new 1996 (the very first reform), new current orthography (2017), and the dpa version (2017)”. These versions include automatic respelling from old to new spelling forms (e.g., Prozeß → Prozess) and of *feste grammatische und lexikale Wendungen*. Using the old orthography or “alte Rechtschreibung” enables you 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>).

The orthography *neue Rechtschreibung* is updated according to the Duden 29, August 2024, and the “Rat für deutsche Rechtschreibung”, Grundlagen der Deutschen Rechtschreibung (2017), including the Eszett-Schreibung (ß/ß).

The German lexicon is based on over ca. 300,000 expanded catchwords (konjugierte Stichwörter), and includes all German toponyms (Ortsnamen), over 13,000 autocorrections (Umschreibungen) and an extensive medical lexicon. Moreover, spell checking is strict, we don't approve errors like: *Oberklasse-Wagen, Oberstufe-Schüler, Klasse-Bücher*. It has to be: *Oberklassenwagen, Oberstufenschüler, Klassenbücher*.

### **Orthographic reform 1996, 2006 & 2017**

**Swiss German** (lexicon size 1,368,500, selection February 2025, Swiss additions to German) (4x)

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

### **Orthographic reform 1996, 2006 & 2017**

**Austrian German** (lexicon size 1,342,500, selection February 2025, Austrian additions to German) (4x)

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

### **Orthographic reform 1996, 2006 & 2017**

**Spanish Peninsular, Argentine, Mexican & Latin American** (lexicon size over 996,700, selection February 2025)

The spelling is according to the new orthographical rules presented in the latest (la última edición) of the Ortografía de la lengua española (2010). Includes respelling of a set of orthographical changes and common errors, e.g.,

exteniente coronel → ex teniente coronel,  
ex presidente brasileño → expresidente brasileño,  
anti-mafia → antimafia,  
Adam y Eva → Adán y Eva,  
Edinburgo → Edimburgo.

### **Orthographic reform 2010, 2014**

**Italian** (lexicon size 998,200, selection February 2025)

The spelling is according Lo Zingarelli 2014. Includes pronomial forms, and an extensive geographical lexicon (comuni e luoghi italiani), and a set of multiple word corrections, e.g., *le USA* → *gli USA*, *il pneumatico* → *lo pneumatico*, *vicino Roma* → *vicino a Roma*.

**Swedish** (lexicon size over 2,146,000 words, selection February 2025)

Includes geographical and proper names, SI unit correction and punctuation correction (not «*blod, svett och trådar*», but »*blod, svett och tårar*»); orthography according to Svenska Akademiens ordlista över svenska språket.

**Portuguese** (lexicon size over 1.715 million words, selection July 2023) (4x)

Iberian and Brazilian Portuguese are very different in terms of use of verb tenses and idiom. Often Brazilian Portuguese is unacceptable for Iberian Portuguese publications, and the reverse is a source of misunderstanding too. Independently of orthography dictionaries need to be different. Therefore Iberian and Brazilian versions according to the previous and acordo ortográfico, have been compiled. These versions include respelling either between Iberian Portuguese and Brazilian Portuguese or between the previous and acordo ortográfico. *O presidente de Portugal, Aníbal Cavaco Silva, promulgou o acordo ortográfico da língua portuguesa, ratificado no Parlamento do país em maio, informaram hoje à Agência Efe fontes da Presidência. ....*

Examples: *equipolente* versus *equiipolente* or *boleia* versus *boléia* or *ação* versus *acção*.

**Orthographic reform 2009, 2010**

**Dutch** (Nederlands, lexicon size 893,500, selection January 2025) (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 over 26,000 collocations and (respelling) autocorrections from the previous to the new orthography is included. This set includes multiple word alternatives for weird combinations such as “door de regen en de wind” → 1) *door weer en wind*, or 2) *in de regen en de wind*, a linguistic mutilation of (*come*) *rain and shine*.

**Orthographic reform 1995 & 2005, 2015**

**Flemish** (Vlaams, lexicon size 918,200, selection January 2025) (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 over 25,300 collocations and (respelling) autocorrections from the previous to the new orthography is included. This set includes multiple word alternatives for weird combinations such as *kost duur* → 1) *is duur*, or 2) *kost veel*, a linguistic mutilation of *is expensive*.

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.

**Orthographic reform 1995 & 2005, 2015**

**Surinam Dutch** (Surinaams-Nederlands, lexicon size 894,300, selection January 2025)

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.

**Orthographic reform 1995 & 2005, 2015**

**Catalan** (lexicon size 1,454,100, selection February 2023)

The lexicon includes all combinations of *pronominals*. The spelling agrees with *Diccionari ortogràfic i de pronúncia, Enciclopèdia Catalana*. The nova ortografia — not *aritmia* but *arrítmia*, not *angiospasm* but *angioespasm* — agrees with the Institut d'Estudis Catalans whose spelling reform has been entered with a transition

period of five years.

### **Orthographic reform October 2016**

**Danish** (lexicon size over 1,191,000, selection February 2025)

The spelling agrees with the Contemporary Danish spelling according to Dansk Sprognævn, med ændrede ord- og staveformer; with recent orthographic changes (2012). It includes present-day idiom, and a set of multiple word corrections, e.g., *æblerne lægger i skålen* → *æblerne ligger i skålen*, *web siderne* → *websiteserne*, etc.

### **Orthographic reform 2012 (update 2015)**

**Norwegian** (lexicon size Bokmål 1,374,000 selection December 2024)

The spelling agrees with the Contemporary Norwegian spelling according to Tanums Store Rettskrivningsordbok and the Norwegian Språkrådet. It includes present-day idiom, and a set of multiple word corrections, e.g., *i Møre* → *på Møre* (*always notify*), *på Møre og Romsdal* → *i Møre og Romsdal*, etc.

### **Orthographic reform 2023**

**Nynorsk** (lexicon size Nynorsk 617,600, selection December 2024)

The spelling agrees with the Contemporary Nynorsk spelling according to Det Norske Samlaget.

### **Orthographic reform 2023**

**Saami** (lexicon size over 1.6 million, selection August 2014)

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 5.56 million words, selection February 2025)

The spelling and tagging (taivutustyypit) agrees with the Contemporary Finnish, Kielitoimiston sanakirja, 2012. The lexicon has extensively been tagged with declension and conjugation classifications. This is a requirement given the compound nature of the Finnish language.

**Afrikaans** (lexicon size 341,000, selection August 2024)

The lexicon agrees with the spelling rules of the Suid-Afrikaanse Taalkommissie. It matches to present-day idiom of the South African society, including a wide variety of neologisms, geographical, business, and social words. The spell checker includes mechanism to proof neologisms by examination of component parts. This mechanism doubles the effective size of the lexicon.

**Latin** (lexicon size 450,000, selection August, 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 3.84 million, selection February 2023)

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 1.390 million, selection February 2018)

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 over 2,235,000, selection January 2020)

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 848,000, selection December 2023)

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 941,000, selection September 2020)

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

**Latvian** (lexicon size 1,207,000, selection September 2020)

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 1.9 million, selection December 2017)

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, ż, Ż.

**Frisian** (lexicon size 444,500, selection April 2024)

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. Orthography in agreement with the “offisjele stavering fan de Fryske taal 2014” published by the of the Fryske Akademy, e.g., ienentweintichste-iuwsk not ienentweintichste-ieusk.

**Orthographic reform 2014**

**Galician** (lexicon size 281,900, selection July 2024)

The Galician language is now spoken in Spanish Galicia, situated north of Portugal. It is a Romance language related to Portuguese. Spelling according “Dicionário Século21 da língua galega”.

**Hungarian** (lexicon size over 5 million words, selection December 2020)

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 1,810,000, selection July 2024)

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 600,000, selection January 2009)

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 788,100, selection August 2023)

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 Loïs Alibèrt (2000).

**Esperanto** (lexicon size 300,000, selection September 2020)

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 1,860,000, selection November 2015)

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 1,000,000, selection June 2009)

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>, the s-comma below, the t-comma below.

**Bulgarian** (lexicon size 840,000, selection February 2016)

The Bulgarian language is written in the Cyrillic alphabet.

**Faeroese** (lexicon size 579,200, selection March 2023)

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

**Bahasa Indonesia** (lexicon size 77,000, selection July 2020)

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.

**Slovenian** (lexicon size 754,300, selection August 2023)

The Slovenian 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. Slovenian is highly inflected and nearly every noun has an adjective form too.

**Croatian** (lexicon size 633,000, selection April 2016)

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 650,000, selection April 2016)

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

**Serbian Cyrillic** (lexicon size 658,000, selection April 2016)

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

**Byelorussian** (lexicon size 1,600,000, selection February 2009)

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 1.15 million, selection August 2024)

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

**Ukrainian** (lexicon size over 1,154 million words, selection August 2022)

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 61,000, selection September 2009)

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 August 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 925,000, selection January 2016)

The Welsh language is the Celtic language of Wales, spoken by about 500,000 people (mainly bilingual in English). Multiple word correction mechanism supports mutations of consonants, e.g., *fy car* → *fy nghar*.

**Greenlandic** (lexicon size 3,476,600, selection April 2025)

is an East Inuit language spoken by 57,000 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 322,000, selection April 2016)

The Macedonian language is written in the Cyrillic alphabet.

**Albanian** (lexicon size 585,000, selection April 2011)

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 171,000, selection May 2020)

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

**Zulu** (lexicon size 372,750 selection August 2024)

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

**Arabic** (lexicon size ca. 5 million, selection February 2025)

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 132,000, selection May 2010)

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

**Hebrew** (lexicon size ca. 5.5 million, selection August 2023)

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.

**Persian/Farsi** (lexicon size 458,800, selection August 2023)

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

**Urdu** (lexicon size 133,000 , selection December 2023)

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

**Breton** (lexicon size 653,000, selection July 2022)

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

**Hindi** (lexicon size 541.000, selection December 2023)

The Hindi language is spoken in northern and central India. Written Hindi is relatively standardized over the whole Hindi language area. It is an Indo-Aryan language. Although related to Urdu, Hindi does not favour the use of Persian and Arabic loanwords. Hindi is written in the Devanagari script, it includes a lot of complex characters, consisting of vowels, consonants, vowel-signs (matras), numerals, and diacritical marks.

**Marathi** (lexicon size 1,681.100, selection December 2023)

The Marathi language is spoken in the Mahatashtra state of India. It is an Indo-Aryan language written in the Devanagari script.

**Nepalese** (lexicon size 130.000, selection December 2010)

The Nepalese language (Nepali) is spoken in the Himalayan state of Nepal between India and China. Nepalese is written in the Devanagari script.

**Kurdish (Northern)** (lexicon size 90,000, selection July 2009)

belongs to the Iranian group of languages. Kurdish is spoken in Turkey, Iraq, Iran, Armenia, Georgia and Azerbaijan. The latin script is used for the Northern variety of Kurdish.

**Malayalam** (lexicon size 781,000, selection April 2021)

The Malayalam language is spoken in Kerala, a state in the south of India. It is a Dravidian language written in the Malayalam script, a descendant of the Brahmi script.

**Bengali** (lexicon size 585,000, selection March 2021)

The Bengali language is spoken in Bangladesh. It is a Indo-Aryan language written in the Bengali script, a descendant of the Brahmi script.

**Gujarati** (lexicon size 189,000, selection October 2018)



The Gujarati language is spoken in the Indian state of Gujarat. It is a Indo-Aryan language written in the Gujarati script, a descendant of the Brahmi script.

**Tamil** (lexicon size 1,250,000, selection February 2021)

The Tamil language is spoken in southern India (Tamil Nadu) and Sri Lanka. It is a Dravidian language written in the Tamil script, a descendant of the Brahmi script. Tamil has many Indo-Aryan loanwords. Tamil in Sri Lanka incorporates loanwords from the Dutch, Portuguese, and English language.

**Sinhala** (lexicon size 208,000, selection November 2009)

The Sinhala language is spoken in Sri Lanka India. It is an Indo-Aryan branch of the Indo-European languages written in the Sinhala script, a descendant of the Indian Brahmi script. There is some affinity to neighbouring languages. Sinhala has features that may be traced to Dravian influences.

**Punjabi** (lexicon size 94,000, selection January 2018)

The Punjabi language is spoken in the Punjab state of India. It is an Indo-Aryan branch of the Indo-European languages written in the Gurmukhi script, a descendant of the Indian Brahmi script.

**Telugu** (lexicon size 240,000, selection January 2018)

The Telugu language is spoken in Andhra Pradesh, one of the largest states of India. It is a Dravidian of the Indo-European languages written in the Telugu script, a descendant of the Indian Brahmi script.

**Oriya (Odia)** (lexicon size 331,000, selection January 2018)

The Oriya or Odia language is spoken in Odisha state of India. It is an Indo-Aryan branch of the Indo-European languages written in the Kalinga script, a descendant of the Indian Brahmi script.

**Khmer** (lexicon size 30,000, selection November 2009)

The Khmer language is spoken in Cambodia. It is the second most widely spoken Austroasiatic language. As in Thai Khmer sentences are written without spaces. Therefore spell checking strongly depends on segmentation.

**Kazakh** (Cyrillic/Latin) (lexicon size 900,000, selection May 2010)

The Kazakh language is spoken east of the Caspian Sea. It is a Turkic language related to Azerbaijan and Turkish. Kazakh is mainly written in the Cyrillic alphabet in Kazakhstan but a transition to the Latin script has already been brought up by the President of Kazakhstan in 2006. For this reason both Cyrillic and Latin lexicons have been compiled.

**Luxemburgish** (Lëtzebuergesch) (lexicon size 200,000, selection December 2012)

The Lëtzebuergesch language is spoken in the Grand Duchy of Luxembourg. The language/dialect descends from Mosel-Frankish, a dialect, linguistically close to High German and Limburgish. The population of Luxembourg is half a million only.

April, 2025

\*TALO bv, Bussum, The Netherlands