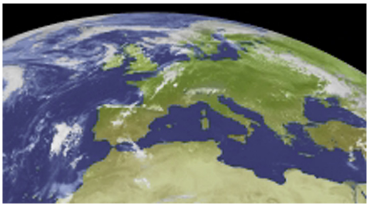
*TALŌ's LANGUAGE TECHNOLOGY

HYPHENATORS
SPELL CHECKERS
DICTIONARIES



Dr. J.C. Woestenburg,
**TALŌ b.v.,*
Lijsterlaan 379,
1403 AZ Bussum,
The Netherlands.

tel: +31 35 69 32 801
e-mail: info@talo.nl
http://www.talo.nl/

Spelling of lexical and grammatical collocations
Enlarged edition
Completely revised
License the best language technology!
Revision April 2025

*TALŌ

the Germanic root of our words *tell*,
tale, like a *notch made on a tally*

Please, view PDF versions of this book with Adobe Reader Windows, Linux (x86) or Mac only. For other platforms or viewers, fonts might miss some glyphs (character images) or PDF engines might incorrectly position some combining (diacritical) glyphs.

23th Edition, April 2025

The greatest care has been taken in compiling this book. However, no responsibility can be accepted by the publisher or author for the accuracy of the information presented.

Content

Preface

The word *TALŌ is a reconstruction of an Old Germanic word meaning a *notch on a tally*. The original marks on a *tally* can be thought of as the very first digital carvings in wood, being similar to bits in a present-day computer language word.

The emergence of computers enabled the development of mathematical methods for processing large volumes of text. The mathematical engineering approach of the 80s, however, does not fit into the psychology of language and doesn't solve problems that are inherent in a language. In other words: the nature and complexity of language were not included in the work done by the mathematicians and engineers. Consequently, the mechanical approach didn't solve the basic problems encountered in hyphenation and spell checking. The diversity and nature of languages force us to reconsider the way technology is set up.

All languages have their roots and linguistic family relations in the distant past.
Scientists have proposed hypotheses about the earliest emergence of speech, but the real origins of speech can not be established. Factual knowledge about languages dates back to about 5000 or 6000 years ago.
Early hypotheses assumed an *onomatopoetic* root as (wo)men tried to imitate animal sounds. Other hypotheses assume that speech has its origins in the expression of emotions or in the communication needed during cooperative efforts (labour). Some theories assume an underlying application of rules of generative innate grammar. More recently, the semantic factor of language has been emphasized. The structure of words doesn't seem to be just the result of a chaotic process, but it also doesn't seem to be just the result of the application of certain rules.

This dominant semantic factor, a trace of meaning, has guided us in the construction of tools for processing written language. The basic elements of meaning have been used to find the smallest building blocks of meaning that help to decide where to hyphenate. These building blocks vastly improve hyphenation as well as spelling. They are applied in a set of tools described in the first chapter of this book.

This book is intended to provide a better view of language. It reviews the different hyphenations in different languages as well as the relations between families of languages and the origins of languages.
When our ancestors, about 1500 generations ago, came up with names for man (*manu- in Proto-Indo-European) and for family members, and later on started riding horses, invented the wheel, and built fortified elevated places (referred to in word combinations containing *bhergh*-), all of these activities and subjects were given names. These ancient roots are still present in current languages despite the fact that through the ages languages have continuously been subject to changes.

Development of relevant, practical and successful language technology requires the knowledge and understanding that a company like *TALŌ has of the evolution and peculiarities of many individual languages.

J.C.Woestenburg, PhD,
Bussum, 28 September 2006

Who is *TALŌ?

*TALŌ is a Netherlands-based software company, specializing in the development of language modules. We conceived, designed, and developed the *TALŌ hyphenation system from a psychological, cognitive, and language perspective. The system is based on a unified structure for syllabification, that works in every alphabetical language. Our first product was called the "Dutch hyphenator", and was integrated into the Dutch version of WordPerfect in 1986. Our next step was building a software generator to create hyphenators for other languages. The creation of hyphenator modules was followed by the development of spelling modules, based on the same language principles.

The development of a hyphenator begins with the thorough research into the language involved, followed by the action of the powerful generator. The reliability of the hyphenator depends on the quality of the research. In a short period of time, determined by the complexity of a language, the generator completes the creation of the hyphenator. The fundamental language model based on morphological principles produces reliable hyphenations. The mechanism of the hyphenator is a process of pattern recognition, based on the psychology of the working brain, and reacts instantaneously. Information is available in a very short time. The combination of this mechanism and the language model results in the hyphenator's high speed. The hyphenators are currently comparable to the end state of a neural network.

The hyphenators follow the rules and exceptions of hyphenation governed by leading academies. They internally take care of every peculiarity of a language. They correctly hyphenate new words and even complex compounds.

The *TALŌ speller/correctors are conceptually different from other 'spellers'. The morphological principles applied to their mechanism stem from our vast experience with hyphenators. Reliability is ensured by the principle of fuzzy distance: an alternative is presented that really resembles the misspelled word, including reversals (e.g. the instead of the) and missing (initial) letters. Each European language has a very distinct character of its own. Therefore, each speller has its own language model, describing its characteristic behaviour, guaranteeing its speed. The most recent spellers go beyond the word boundary leaving the fundamental principles unchanged.

The spellers also follow the rules and exceptions of spelling governed by leading academies. Abbreviations, punctuation marks, fixed formats like dates and numbers (financial data), and Internet addresses are being controlled during the spelling process. The speller has a learning capability but forgets infrequent, unimportant words. Spelling reforms, from old to new and vice versa, can be executed automatically. Spelling changes (for example between leading academies or between American and British English) are an option too. You can create and implement your favourite style for communication with the external world. You can try a search with wild cards in the large lexicons for interesting and surprising information!

Since the first implementation we at *TALŌ have made significant progress in the development of hyphenators and spellers for the European, Asian, and African languages. We have also put considerable effort into improving both the underlying scientific principles and the technical realizations.

The language modules have been grouped together in shared libraries. The differences in language processes are encapsulated to facilitate an integration with the OEM's particular system, either as a runtime linkable library or as a coherent set of sources. The shared libraries can easily be plugged into smaller systems.

At present, the language modules are available for more than 96 languages, with several other modules under development.

January, 2010; August, 2011; January, 2013; December, 2014; May, 2015; January, 2016; October, 2019; December, 2020; June, 2021; December, 2022, April 2025.

Insert *TALŌ in your software

The development of software for page-layout systems takes years. So why don't you integrate hyphenation and spell checking routines of the highest standard?

This is now possible because you have the opportunity to licence the *TALŌ technology!

*TALŌ offers the best language solutions in the world. Syllables, as the building blocks of language, from the basis for the decisions on how to spell a word and where to hyphenate it. Even without user-defined exceptions, the language model is over 99,95% reliable. The *TALŌ technology has its roots in the Netherlands but is unique throughout the world!

Our first product was called the Dutch pattern hyphenator and was integrated into the Dutch version of WordPerfect in 1986. Since then the *TALŌ modules have been implemented in Quark XTensions for QuarkXPress, Plug-Ins for Adobe's InDesign, as well as in editorial systems of newspapers and independent companies, that have developed their own system.

At present, the language modules are available for more than 70 languages, with several other modules under development.

In the following pages, you will find out more about the technical details, or you can surf to *TALŌ's Web site at http://www.talo.nl/.

Both the spelling and the hyphenation software are available for dozens of languages, and this number is increasing each year.

For more information:

TALO b.v.,
Lijsterlaan 379,
1403 AC Bussum, The Netherlands.

tel: +31 35 69 32 801

*TALŌ

the Germanic root of our words *tell*, *tale*, like a *notch made on a tally*

CHAPTER 1

1. **TALŌ's Language Technology**

1.1. **TALŌ's Speller for the European, American, Australian, African and Asian languages**

Rechtschreibung, stavning, ret(t)skriving, orthographe, orthografia, spelling, speltoetsing, pravopis, helyesírás

A new upgrade of version 6.2.4.(6) of the *TALŌ Speller for the European languages has been released. The speller is distributed with editor tools and with over 250 MB of lexicon data. The base of the speller is a Dynamic Link Library (DLL), which can be linked to redactional systems (newspapers) and to nearly any output application, such as used for the prepress industry. They should only insert their own interface between their application and *TALŌ's Dynamic Link Library.

Why *TALŌ's Speller?

The mechanisms of *TALŌ's Speller incorporate the heritage of the common background of the (Indo-)European languages. The derived morphological knowledge helps in seeking relevant alternatives for spelling errors, in a way that outperforms outdated phonetical algorithms that are still in use. Moreover, the speller is tuned into the peculiarities of each individual language by using an intelligent language technology. Version 6.2.4 goes beyond what normal spelling is, right into the domain of grammatical and lexical collocations.

Languages

*TALO's language tools are typically meant for newspapers, magazine and book publishers and other businesses that produce text “en masse” in:

American English,	British English,
Canadian English,	South African English,
Australian English,	Dutch (three spelling variants),
Flemish (three spelling variants),	Frisian,
Afrikaans,	German (old spelling),
German (reformed spelling, 2x),	German (news agencies),

Swiss German (old spelling),	Swiss German (reformed spelling, 2x),
Swiss German (news agencies),	Austrian German (old spelling),
Austrian German (reformed spelling),	Austrian German (news agencies),
French (regular),	French (recommended),
Canadian French (regular),	Canadian French (recommended),
Italian,	Latin,
Spanish,	Catalan (nova ortografia),
Galician,	Basque,
Iberian Portuguese,	Brazilian Portuguese,
Iberian Portuguese acordo),	Brazilian Portuguese acordo,
Danish,	Swedish,
Norwegian (bokmål),	New Norwegian (Nynorsk),
Icelandic,	Finnish,
Faroese,	Sámi,
Estonian,	Lithuanian,
Latvian,	Polish,
Czech,	Slovak,
Slovene,	Croatian,
Russian,	Ukrainian,
Byelorussian,	Bulgarian,
Serbian,	Hungarian,
Macedonian,	Albanian,
Rumanian,	Bosnian,
New Greek,	Turkish,
Bahasa Indonesia,	Bahasa Melayu,
Swahili,	Occitan,

Maltese,	Esperanto,
Irish (Gaelic),	Welsh,
Maori,	Zulu,
Surinam Dutch,	Xhosa,
Arabic,	Hebrew,
Persian/Farsi,	Urdu,
Thai,	Kurdish (Northern),
Hindi,	Marathi,
Nepalese,	Malayalam,
Bengali,	Gujarati,
Tamil,	Sinhala,
Punjabi,	Telugu,
Khmer,	Luxembourgish
Oriya (Odia),	New languages in preparation

*TALO's dictionaries are huge, e.g., the Swedish dictionary consists of over 2,1 million words, the German one covers over 1,330,000 words (mit Fremdwörter), and the Hebrew dictionary consists of 5,5 million words. A great many compounds can be found in *TALŌ's lexicons, and additional functionality helps in viewing the general structure of compounds, to decide whether one should keep compounds together or separated.

Key points

The speller's key points are speed, accuracy in terms of alternatives, language-dependent peculiarities, and proper capitalization. Last but not least is an extensive respelling facility which acts like human memory by using priorities. Additionally, correction functionality for grammatical and lexical collocations and an open ***punctuation*** and ***citation checking*** mechanism has been added. The first mechanism is the watchdog of style. The second mechanism also ***controls capitalization*** of ***sentences***.

Is your speller *false approval compliant?*

Quite often spellers accept strange words too easily. For *TALŌ's Speller, these drawbacks are absent (e.g., “Aschemittwoch” as correct instead of showing “Aschermittwoch” (German), “registratieapparaat” versus “registreerapparaat” (Dutch), and “registreeramptenaar” versus “registrasieamptenaar” (Afrikaans)). A most salient item is that there are no false approvals to camouflage limitations. False approvals are really real spelling mistakes. They are quite often mistakenly accepted as correct, which is an expensive mistake! Is your speller *false approval compliant*?

Crossing the word barrier

Grammatical and lexical collocations are another category of unexpected errors: *a ytterbium, a honest reaction, all us, the snake river* in the Rockies. Wouldn’t it be better to have the answer: *an ytterbium, an honest reaction, all of us, the Snake River* in the Rockies?

*TALŌ’s Speller is distributed on DVD. It also can be made to comply with your special requirements. Demonstration versions can be downloaded from “http://www.talo.nl/download/index.html” — or just call us.

*TALŌ bv, Lijsterlaan 379, 1403 AZ Bussum, the Netherlands
Tel.: +31 (0) 35 69 32 801,
E-mail: info@talo.nl, Website: http://www.talo.nl/

1.1.1. **Speller: languages and sizes of dictionaries**

English [03,04,21,39,84][‡] between 512,550 and 514,000 words (American, British, Canadian, South African, Australian versions selection February 2025), includes a set of collocations and automatic respelling functions between American English, Canadian English, and British English orthographical varieties, e.g., (to UK) *counseling* -> *counselling* or (to US) *counselling* -> *counseling*; (UK & US) *Mao Tse-tung* -> *Mao Zedong* (see the Style Guides of the New York Times and the Economist).
Be careful with expressions as *Thanks God its Friday*! Without an apostrophe it looks a bit strange. Therefore, a set of multiple word corrections is included, e.g., *thank God its Friday* -> *thank God it's Friday*, or with multiple alternatives, *redo it over* -> **1**) *redo it*, **2**) *do it over*, etc.

French [05,17,14,27] over 659,600 words (selection February 2025, including the most extensive geographical lexicon). Two lexicons are available, one according to the previous spelling and one according to the Rectifications de l'orthographe of the *Conseil supérieur de la langue française*, published 6 December 1990. Canadian French versions are available as well. Extensive respelling tools between previous and new spelling forms are available.

German [06,10,25,92] 1,334,300 words (selection February 2025). 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 (2022), and the dpa version (2025)".
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 (2024), including the Eszett-Schreibung (ß to ESZETT UPPERCASE).
The German lexicon is based on over ca. 317.000 expanded catchwords (konjugierte Stichwörter), and includes all German toponyms (Ortsnamen), over

[‡] The number between brackets refer to orthographic modules compiled into the shared library. The numeration agrees with age of a module.

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*.

Swiss German [07,11,26,93] 1,374,900 Swiss additions to German. There are four versions "alt, 1996, neu (2006)-2025, dpa/SDA (2007-2025)" see German.

Austrian German [08,12,28,94] 1.348,900 Austrian additions to German. There are four versions "alt, 1996, neu (2006-2025), dpa (2007-2025)" see German.

Spanish [15] 989,400 - 996,700 words (selection February 2025) 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.*

Italian [16] 998,200 words (selection February 2025) spelling 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., *il pneumatico* -> *lo pneumatico*, *vicino Roma* -> *vicino a Roma*.

Swedish [18] 2,146,200 words (selection February 2025), includes geographical and proper names, SI unit correction and punctuation correction (not «blod, svett och trådar (or dårar)», but »blod, svett och tårar»); orthography according to *Svenska Akademiens ordlista över svenska språket*.

Portuguese [19,20,92,93] over 1,715 million words (Iberian and Brazilian versions, selection July 2023), 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. ...., O Novo Acordo Ortográfico da Língua Portuguesa está em vigor no Brasil desde* *o último dia 1° (2009). Examples: equipolente versus eqüipolente or boleia versus boléia or ação versus acção*.

Dutch [01,02,80] 893,500 words (selection January 2025). The spelling according to the governmental rules (Groene Boekje, Werkgroep 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 23,800 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*.

Flemish [08,09,81] 918,200 words (selection January 2025). The spelling according to the governmental rules (Groene Boekje, Werkgroep 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. 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*

Surinam Dutch [78] 894,300 words (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, Werkgroep 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 (see Dutch).

Catalan (nova ortografia) [29] 1,454,100 words (selection February 2023), the lexicon includes all *pronoms personals* and agrees with the Diccionari ortogràfic i de pronúncia, Enciclopédia Catalana. The nova ortografia, not *arítmia* but *arrítmia*, not *angiospasme* but *angioespasme*, agrees with the Institut d'Estudies Catalans whose spelling reform has been entered with a transition period of five years.

Danish [22] over 1,191,000 words (selection February 2025). Contemporary Danish, spelling according to Dansk Retskrivingsordbogen of the Dansk Sprognævn (2012). It includes present-day idiom, and a set of multiple word corrections, e.g., *æblerne lægger i skålen* -> *æblerne ligger i skålen*, *web siterne* -> *websiterne*, etc.

Norwegian, Nynorsk [19,20] Bokmål 1,374,000 words (selection December 2024). Contemporary Norwegian, spelling according to Tanums Store Rettskrivningsordbok) and Nynorsk 617,600 words (selection March 2024). 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.

Sámi [54] 1,612,000 words, (selection March 2019). The spelling agrees with the Nord Sámi language as spoken in Finnmark county in the North of Norway.

Finnish [25] over 5.56 million words (selection February 2025). Spelling and templates (taivutustyypit) of the revised modules agree 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 [31] 341,550 words (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 [30] 450,000 words (selection August, 2007). The Latin lexicon has been compiled from classical, medieval, clerical, vulgate, and scientific texts. Since names from the classical period and from the clerical (and Biblical) world recur in texts, they have been included in the dictionary.

Basque [51] 3,84 million words (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.*. nn **Russian [40]** 1,390,000 words (selection February 2018). The Russian language goes back to Old Church Slavic, but there also exists a literacy tradition less tied to the church and Old Church Slavic. The last extensive spelling reform occurred in 1917.

Estonian [38] 2,235,000 words (selection January 2020). The Estonian language belongs to the Finno-Ugric family of languages. It is closely related to Finnish, and, like Finnish, prepositions are attached to the end of the word.

Icelandic [41] 848,300 words (selection December 2023). The Icelandic language is a North Germanic (Scandinavian) language, since 1935 the official language of Iceland. Historical morphological characteristics have been preserved.

Lithuanian [46] 941,000 words (selection September 2020). The Lithuanian language like the Latvian language belongs to the Baltic family of languages. Lithuanian uses the Latin alphabet with diacritics, including as <ė>, <į>, <ų>. Lithuanian is highly inflected.

Latvian [44] 1,207,000 words (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 [47] 1.964 million words (selection March 2025). The Polish language is a West Slavic language spoken by ca. 42 million speakers. It uses the Latin alphabet with diacritic marks and special characters: ł, Ł, ż, Ż.

Frisian [49] 444,500 words (selection April 2024). The Frisian language is spoken by approx. 300,000 speakers in the Dutch province of Frisia. It has been standardized due 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".

Galician [58] 281,900 words (selection July 2024). The Galician language is spoken in Spanish Galicia, situated north of Portugal. It is a Romance language related to Portuguese. Spelling according “Dicionário Século21 da lingua galega”..

Hungarian [42] over 5 million words (selection December 2020). The Hungarian language belongs to the Uralic family of languages. It is the official language of Hungary, and is somewhat related to the Finno-Ugric languages. The orthography includes characters with the Hungarumlaut: <ő>, <ű>.

Czech [52] 1,810,000 words (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 [81] 775,000 words (selection March 2009). The Upper Sorbian language is a West Slavic language spoken in the South Eastern part of the former German Democratic Republic. Spelling agrees with Hornjoserbskeje rěčneje komisje hač do junija 2005.

Maltese [66] 845,000 words (selection January 2006). The Maltese language is a Semitic language using the Roman alphabet, including <ċ> <ħ> <ġ> and <ż>. The speller includes checks for proper use of assimilations of the article.

New Greek [45] 788,100 words (selection March 2025). 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 only uses a limited number of accents and diaereses over vowels

Occitan [65] 210,000 words (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 [67] 300,000 words (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 [43] 1,86 million words (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 letter *i* is not a lower case of the majuscule letter *I*, a major problem to many text processing systems.

Romanian [55] 1,000,000 words (selection June 2009). The Romanian language belongs to the Roman languages. It includes a few additional characters such as the a-breve, i-circumflex, the s-cedille, the t-sedille, s-comma below and t-comma below.

Bulgarian [70] 840,000 words (selection February 2016). The Bulgarian language uses the Cyrillic alphabet.

Faroese [53] 579,200 words (selection March 2023). The Faroese language is spoken by 50,000 inhabitants of the Faroer Islands. It is based on the old Norse as is the Icelandic language.

Bahasa Indonesia [63] 77,000 words (selection May 2020). Bahasa Indonesia 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 [56] 754,300 words (selection August 2023). The Slovene language is spoken in the Republic of Slovenia, situated between Austria, Hungary, Croatia, and Italy. It is a south Slavic language using the Latin alphabet, including a few Slavic characters such as <č>, <š>, <ž> and the diagraphs Lj and Nj. Slovene is highly inflected and nearly every noun has an adjective form too.

Croatian [60] 633,000 words (selection April 2016). The Croatian language, formerly named Serbo-Croatian, is closely related to Serbian. The Croatian language uses the Latin alphabet, including a few typical Slavic characters such as <č>, <ć>, <š>, <ž>, and digraphs <Lj> and <Nj>.

Bosnian [71] 650,000 words (selection April 2016). The Bosnian language, formerly named Serbo-Croatian, is closely related to Serbian and Croatian.

Serbian Cyrillic [62] 658,000 words (selection April 2016). The Serbian language uses the Cyrillic alphabet, including typical Serbian characters Dž, Lj, Nj (Џ, Љ, Њ).

Byelorussian [69] 1,600,000 words (selection February 2008). The Byelorussian language uses the Cyrillic alphabet, like the Russian language, but for centuries the language has been heavily influenced by Polish. Today Byelorussian plays a lesser role compared to the Russian language in the Byelorussian Republic.

Slovak [61] 1.15 million words (selection August 2024). The Slovak language is closely related to Czech, but a few characters differ.

Ukrainian [48] over 1.154 million words (selection August 2022). The Ukrainian language uses the Cyrillic alphabet, but the language has been heavily influenced for centuries by Polish.

Swahili [50] 80,000 words (selection February 2009). The Swahili language is spoken along the East Coast of Africa. It is the lingua franca of many coastal nations. It uses the Latin alphabet.

Bahasa Melayu [64] 62,000 words (selection September 2009). Bahasa Melayu is the standard language of Malaysia. It has a common root with Bahasa Indonesia. However, Bahasa Melayu was heavily influenced by English while Bahasa Indonesia was influenced by Dutch during the colonial age.

Irish (Gaelic) [73]. 325,000 words (selection August 2007). The Gaelic language is a Celtic language spoken in Western Ireland. A slightly different variant is spoken in the Highlands of Scotland.

Welsh [74]. 921,300 words (selection July 2015). The Welsh language is the Celtic language of Wales, spoken by about 500,000 people (mainly bilingual in English). The current lexicon supports hyphens and apostrophes in words. The expression system supports re-writing of erroneous mutations, e.g., "hen gwlad" >> "hen wlad" or "o Castell" >> "o Gastell".

Greenlandic/Kalaallisut [69] 3,476,600 words (selection April 2025), is an East Inuit language spoken by 57,000 Greenlanders. The Greenlandic language adds particle to particle to words leading to a single word sentence. The Latin alphabet is used whereas the Canadian Inuit use their own script.

Macedonian [59] 324,000 words (selection April 2016). The Macedonian language uses the Cyrillic alphabet.

Albanian [57] 585,000 words (selection December 2017). The Albanian language uses the Latin alphabet. The Albanians call their language *shqip* and their country *Shqipëria*.

Maori [75] 30.000 words (selection March 2004). The Maori language is spoken in New Zealand and uses the Latin alphabet. A macron is placed above the vowels to differentiate between long and short vowels.

Zulu [79] 372.750 words (selection August 2024). The Zulu language is spoken in the Republic of South Africa and uses the Latin alphabet.

Xhosa [82] 171.000 words (selection May 2020). The Xhosa language is spoken in the Republic of South Africa and uses the Latin alphabet.

Arabic [83] ca. 5 million words (selection February 2025). The Arabic language has its own script and the orthography is mainly based on consonantal roots. These roots unfold into millions of words. The current version supports Arabic punctuation corrections, included the Arabic Comma, Arabic Question mark, and the Arabic Semicolon.

Azerbaijanian [85] 132.000 words (selection May 2010). The Azerbaijanian language is written in the Latin alphabet. It has much in common with Turkish.

Hebrew [86] ca. 5.5 million words (selection March 2025). The Hebrew language is written in Hebrew characters, mainly consonants. The orthography is based on roots of 3 radicals, which unfold into millions of words. In agreement with the "Haaretz" a separate Hebrew punctuation category has been added. Hebrew punctuations are not mirrored.

Persian/Farsi [87] 458.800 words (selection August 2023). The Persian language is written in the Arabic script, but being an Indo-European language vowels are important.

Urdu [88] 133.000 words (selection December 2023). The Urdu language is closely related to Hindi, but written in the Arabic script. Urdu and Hindi are Indo-European languages.

Breton [90] 653.000 words (selection July 2022). The Breton language is spoken in French Bretagne. It is a Celtic language once related to extincted Cornish in the United Kingdom.

Thai [68] 80.000 words (selection March 2008). The Thai language is the official language of Thailand. Thai has its own script. It is a syllable script and most vowels are written above the consonants. Thai itself is a tone language and the tone marks are also written above consonants. If a syllable has written vowels the tone marks float above the vowels. The words in a sentence are written without spaces and therefore a sentence has to be segmented (hyphenated) prior to spell checking.

Tagalog/Pilipino [72] 25.000 words (selection August 2004). The Tagalog language is the national language of the Philippines. It is an Austronesian language influenced by the Spanish language. 70% of the Philippine population speak Tagalog.

Vietnamese [89] 75,000 syllable words (selection February 2007). The Vietnamese language is the national language of Vietnam. Vietnamese is a tone language using the Latin script plus a lot of tonal marks. The Vietnamese mostly write single syllables but concepts (words) consist of more then one syllable separated by spaces.

Setswana [91] 50,000 words (selection August 2007). The Setswana language is one the languages of the Republic of South Africa. It is a Bantu language written in the Latin script. Setswana is spoken in Botswana and in South Africa in the Northern Cape, the central and western Free State and in the North-West Province.

Kurdish (Northern) [95] 90,000 words (selection July 2009). The Kurdish language 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.

Hindi [96] 541,000 words, 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. Althrough 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 [97] 1,681,100 words, 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 [98] 130,000 words, 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.

Malayalam [99] 781,000 words, 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. The Malayalam module supports chillu letters.

Bengali [100] 585,000 words, selection April 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 [101] 189,000 words, selection January 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 [102] 1,250,000 words, 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 loadwords from the Dutch, Portuguese, and English language.

Sinhala [103] 208,000 words, selection December 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 Dravidian influences.

Punjabi [104] 94,000 words, selection January 2018. The Punjabi language is spoken in 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 [105] 240,000 words, 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 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 [106] 30,000 words, 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 (see Hyphenator languages).

Luxembourgish [107] 200,000 words, selection December 2012. The Lëtzebuergesch language is spoken in the Grand Duchy of Luxembourg. The language/dialect descents from Mosel-Frankish, a dialect, linguistically close to High German and Limburgish. The population of Luxembourg is half a million only.

1.2. TALŌ's Hyphenator for the European, American, Australian, African and Asian languages

Silbentrennung, avstavning, stavenøglen, separación silábica, hifenização, woordafbreking, orðskipihlutinn, tavujako, stavelsesdeling, děleni

Although the *TALŌ hyphenators are not restricted to the European Community, the central point of activity for all languages lies in this community. This multilingual area is in need of proper hyphenation.

Refined

*TALŌ has refined the principles on which the hyphenators have been built. The general structure of the hyphenators is the same for all languages, but the realization of each hyphenator is based on a specific language model that exactly describes the solutions for that specific language. This language model is different for different languages. Moreover, the hyphenators have been optimized for increased performance. Response time has been reduced by a factor of 5, and it could be reduced even more in specific cases, while even quality could be improved further.
The hyphenator has been ported to Unices and is available for Unicode too.

New languages

New languages have been added. Therefore, the following languages or variants within a language are available:

English (UK, EU),	English (American, Canadian),
English (Australian, New Zealand),	English (South Africa),
Dutch,	Flemish, Surinam Dutch,
German (old spelling),	German (reformed spelling, 2x),
Swiss German (old spelling),	Swiss German (reformed spelling, 2x),
Asutrian German (old spelling),	Austrian German (reformed spelling),
French (phonological),	French (etymological),
Canadian French (phonological),	Canadian French (etymological),
Italian,	Spanish,

Catalan (nova ortografia),	Basque,
Portuguese (Iberian),	Portuguese (Brazilian),
Portuguese acordo (Iberian),	Portuguese acordo (Brazilian),
Danish,	Swedish (ck~, c~k, SAOL varieties),
Norwegian (consonant principles),	Norwegian (morphological principles),
Nynorsk (consonant principles),	Nynorsk (morphological principles),
Finnish,	Icelandic,
Estonian,	Afrikaans,
Greek,	Lithuanian,
Latvian,	Polish,
Slovene,	Russian,
Czech,	Turkish,
Azerbaijanian,	Basque,
Byelorussian,	Ukrainian,
Bulgarian,	Rumanian,
Macedonian,	Hungarian,
Serbian,	Croatian,
Bosnian,	Macedonian,
Frisian,	Galician,
Rhaeto-Romance,	Slovak,
Bahasa Melayu,	Bahasa Indonesia,
Thai,	Pilipino/Tagalog,
Greenlandic,	Maltese,
Sámi,	Irish (Gaelic),
Zulu,	Xhosa,
Azerbaijanian,	Swahili,

Kurdish (Northern),	Khmer (Cambodia),
Kazakh (Latin),	Latin,
Welsh,	Hindi (India)
Malayalam (India),	Tamil (India & Sri Lanka),
Bengali (India),	Marathi (India),
New languages in preparation	

Regardless of the language, our language model offers outstanding performance.

Special cases

Special cases of hyphenation are supported. For the pre-reform German language, consonant doubling (Schiffahrt -> Schiff-fahrt) and ck -> c-k (the last item for the old spelling only); for Swedish and Norwegian, consonant doubling (bussjåfør -> buss-sjå-før), for Dutch all spelling changes (e.g., chaletje -> cha-let-tje or papaatje -> pa-pa-tje), for Hungarian bi- and tri-character consonant doubling (e.g., asszimiláció -> asz-szimiláció) are included. The peculiarities of other languages are also supported: for Italian, elision hyphenation; and for French, *muettes* (silents) and long vowel groups in verbs. The hyphenators use the default script of a language, e.g., for New Greek, the WCP-1253 codepage is applied. A native Unicode version is available too.

Exceptions are not really needed, but...

Although exceptions are hardly needed, our exception functions even support special hyphenations, which is unmistakably an advantage for all languages that have a large number of compounds.

Moreover, different types of hyphenation functions are available, all optimized for their specific application. A word can be hyphenated with or without spelling changes (note that all types are marked), or a single hyphen can be placed in the best possible location, taking spelling into account.

A Win x86 or x64 DLL or an executable for developers

The *TALŌ Hyphenator will also be available both as a Windows 32- and 64-bits DLL directly attachable to an OEM program during run time and as a set of functions that can be linked to an OEM program. Shared libraries for Mac OS X and Linux have been developed too. Please inquire for the Windows x86 or x64 demo hyphenator.

*TALŌ bv, Lijsterlaan 379, 1403 AZ Bussum, the Netherlands
Tel.: +31 (0) 35 69 32 801,
E-mail: info@talo.nl, Website: http://www.talo.nl/

1.2.1. **Hyphenator: languages and varieties**

Dutch, Flemish, Surinam Dutch [1,56,57] (update January, 2025) supports the generally accepted spelling (the Netherlands), progressive spelling (Belgium), and the 1996 and October 2005 spelling reforms — four principles have been integrated in one hyphenator. Supports the Belgium, Surinam and Dutch idiom. The hyphenator recognizes compound boundaries and covers the Dutch idiom in the most extensive way.

English [2,3,58,59,60,61] (update January, 2025) supports phonetical hyphenation according to the world's most trusted dictionaries: Webster's New Twentieth Century Unabridged Dictionary (2nd edition), Webster's Third New International Dictionary and Longman's Dictionary of Contemporary English; hyphenators for the British, South African, Australian, New Zealand, Canadian, and American English are available[1]. The hyphenator handles the irregularity of the alternation of English strong and weak syllables.

German alt [4] and reformed [5] (update October, 2024) supports every characteristically German hyphenation according to the most recent Duden Rechtschreibung August 2006-2024 and Wahrig 2009. For German reformed two hyphenation styles have been implemented, one in agreement with the Duden(s) 1996-2004, and the other prefers the Duden 2024 hyphenations, if meaningful strict eingedeutschte syllables are used. The German hyphenator recognizes compound boundaries independent of the spelling reform. The new feature for "der Verwendung von Großbuchstaben ESZETT für ẞ" correctly hyphenates both "Schreibungsweisen". A special effort has been made to support medical and other scientific domains. The German hyphenators have been compared to over two million German, Swiss German & Austrian German words as an independent estimate of accuracy.

Swiss German old [6] and reformed [7] (update October, 2024) responds accurately to the typical Swiss German deviations and local idiom (including the ß to ss transcription).

Austrian German old [62] and reformed [63] (update October, 2024) responds accurately to the typical Austrian German deviations and local idiom.

French [8,9] (two versions, update February, 2025) accepts etymological syllabification according to Grevisse's "le bon usage." A second version accepts

[1] Hyphenation agrees with The Oxford Colour Spelling Dictionary (1995). Oxford's Dictionary, however, treats syllabication less extensively than the other dictionaries.

phonetical hyphenation rules recommended by the leading French linguist Nina Catach in Paris. Both versions use the new double layer technique to enable or disable hyphenation of muettes. Covers nearly all of the French idiom.

Canadian French [64,65] (two versions, update February, 2025) 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 versions use the new double layer technique to enable or disable hyphenation of muettes. Covers nearly all of the Canadian French idiom.

Spanish [10] Peninsular, Argentine, Mexican & Latin American (update March 2020) supports the official hyphenation rules as published by large dictionary publishers; covers the complete Spanish and Latin American idiom.
Be carefull with foreign words in Spanish! Do not hyphenate as "Burn the Wit-|ch" (El País, 23, 4-6-2016). It is not a syllable!

Italian [11] (update October 2024) supports phonetical hyphenation according to the Istituto Geografico de Agostini, including hyphenation of elisions (al-l'I.talia), conjugations, and declensions.

Iberian [12, plus acordo 68] and Brazilian [13, plus acordo 69] Portuguese (update April 2022) based on the vowel as the syllabic unit, but falling diphthongs and final diphthongs are kept unbroken. Both idioms are supported by a single hyphenator engine. Doubling of the hyphen is supported (repetir o hífen na linha sequinte).

Czech [14] (update July 2024), Slovak [35] (Update August 2024) supports the reformed spelling. Like in every Slavic language, a number of additive vowels and consonants exists, that have a large impact on hyphenation. Syllables that solely consist of consonants are supported (ji-tr-nice).

Swedish [15,16] (update February 2024) accepts the mekaniska principen, but compounded words are divided into their morphological roots. The considerable number of neologisms, and newly created forms, is a reason to use the *TALŌ hyphenator. You can switch between c-k or ck- hyphenation, and between within-word vowel-vowel hyphenation off or on. A third hyphenator model also supports morphological hyphenation as specified in the Svenska Akademiens ordlista över svenska språket (SAOL), however, *Mediespråksgruppen (tt.se) tar avstånd från SAOL:s avstavningssystem*, yet it has been introduced into the schooling system many years ago. The fact that Mediespråksgruppen rejected SAOL hyphenation might be related to technical hindrances to build a hyphenator just like that. Still it is possible given the use of a proper language model. The latest Swedish hyphenator (May 2023) agrees with SAOL 13/14 morfologiska avstavningar.

Finnish [17] (update June 2021) 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 its large inflection structure. Its resemblance to the neighbouring Estonian [23] might amaze you. The April 2018 hyphenator got a new extended language model, and learned the hyphenation principles embedded in more than a million words.

Catalan (nova ortografia) [18] (update May 2017) supports the mixed French and Spanish origins of the Catalan language. A peculiarity of Catalan, needing special care, is the l geminada (l·l). This version supports the spelling reform of October 2016.

Danish [19] (update April 2024) accepts the hyphenation rules of the Dansk Sprogncvns Retskrivningsordbog (2012), including the latest hyphenation changes. Compounds and newly created forms are supported; it even hyphenates Norwegian according to consonant rules.

Norwegian/Nynorsk [20,21] (update February 2024) accepts consonant rules (20) or the morphological rules of the Nordisk institutt of the University of Bergen (21). The different sets of rules using principles of pattern recognition also apply to both Bokmål and Nynorsk; each hyphenator recognizes both languages.

Icelandic [22] (update November 2023) accepts morphological rules which separate the attached article and nominative, dative, accusive, and genitive cases and is capable of dividing a pile-up of compounds.

Estonian [23] (Update January 2020) behaves like the Finnish hyphenator and is capable of hyphenating Estonian compounds and diphthongs correctly. However, there are more diphthongs in the Estonian language than in Finnish, which increases complexity.

New Greek [24] (Update February 2016) is tuned to the Greek script, the Elot codepage. It hyphenates modern Greek, but also the Classical Greek varieties. Present-day Greek is evolved and flavoured with diacritics.

Polish [25] (Update February 2016) hyphenation of the Polish language is hindered by an immense amount of consonants, quite often unpronounceable for non-Polish speakers. However, the hyphenator has been fully adapted to handle these difficult syllables.

Latvian [26] (Update February 2016) is tuned to the properties of Baltic languages. Words take many forms. Latvian uses additional consonants and vowels, which are recognized by the hyphenator.

Azerbaijanian [27] (Update February 2016), also called Azerbaijan, is one of the new Transcaucasian republics that used to be part of the USSR and is independent today. Azerbaijanian is related to Turkish. The Azerbaijani now use a Latin script. There is no standard script yet, but this does not affect the applicability of *TALŌ's hyphenator principles.

Turkish [28] (Update February 2016), present-day Turkish is spoken in SW Asia, but in the past the Turkish region extended to 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 [29] (Update February 2016) is one of the Baltic languages that is extensively declined. Lithuanian uses (semi-)diphthongs, palatals, and affricates, which have been taken into account for hyphenation.

Afrikaans [30] (Update August 2024). 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; however, Afrikanization of spelling has made hyphenation irregular, which in turn calls for accurate tools. The Afrikaans hyphenator takes all Afrikaans peculiarities into consideration, including diaeresis hyphenation.

Russian [31] (Update July 2014) accepts Cyrillic characters, but this 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 [32] (Update February 2016). 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 to all those peculiarities of real-life language.

Hungarian [34] (Update February 2016). The Hungarian language has lost many of its Uralic characteristics and many words have been borrowed from the Turkic (subfamily of Altaic languages) and European languages. The language is flavoured with compounds and special hyphenations (briddzsel -> bridsz-dszel).

Bahasa Indonesia [33], Bahasa Melayu [49] (Update July 2014). The Bahasa Indonesia (Standard Indonesian) and Bahasa Melayu (Standard Malayan) are Austronesian languages 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.

Byelorussian [37] (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 uses the Cyrillic alphabet.

Ukrainian [38] (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.

Bulgarian [39] (Update July 2014) is spoken by 90 % of the population of Bulgaria. The modern Bulgarian alphabet is the same as the Russian alphabet.

Rumanian [40] (Update June 2009) is the national language of Romania. It is a Romance language that uses the Latin alphabet. One third of all Romanian words are of French origin. The hyphenator accepts words with a s-comma below and t-comma below.

Serbian [41] (Update July 2014) or *srpski jezik* uses the Cyrillic alphabet. Serbian is closely related to Croatian, however, Serbian characters are written with single symbols Џ, Љ, Њ (Dž, Lj, Nj). Like words in any Slavic language Serbian words can have many prefixes to be hyphenated.

Croatian [42] (Update July 2014) or *hrvatski jezik* uses 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 [55] (Update July 2014) or *Bosanski Jezik* exists since Bosnia and Herzegovina became independent. Bosnian has developed its own identity, using the Latin alphabet but is closely related to Croatian.

Macedonian [57] (July 2007) is the principal language of the new nation of Macedonia, it is closely related to Bulgarian and uses the Cyrillic alphabet.

Galician [44] (Update July 2024). The Galician language is now spoken in Spanish Galicia, situated north of Portugal. It is a Romance language related to Portuguese. The orthography differs slightly from Spanish.

Frisian [43] (Update March 2024) or Frysk is spoken in Friesland the northernmost province of The Netherlands. Frisian is closer to English than to Dutch.

Rhaeto-Romance [45] (February 2002) is the collective of three Romance dialects spoken in the northeastern Italy and southeastern Switzerland.

Tagalog (Pilipino) [50] (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.

Greenlandic/Kalaallisut [47] (March 2025) is an Inuit/Eskimo language spoken in Greenland. Greenlandic uses the Latin alphabet. Words can be very long and one word can be a complete sentence.

Slovene [36] (Update May 2015) or *Slovenski jezik* uses 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 [54] (Update October 2019). 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.

Maltese [52] (March 2005) is one of the official languages of the island(s) of Malta, it is a Semitic language that uses the Latin alphabet, including <ċ> <ħ> <ġ> and <ż>, the variety of root words has a great impact on hyphenation.

Sámi [48] (Update July 2014). The hyphenation agrees with the Nord Sámi language as spoken in Finnmark county in the north of Norway.

Hebrew [66] (December 2006) is written in Hebrew consonants only and therefore hyphenation is partially uncertain. Within this uncertainty the hyphenator accepts graphical hyphenations.

Irish/Gaelic [67] (December 2006) is a Celtic language mainly spoken in Ireland.

Zulu [70] (February 2016) is a Bantu language mainly spoken in the Republic of South Africa. Zulu is one of the 11 South African languages and is very different from Afrikaans and the other Indo-European language and so is hyphenation: be-nga-ka-la-li, ma-fu-ngwa-se.

Xhosa [71] (September 2008) is a Bantu language mainly spoken in the Transkei, Ciskei and Eastern Cape regions of the Republic of South Africa. Xhosa is one of the 11 official South African languages. It is very different from Afrikaans and the other Indo-European language, and so is hyphenation: i-si-Mpo-ndo, ye-Bha-nki.

Swahili [72] (March 2009) is a language 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.

Kurdish (Northern) [73] (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.

Khmer (Cambodia) (November 2009) belongs to the Austroasiatic languages. Khmer has its own script known as Aksar Khmer. In Khmer no spaces are inserted between words. Yet words have to be segmented, even unknown words.

Kazakh (Latin) (Update May 2010) belongs to the Turkic family languages. Kazakh is written in the Cyrillic, Arabic or Latin script. An official transition to the Latin script could happen in a 10 to 12 year period. Despite being developed for the Latin script Cyrillic hyphenation is nearby.

Latin (March 2011) is an extinct language, but still spoken in the Roman Catholic church. Together with Greek, Latin had an enormous influence on nearly any European language. The Latin hyphenator divides Latin prefixes, suffixes, and enclitics.

Welsh (July 2015) is a Celtic language which is hyphenated according to morphological principles. In addition a lot of sound changes (mutations) replicates these principles.

Hindi (May 2021) belongs to the Indo-Aryan languages spoken in the Republic of India. It is written in the Devanagari script. Dependent vowels are written in conjunct with consonants. A new language model for hyphenation has been developed which prevents broken syllables (like "nex·t"). The principles of the Indic hyphenators are position independent.

Malayalam (April 2021) is a Dravidian languages using the Malayalam script, which has common characteristics with the other languages of India. There exist a traditional and a reformed orthography. The Malayalam language model prevent broken dependent vowels and keeps final closed syllables together (so not hyphenated as "nex·t").

Tamil (April 2021) is spoken in south eastern part of India (Tamil Nadu) and in Sri Lanka. It is a Dravidian language related to Malayalam, but the orthography is quite different. Double consonants (t·t, k·k) are hyphenated in between. They are marked with a Virama sign. Hyphenation is modelled by a Tamil language model.

Bengali (May 2021) is an Indo-Aryan language spoken in the region of Bengal of India. The Bengali Script also applies to Assamese Language located in the Assam region of India. The Bengali hyphenator can hyphenate Assamese too.

Marathi (May 2021) is an Indo-Aryan languages which is written in the Devanagari script, just as Hindi. However, there is a large difference with Hindi, in words and grammar.

1.3. **Hyphenation and spelling worries: how to de-Babel-ize?**

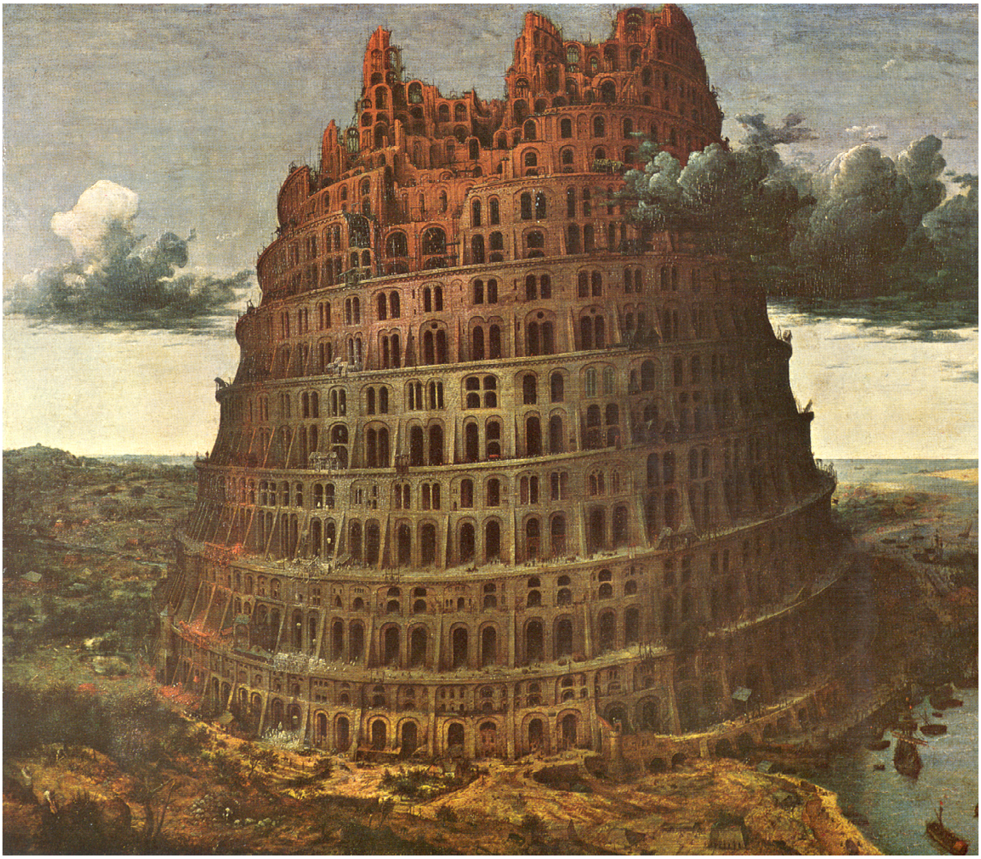


Fig. 1.1: The Tower of Babel by Pieter Brueghel the Elder (1563)

All tongues became different after the Babylonian Ziggurats were hit by God's fury (fig. 1.1). Nowadays an even greater variety of languages exists, but recent studies of the mitochondrial DNA genes demonstrate a greater common background in humans that is in line with earlier comparative language studies[1].
Why are humans so successful in communication? By which language were the 7 daughters of Eve communicating? The answer remains open. Did they live simultaneous and within reach to each other? The answer is no. Ursula lived 45,000 years ago, Xenia 25,000, Helena 20,000, Velda 17,000, Tara 17,000, Katrine 15,000, and Jasmine at the end of the Ice Age.

Which language technology summarizes the common background of our ancestors?

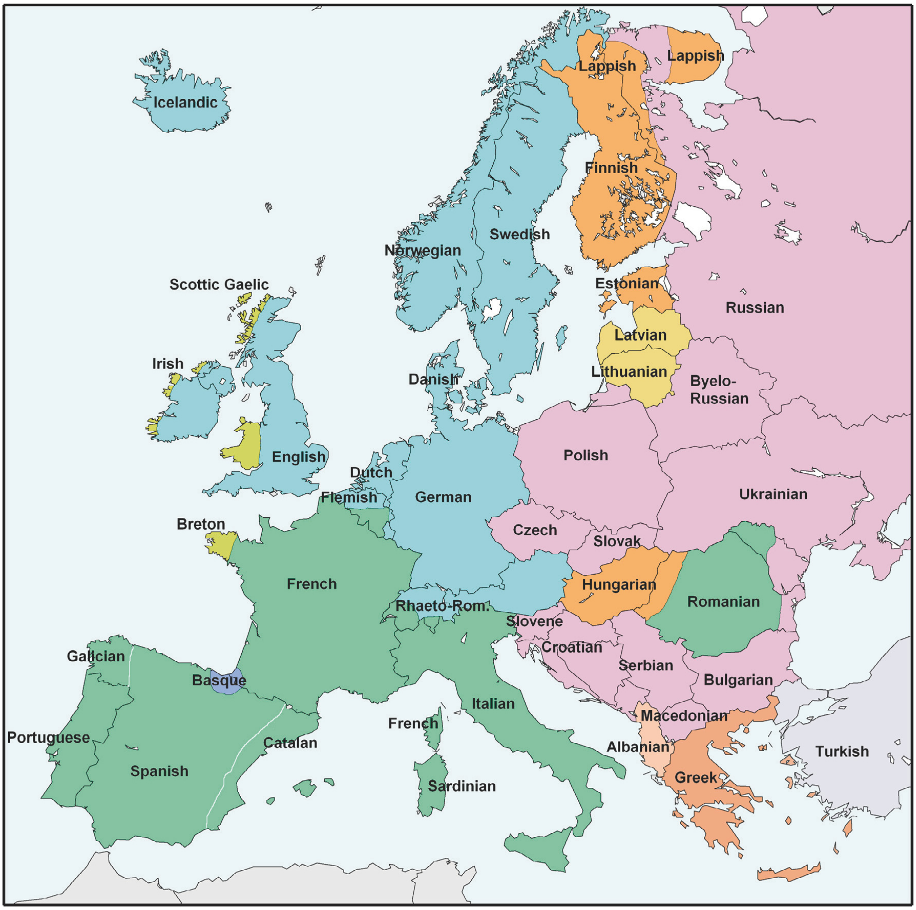


Fig. 1.2: The languages of Europe: Most of the European languages belong to the Indo-European linguistic family, except for the Finno-Ugric languages Finnish, Lappish, Estonian and Hungarian, Turkish and Basque. The language borders are not always similar to the country's borders[†].

Today's great variety in languages, each having its own peculiarities, has significant consequences for applications in publishing. However, the diversity in tongues does not necessarily mean that hyphenation and orthography problems are unsolvable. Probably it is not the study of computer technology that should determine the lingual processes of applications in publishing, but rather

[†] Inspired by De Grote Taalatlas (The Atlas of Languages), Schuyt & Co, 1998. "The Atlas of Languages" not only puts the European languages together, but also outlines relations between all languages of the world.

the applications should be based on a broad survey of languages that guides language technology in a new direction. *TALŌ’s spellers and hyphenators are based on language models, the basic elements forming a database of *quarks*, or particles of meaning.

1.4. **The complexity of languages**

Languages use a complex mechanism to make meaningful words. Some languages catenate particle by particle, some languages have sound changes, other languages are basically analytic and grammatical structure is determined by word order. These mechanisms have a significant impact on language technology.

Finnish is a non-analytic language and combinations of words make thousands of new words, each of them chained with particles added to the end of the word. The function of these particles looks like the function of prepositions.

In Finnish hyphenation, compounds are divided into their basic elements, but it is not always the lemma which divides the word.

Jumala (god) in the word *jumalakeskeinen* is divided in *jumala-keskeinen* (theocentric). However, the word *jumala**n**ilma* is the dreadful weather and is God’s possessive case, divided as *jumalan-ilma*. Considering these fine differences in orthography, obviously some spellers are not able to differentiate between *jumalailma*, *jumala-ilma* and *jumalanilma*.

Other examples are *kruunu~kurki* (crown crane) versus *kruunun~prinssi* (Crown Prince); *kuoli~aaksi* (to death) versus *kuolin~apu* (euthanasia); *kuvakasetti* (video cassette) versus *kuvanheitin* (projector) — *kuvaheitin* would be a spelling mistake.

jumalakeskeinen	
jumala**n**ilma	NOT jumalailma
kruunukurki	
kruunu**n**prinssi	NOT kruunuprinssi
kuoliaaksi	
kuoli**n**apu	NOT kuoliapu
kuvakasetti	
kuva**n**heitin	NOT kuvaheitan

German compound formation confuses spellers and hyphenators, too. In some compounds the possessive case is used, whereas in other compounds it is not: *Bahn~hofvorstand* (railway station board) versus *Bahnhof**s**buchhandleitung* (manual for the railway station). For hyphenation, the compound word *Empfangnahme* (the noun case for receiving) follows the main rules, but the word *Empfang**s**antenne* (receiving antenna) conflicts with the main rules and therefore causes variation in the hyphen positions. Spelling cannot logically be derived from lemmas. The possessive case is irregular, and compounds usually denote more than their constituents. *Geisteswissenschaft* (social sciences) is a regular German word, but despite its German look-alike nature, *Geisternwissenschaft* is not. The latter word is accepted by many spellers as a correct orthography: however, the word is nonsense, and it has to be flagged by a speller!

Bahnhofvorstand	
Bahnhof**s**buchhandleitung	NOT Bahnhofbuchhandleitung
Empfangnahme	
Empfang**s**antenne	NOT Empfangantenne
Geiste**r**erscheinung	
Geiste**s**wissenschaft	NOT Geisternwissenschaft

Similar confusion in compounds occurs in Estonian, Dutch, Danish, Norwegian, Swedish, Icelandic, Frisian, Afrikaans, Hungarian, and other languages that use compounds.

The English language was influenced by the Anglo-Saxons, Danes, Norwegian Vikings, and Norman French-civilized Vikings and is considered to belong to the Germanic languages.
The orthography of compounds as “kangarooapple” versus “kangaroo beetle” or “load line” versus “loadstone” (magnetite) differs in respect of an uncertainty about where to use a space or a hyphen. We now detect the omission of a space and errors in hyphen usage; in future, we will analyse compounds and expressions beyond the boundary of unjustified and justified spaces.
American English — like British English —, the communication vehicle between so many people — has its own outstanding properties.

On the other hand a lot of similarities between languages exist (see fig. 1.3).

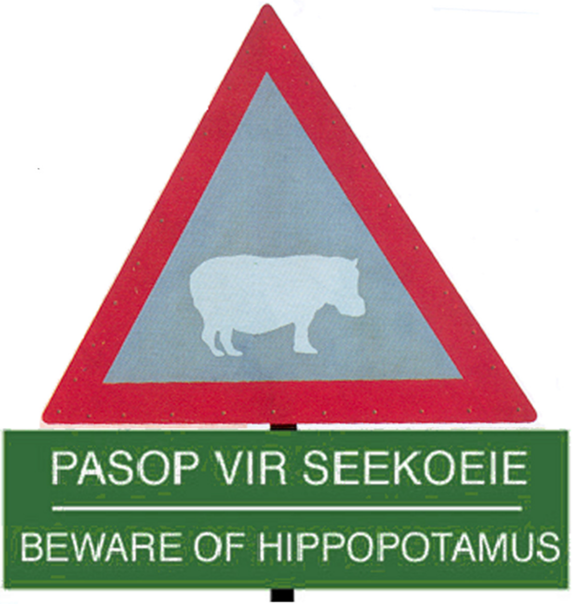


Fig. 1.3: There are two ways of naming hippos; the Greek word *hippopotamos*, earlier *hippos ho potamios* 'river horse', *hippos* 'horse, *potamos* 'river' and the translated version of the Greek word became in French *hippopotame*, in Catalan *hipopòtam*, in Spanish and Portuguese *hipopótamo*, in Basque *hipopotamo*, in Italian *ippopotamo*, in Polish *hipopotam*, so all these languages have kept the original Greek word; the Swedish *flodhäst*, the Danish and Norwegian *flodhest*, mean river horse, the Dutch *nijlpaard*, the Frisian *nylhoars*, the German *Nilpferd*, the Latvian *nilzirgs*, the Serbo-Croatian *nilski konj* add the source of origin, the river Nile, for Afrikanders it simply is a cow that loves the see.

The linear model

Compounds not only have an influence on hyphenation (before or after), but they also have consequences with respect to the theory applied in prepress technology, the foundation on which the technology is based[2,3].

Most hyphenation technologies applied by American companies are based on a linear model. This model is freely available from Donald Knuth's TEX environment. This model applies the concept of uncertainty in order to avoid risky hyphenations. However, uncertainty simply implies making errors!
Uncertainty of hyphenation is equally spread over the word that is to be hyphenated. The linear model steps through the word-matching patterns, summing up the uncertainty of the patterns. Finally, only patterns with acceptable certainty

are applied.

This model is falsely applied to the Germanic and Finno-Ugric languages. The model is imperfect for other languages, too. The main problem is caused by the fact that each part of the compound has its own and independent distribution of uncertainty. So when the two sets of distributions encounter, uncertainty of hyphenation increases. The main cause of this problem is caused by word formation properties. At the compound boundary, uncertainty should be reduced, but the linear model absolutely fails to estimate the boundaries correctly.

Why you should use lingual *Quarks*!

*TALŌ's hyphenation technology is not based upon linearity and uncertainty. The hyphenation technology is based on a language model that describes the possibilities and impossibilities within a target language. On the basis of this model, particles of meaning, just like *quarks* in physics, are determined, and yet these quarks are related in terms of meaning to its very origin. In James Joyce's words:

> *— Three quarks for Muster Mark*
> *Sure he hasn't got much of a bark*
> *And Sure any he has it's all beside the mark*
> *But O, Wreneagle Almighty, wouldn't un be a sky of a lark*
> *............*
> *............ That song sang seaswans.*
> James Joyce's Finnegans Wake, 1939[4].

These lingual *quarks* are the smallest carriers of meaning in words. When hyphenation is applied, these *quarks* are detected within the context of the language model. They tell us how to spell and where to hyphenate, with certainty!

Every language has its own peculiarities: e.g., the German/Dutch *sch* grapheme has become *sk* in Afrikaans and Danish (Schiff/schip versus skip/skib). So German *quarks* differ from Danish *quarks*. Instead of using a blind pattern recognition system, *TALŌ has developed a language model for each language. These language models are tuned into solving the boundary of compounds by recognition of lingual quarks. The results are highly accurate hyphenators that do not need a dictionary containing exceptions. The hyphenators are also capable of hyphenating future neologisms because these new words are also constituents of already-known lingual quarks. Applying this technology, *TALŌ has created the Hyphenator XT and Smart Hyphen. This hyphenator includes nearly all European languages. For the Asean market Bahasa Indonesia, Bahasa Melayu, Tagalog and Thai have been added.

The fundamentals of the spellers are also based on the language model. The model enables the speller to estimate alternatives that are look-alikes to a potential error. However, as demonstrated above, word formation is too irregular and is often linked to a special meaning playing an important role. Moreover, words and their formation are also based on history. Therefore, huge corpora, dictionaries, are a prerequisite, either to get the correct case or the nearest example. Huge corpora call for extremely fast access, which again is based on the language model, thus giving the optimal entries. Several improvements have been added to let the speller learn from past experience and to forget irrelevant information.

Dr. Jaap Woestenburg, *TALŌ bv, Bussum, The Netherlands

1.5. **References**

1 Bryan Sykes, 2001. The Seven Daughters of Eve: The Science That Reveals Our Genetic Ancestry, W.W. Norton.

2 J.C.Woestenburg, 2005. A Note on Hyphenation, *TALŌ bv, Bussum.

3 J.C.Woestenburg, 2006. Hyphenation and spellchecking in InDesign, Smart Hyphen & Smart Speller WoodWing Publishing Conference, Cancun, Mexico, November, 9 - 10, 2006.

4 James Joyce, 1939. Finnegans Wake, Faber and Faber, London.

CHAPTER 2

2. The hyphenation rules for the European languages

The Greek word *huphen* literally means under one. It is derived from *hupo* under, and *hen* the neuter accusative case of heis, a **mark**. A hyphen marks a syllable boundary. It is used to break up words at the end of a text line.

The hyphenation principles differ from language to language. Some of these differences are related to differences in the nature of syllables. Other differences arise from the unifying nature of compound words.

In order to illustrate what hyphenation is, the basic rules for the European languages are presented on the fly. For some languages more than one hyphenation principle is supported.

2.1. Overview Hyphenation Rules

In the following you will find an overview of the hyphenation rules per language module, in alphabetical order.

2.1.1. The Afrikaans hyphenation rules (Skeiding van Woorddele)

Afrikaans occupies an important position in the row of languages in the Republic of South Africa. It is spoken by 15% of the population and even overshadows English (9%) in number of native speakers[1]. However, disputes are ongoing concerning its position in public life. On the other hand, the economic position of the Afrikaners is without doubt one of dominance.
Although Afrikaans has developed independently during 4 centuries, Afrikaans hyphenation has much in common with the Dutch languages. The edition of the Taalkommissie “Afrikaanse Woordelys en spelreëls” recommends the following guide lines for the hyphenation of Afrikaans words. It is primarily considered as a typographical matter and probably hyphenation problems let “die Taalkommissie” to the recommendation “not to hyphenate when possible”. The following guide lines exist[2]:

- In many cases the pronunciation of the syllables guides hyphenation. This is relevant for a number of compounds were the original compound boundary is lost: *aar-tap-pel, tran-sak-sie, hui-ge-laar, va-naand, voor-taan, etc.*
- Readability is important, especially in respect to prefixes. We hyphenate *her-enig* and not *here-nig*, *ver-ewig* and not *vere-wig*.

- A single consonant between vowels is transferred to the next syllable. The ch, th and gh are considered as a single vowel: *ba-nier, brui-ne-rig, da-gha, Bou-cher, spa-ghet-ti, So-tho*.

- A double consonant between vowels is hyphenated between the two consonants: *as-pi-rant, ag-nos-ties, drif-tig, kal-meer, plan-kie*. exceptions are the double consonant sj which is not separated: *bro-sju-re, de-ta-sjement, ma-sjien*, except for compounds: *dis-junk-tief, plaats-ja-pie*.

- . Hyphenation of the diminutive tjie depends on the lemma: *gordyn-tjie, nael-tjie, but kaart-jie, eelt-jie*.

- Hyphenation of three consonant between vowels depends on the combination: *ak-trise, ban-krot, in-fluensa, kon-stant, on-skuur, against simp-toom, tink-tuur, intelligent-sia*. For some words two possibilities are acceptable: *bar-stens or bars-tens, bron-stig* or *brons-tig, vor-stin* or *vors-tin*.

The real hyphenation problem is found in compounds. Afrikaans closely resembles the Dutch languages with an overwhelming number of neologisms, being highly irregular: *be-stuur-stel-sel, be-stuurs-ta-fel, blom-aar, blom-ar-ti-sjok, bol-ag-tig*. The typical Afrikaans orthography adds an additional number of irregularities to the hyphenation problem.

The solution depends on an accurate language model for the Afrikaans language, which reduces conflicts due to the irregularities.

2.1.2. The Basque hyphenation rules (hitzen ebaketa silabikoa)

Basque is one of Europe's most exotic minority languages, probably genetically unrelated to any other language in the world. It is spoken by two-thirds of a million people, constituted in what is now called Euskal Herria — the Basque homeland. Basque is a highly inflected language in which the prepositions are added to the end of nouns. Adjectives are inflected too[3].

- Long consonant groups do not exist in the Basque language, so in general one consonant between vowels belongs to the next syllable: *ta-berne, e-li-za, Do-nastia, za-har, etc*.

- two consonants are usually hyphenated in between: *on-doan, en-paran-tza, eus-karaz, taber-na, joan-go, etc*. The rr, tz, ts, and tx are never divided. These combinations almost sound like a **t** followed by **z**/**s**/**x**, but each combination makes a single sound[4]: *hondar-tzan* (at/on the beach), *hondar-tzatik* (from the beach), *e-txean* (at home), *hizkun-tza, zor-tzi-hi-ru, etc*.

- Diphthongs are never divided. There are two groups of diphthongs: ai, ei, and oi and au, and eu. The *ai* in *Bai-one*, the *eu* in *eus-ka-raz*, and the *au* in *hau-xe* is kept together.

Word particles can be added to each other, forming long compounds. These particles modulate meaning and they can be attached to nearly all basic words: *estu, estualdi, estuario, estutasun, estudiamte, estudia(tu), estudio, estugarri, estugune, estuki, estune, estura, estu(tu)*. Moreover, each of these words can be fully inflected: *es-tu* (congestion), *es-tua, es-tu-ak, es-tu-an, es-tu-aren, es-tu-e-nak, es-tu-e-ta-ko*, and there are a lot more of these inflections. Inflection depends on living or non-living word properties.

2.1.3. The Catalan hyphenation rules (trencar un mot al final de ratlla, nova ortografia)

Catalan is spoken by a population of 7,000,0000 people around the Mediterranean coast, the South of France and north eastern part of Spain. The Catalan language has been suppressed for centuries. During and after the French Revolution Catalans were persecuted on behalf of this language usage. The Spanish State ended the interdiction in 1939[5]. The princedom of Andorra has been the only state which officially accepted Catalan by law. The Catalan language has common elements both with the French and Spanish language. Words are hyphenated as follows[6]

- When a word contains one consonant between two vowels, the hyphen comes before the consonant: *ca-ber, cai-man, cà-tar, ca-sa, etc*

- When a word contains two consonants between two vowels, the word is hyphenated between the two consonants: *can-vi, cis-ma, cis-tell*, except for the dissyllables *p, b, c* or *g* followed by *l* or *r* and the *t* or *d* followed by an *r*, which are hyphenated before the dissyllable: *co-bla, so-bre, fi-bra, jo-glar, pe-ple, etc*.

When there are three or more consonants between vowels the hyphen comes after the second consonant, except for dissyllables of the preceding section: *abs-ten-ció, subs-tàn-ci-a, com-plir, res-clo-sa, sas-tre, mons-tre, etc*.

The dissyllables *rr, ss, l·l* and the intervocalic ix between vowels are hyphenated between the two consonants: *car-rer, pos-ses-si-óm, rei-xa, col-le-gi*.

The digraphs *gu* (guerra), *qu* (esquerrà, quitrà), *ig* (rebuig), *ix* (a final ending as in peix, coix), *ny* (allunyar), *ll* (palla) are hyphenated before the digraph: *al-guer, es-quer-rà, a-llu-nyar, etc*.

Prefixes and compositions are divided like: *an-hidre, con-hort, in-alternable, des-eixit, ex-ornar, sub-altern, nos-altres, vos-altres*.

Syllables containing two or three vowels, like diphthongs (a combination of strong and weak vowels) are not hyphenated. Two classes exist:
a) rising diphthongs (*i* and *u* followed by a strong vowel *a, e* or *o*): *io-de, io-gurt, hie-na, jo-ia, cre-uen*.
b) falling diphthongs (the strong vocals *a* or *e*, followed by a weak *i* or *u*: *em-pai-tar, fei-na, al-moi-na, su-au, seu-re, etc*.

2.1.4. The Danish Hyphenation rules (orddeling men ofta stave.nøglen)

The Danish hyphenation rules are quite similar to the hyphenation rules of the other Scandinavian languages: the extensive use of compounds and the undefined article attached to the end of the noun. Given the extensive use of neologisms words can take nearly any form. Like in the Swedish and the Norwegian languages more than one principle of hyphenation is allowed, consonant hyphenation and hyphenation according of meaningful word parts (orddelingsprincipperne)[7].

- The most important rule is that words are divided into their compound elements (orddeling ved betydningsbærende orddel): *bage-pulver, grill-bar, kakkel-ovn, spå-kvinde, mælke-karton, studenter-eksamen, stik-prøve, øl-oplukker, etc*. Do **not** make hyphenations as *lampe-tarm*![8]

- Words are hyphenated after a prefix: *an-klage, af-brud, a-typisk, be-undre, bi-falde, des-orienteret, eks-portere, er-kende, for-stå, fore-skrive, gen-rejse, in-direkte, sam-handel, u-gerning, und-gå, ur-komiske, van-ære, etc*. There are small differences: *an-alfabet* and *a-nalyse*, *a-norak* and *an-ordning* (note: Although a single vowel is a correct syllable, it should not be used in printing).

- Words are hyphenated before a suffix beginning with a consonant: *styr-bar, hellig-dom, mand-haftig, uviden-hed, demo-krati, far-lig, tro-skab, bis-pe-dømme, frede-lig, femino-logi, stands-mæssig, forsøgs-vis, etc*.

- Certain categories of suffixes beginning with a vowel are also hyphenated before the suffix: *grin-agtig, lærer-inde*. It is also allowed to hyphenate before suffixes as *bestyr-else, telefon-ere, skib-et*, but the hyphen usually comes before the consonant: *besty-relse, telefo-nere, ski-bet*.

- One should hyphenate between two vowels which are not diphthongs: *bureau-et, di-æt, fi-asko, spi-on, tre-er, akti-er, etc*.

a Words containing a single consonant between vowels are hyphenated before the consonant : *a-lene, ba-jer, besty-relse, se-xet, pro-cent, kro-ne, fo-nem, etc*.

b Words containing two consonants between vowels are hyphenated between the consonants: *ar-ving, dan-ne, dat-ter, æb-le, grif-fel, etc*. However, the consonant group *ch, ck st sk dh, gh, sc, sch, sh, sj* is not divided: *bro-chen, dou-che, check-en, bud-dhisme, spa-ghetti, fa-scisme, di-sciplin, rut-sche, gei-sha, bol-sje, etc*.

c Words containing double consonants with an r and especially those originating from French are kept together: *pro-blem, neu-tral, ag-gressiv, ak-klimatisere, sup-pleant*.

d When words contain three and more consonants between vowels at least one consonant goes to the next syllable: *bredskuld-ret, brevveks-ling, gurg-le, etc*.

Dividing attached articles

The Danish hyphenator follows the consonant style of hyphenation for non-compound words. Prefixes are separated, but suffixes which are not felt in speech are hyphenated according to the rule 5, e.g., *afbæ-rerne* and not *afbær-erne*.

The latest hyphenator version (July, 2013) includes the hyphenation changes of the Dansk Sprognævn (2012).

2.1.5. The Dutch hyphenation rules (Woordafbreking)

The edition of the Dutch government's "Woordenlijst van de Nederlandse taal" recommends the following rules for the hyphenation of Dutch words[9a,b].

- Between two vowels (or groups of vowels) which immediately follow each other and which are not double vowels: *be-amen, brij-achtig, bui-ig, kri-oelen zwaai-en, draai-ing etc*;

- Before a word or stem of a word, being part of a compound word like: *doorn-struik, kwaad-achtig, weer-spannig, ziels-bedroefd etc.*;
- After a prefix: *be-horen, er-kennen, ge-lag, her-ademen, on-gelukkig, ont-zien, ver-kennen, voort-razen, etc.*;
- Before the suffixes -aard and -achtig and before suffixes beginning with a consonant: *blood-aard, blauw-achtig, boom-pje, dek-sel, naai-ster, etc*. Exceptions: Words as *do-laard, grijn-zaard* en *vein-zaard*. Hyphenation of the superlatives *beste, meeste* and *dwaaste* are dealt with rules presented below.

When hyphenation is not determined by the above rules the following rules apply:

- One consonant between vowels (ch is considered a consonant) will be transported to the next syllable: *ve-len, wa-ren, na-men, la-chen, etc*.
- When a word contains two consonants between vowels, the hyphen will be placed between the consonants and the second consonant will be transported to the next line: *konin-gen, bes-te, var-ken, bur-ger, zus-ter, sys-teem, etc*. **but** it is *pa-sja, cra-shen, fini-shen*, and *pu-shen* according to the new Dutch spelling rules of October 2005[9a,b].
- When a word contains more than two consonants, as many consonants as possible will be transported to the next line: *ek-ster, ad-mi-ni-stra-tie, art-sen, beamb-ten, erw-ten, etc*.

Notes!

- A single x between vowels is not hyphenated (*exa-men, exo-dus*), except when the *x* is integrated in the first part of a compound word (*telex-apparaat*).
- When hyphenating words containing vowels with two dots, the dots will disappear from the beginning syllable: *beëindiging* → *be-eindigen*, *definiëring* → *defini-ering*, etc.

Dividing compounds

To a native Dutch speaker, hyphenation of compound words like verf-laag or beurs-prognoses comes automatically, as he will create separate parts, probably through mental linguistic pattern recognition process. A computer program, however, which is by no means a specialised system, does not comprehend the meaning of these words, thus possibly causing the following hyphenations: *ver-flaag* or *beur-sprognoses*. The computer will be even more confused when an **s** or **e(n)** comes between two parts of a compound word: *afdeling[s]chef, paard[en]bloem, eend[en]kroos, koud[e]oorlogsdenken etc*. The extra character is sometimes placed in the former syllable, or an extra syllable will be formed. The interpretation of this linking-s is not uniform, note the words: *ring-steken, belasting-stelsel*. The words staat and recht in *staats-wege, staat-siekleed, rechts-wege, recht-streeks* emphasize the hyphenation problem of the Dutch language. Some erroneous hyphenations disturb the meaning of a word: *recht-sextremistisch*, or *binnenvaarts-chip*.

Special hyphenations

Diminutive formations like (*-tje*) with a stem in which the vowel is doubled are hyphenated without doubling the vowel: *papaatje becomes papa-tje*, *mamaatje* becomes *mama-tje*, *dynamootje* becomes *dynamo-tje*, *pianootje* becomes *pi-ano-tje*, *menuutje* becomes *menu-tje*, *slaatje* becomes *sla-tje*, *skietje* becomes *ski-tje*, but *dineetje* becomes *diner-tje*. On occasion apostrophes are hyphenated: *AOW'er* becomes *AOW-er*, *sms'en* becomes *sms-en*. Some words have multiple possibilities, *diplomaatje* could originate from *diploma-tje* or *diplomaat-je* (a certificate or diplomat) or *vlootje* could originate from *vlo-tje* or *vloot-je* (flea or fleet).

Aesthetics

Some syllables consist of a single vowel at the beginning of a word: *a-gent, a-nemoon, a-gaat*. These hyphenations do not improve the readability of a text. It is therefore wise to avoid them. The same applies to words like *motora-gent* and *bosa-nemoon* which should be hyphenated as follows: *motor-agent* and *bos-anemoon*.

2.1.6. The English hyphenation rules

English words are hyphenated phonetically[10,11,12]. When pronunciation changes hyphenation changes accordingly and accurately parallel the pronunciation, e.g., *bi-o-log-ic-al, bi-ol-o-gist, bi-o-nom-ics* and *bi-on-o-my* or *ab-sorp-ti-om-e-ter* and *chro-nom-e-ter* but *cen-ti-me-ter* and *hec-to-me-ter*.

These changes in pronunciation, even within related words, make it impossible to have standardised hyphenation of prefixes: *pre-se-lect, pre-sen-sion, pres-en-*

ta-tion, pre-sent-i-ment, pre-serv-a-tive pres-er-va-tion. The same applies to suffixes, e.g., (-cally): *pre-his-tor-ic-al-ly, pro-lif-i-cal-ly*.

There even are problems with frequently occurring suffixes such as -ed and -ing which are often used as hyphenations. The syllables in the words *a-maz-ed-ly* and *a-mazed*, however, are quite different.

Some ambiguities must also be taken into account: to **de-sert** is not a **des-ert**. These indistinguishable homonyms should not be hyphenated.

Hyphenation is very irregular and some dictionaries advise users, in case of uncertainties not to hyphenate. This might apply to non-native speakers of the English language. However, in the case of the *TALO-hyphenators one does not necessarily have to master a language since its linguistic patterns are programmed to detect the fine differences in pronunciation and syllabification.

Despite irregularities, guidelines do exist!!!

- Words that contain the combination vowel-consonant-vowel but do not sound like a single syllable are not hyphenated, e.g., base, were, wife. These words are classified as one-syllable words.

- When a short vowel is followed by one consonant, the word is hyphenated after the consonant (ch, ck, sh, ph, th are considered as being a single consonant): *per-il, jeal-ous, devel-op, min-ic, pun-ish, priv-ilege, prim-i-tive, pick-erel, gov-ernor, etc.* whereas we write *divi-sion*.

- Words with a short vowel followed by two or more consonants are most frequently hyphenated after the first consonant: *Feb-ruary, his-tory, terres-trial, dis-cipline, prob-lem, ob-scure, coun-try, troub-le, pub-lish, trick-ling, strug-gling, aph-tha, etc*. The past particle -ing is most frequently split from verbs suchas: pull, will, ebb, discuss (*pull-ing, will-ing, ebb-ing, dis-cuss-ing*) but the past particle of travel and rob are hyphenated as *travel-ling* and *rob-bing*.

- Words with long vowels or double vowels at the end of a syllable are hyphenated after the vowel: *na-tion, rea-son, deci-pher abdica-tion, fi-nite, bi-ble, va-grant, etc*.

- Prefixes and suffixes are frequently used to separate syllables: *mis-be-have, de-scend, sing-er, long-er, long-est, announc-ing, accord-ing, as-sur-ance, clear-ance, perform-ance, depend-ent, Roman-ism, Puritan-ism, social-ism, natural-ist, national-ist, but admi-rable, favo-rable, memo-*

rable, du-rable, starva-tion, observa-tion, catholi-cism, criti-cism, etc.

Phonetical hyphenation, according to the pronunciation of syllables, is accepted by nearly all English Dictionaries, such as the huge *Webster's New Twentieth Century Unabridged Dictionary*[10a,b]. The vast majority of words in *Longman Dictionary of Contemporary English*[11] also agrees with Webster's dictionary.

Confusing examples

Some words which are spelled similarly have different hyphen locations. They cannot be hyphenated automatically. Examples are:

record	rec-ord or re-cord
crater	cra-ter or crat-er
elder	eld-er or elder
vice	vice or vi-ce ver-sa
nestling	nes-tling or nest-ling
learned	learned or learn-ed
sleeved	sleeved or sleev-ed
soles	soles or so-les

These words are not hyphenated. If necessary the user can place soft hyphens in those words.

Aesthetics

Some syllables consist of a single vowel at the beginning or the end of a word e.g., *a-ble, a-board, a-bolish, an-y*. These hyphenations do not improve the readability of the text and should be avoided,

Hyphenations such as *passo-ver, une-ven* disagree with aesthetics. These words will be hyphenated as *pass-over* and *un-even*.

2.1.7. The Estonian hyphenation rules (silbitusreeglid)

Estonian like Finnish, belongs to Finno-Ugric languages. The hyphenation rules are comparable to those of the Finnish language. Compared to Finnish, there are more diphthongs in Estonian (*õa, õe, õi, õo, õu*), but word endings do not use vowel harmony: *aberratsiooni, aberratsioonidest, tüüpi, tüüpidest*. Like Finnish compounds complicate hyphenation: *ai-nu-kri-tee-riu-mi-na* (and **not** *ai-nuk-ri-tee-riu-mi-na*), *ai-nu-voi-ma-li-kult*, *aja-graa-fi-ku* (and **not** *ajag-raa-fi-ku*).
(For principles see *the Finnish hyphenation rules*).

2.1.8. The Finnish hyphenation rules (tavujako, tavutussäännöt)

Finnish belongs to Finno-Ugric languages. Finnish is one of the Baltic-Finnic languages. Besides Finnish only Estonian has a standardized form. The nature of the relation between Finnish and Estonian is apparent even to the most casual observer. This applies even to hyphenation. Estonian has one extra vowel **o**, which is hard to pronounce by the Finns.

The Finnish language is characterized by an overwhelming use of compounded words. Word-endings to lemmas modify the word-meaning. Moreover, the principles of vowel harmony determine the choice of alternate endings. Therefore the word length can be considerable. Finnish is characterized by doubling of character to express long consonants and vowels. Words with short and long vowels have different meaning: *tuli* (fire), *tuuli* (wind)[13]. Long consonant groups don't occur. On the other hand, a lot of falling and rising diphthongs exist.

The Finnish hyphenation is based on two criteria:

– a syllable boundary occurs before every sequence of consonant plus vowel and

– a syllable boundary occurs between vowels which do not form a diphthong.

Words are hyphenated as follows:

- The Simplizia or the non-compounded words are hyphenated as:
- for a single consonant between vowels: *a-se-ma* (position), *au-to* (auto), *pu-he-lin* (telephone);
- before the last consonant for a consonant group: *ank-ka* (duck), *kart-ta* (geo. map). Two consonants between vowels are hyphenated between the two consonants: *hoh-taa, hoiken-taa, hoik-kuus*.
- The most important rule is that compounded words (Komposita) are divided into their compound elements: *maan-tie* (high way), *syys-kuu* (September), *idän-puoleinen* (easterly), *kukka-kauppa* (flowers shop), *maailman-aika* (Greenwich Mean Time), *moni-avioinen* (polygamous), *pahaa-aavis-tamaton* (without suspicion), *puu-sepän-tehdas* (carpenter's works), *ros-ka-joukko* (rabble, riff-raff).
- Two distinct vowels can be hyphenated: *pi-an, no-pe-a* (fast), but the diphthongs *ai, ei, oi, ui, äi, öi, au, eu, iu, ou, äy, öy, ie, uo* and *yö* are never di-

vided: *tuom-me, vien-ti, Suo-mi*.

2.1.9. The French hyphenation rules (syllabisation)

Hyphenation rules hardly exist in the French language. Le Bon Usage has been used for the rules in general, L'Orthographie Française and information from the Équipe CNRS-HESO for more details.
Word are hyphenated as follows:

- When a word contains a consonant between two vowels, the hyphen comes before the consonant: cha-peau, cou-teau, cha-ri-té. The dissyllables gn, ch, th and ph are not hyphenated: *mi-gnon, appro-cher, li-tho, pathos, gra-phie*. Open consonant groups are never hyphenated: *sa-ble, ora-cle, prê-tre, pro-pre*. Hyphenation is not possible when the *x* is pronounced as *ks* or *gz*: *taxi, hexa-go-ne* compared to *soi-xan-te*.
  When a word contains two consonants between two vowels, the word is hyphenated between the two consonants: *fer-mer, es-pion, tes-son, aller, stag-nant* and *tex-te*. The dissyllables and the consonant-liquid groups are kept unified: *at-mos-phè-re, a-pos-tro-phe, exam-ple* and *appro-cher*. The double l "intervocal ill" in the words *tailleur, cueillir*, and *habiller* are also kept unified, which should not be mistaken for the ill with double l in the words: *vil-la-ges* and *tran-quil-li-té*.

- When there are three consonants, the hyphen comes after the second consonant: *obs-ti-né, comp-té*. Exceptions are the dissyllables: *mar-cher, mor-phi-ne* and the consonant-liquid group: *ar-bre, ap-plau-dir*.

- When there are four consonants, the hyphen yet again comes after the second one: *ins-truit*.

- Vowels are not separated, e.g., in the words: *es-pion* (semi-vowel), *as-seoir, fée-rie* (silent e), *beau, paon* (dissylables). Exceptions are the prefixes: *ré-tro-ac-tif, an-ti-a-é-rien*, and between o/a in *o-a-sis*. When one of the vowels is stressed, hyphenation is always possible: *a-é-rien, a-è-de, po-è-te*.

Note!

Following has been taken from the letter Communication de Liège from Madame Nina Catach and her team Équipe CNRS-HESO (1990)[14].

- Préfixes. When the prefix is unstressed, hyphenation is normal: *té-lé-*

spec-ta-teur, dé-struc-tu-rer, but it is *des-truc-tion*.

- Apostrophe. When the word contains an apostrophe, there is no hyphen between the apostrophe and the following word: *au-jour-d'hui, tout-à-l'é-gout, pour l'a-voir*.
- Consonant-liquid groups. Hyphens are determined between the semi-vowel and the vowel that follows: *cli-ent, clou-ait, cri-ait, tru-and, clou-a, except for huit, flui-de, truis-me*.
- Semivowels and vowel groups. Hyphenation is not possible for words containing a y (intervocal y, yi) as in the words: *em-ployeur, trayait, es-suyas-se, es-suyions, ab-baye, but ma-yon-nai-se, co-ba-ye*.
- Nasal vowels. Hyphens come before the vowel in words like *en-ivrer* but *é-namourer* (stressed vowel).
- Internal and final silent vowels plus silent *e*: *as-seoir, es-suient, cri-aient, prient, homme* (consonant + silent vowel); however, words like: *ai-guë* and *am-bi-guë* are not hyphenated.

Etymological and phonological rules

Automatic syllabification programmes exist to divide words in either the etymological or the phonological way. The two programmes follow general rules: *Manu-scrit* and *manus-crit*, *atmo-sphère* and *atmos-phère*.

Hyphenation in publishing matters

Newspapers accept different rules as to avoid blanks in small text boxes. Sometimes you will find isolated characters at the beginning (*a-é-rien*) and more frequently refusals of silent vowels followed by a consonant-liquid group at the end of a word (*pe-lot-te, noi-râ-tre*). These words are nearly always hyphenated in books, especially in plural (*pe-lot-tes*). The *TALO-hyphenator offers the possibility to avoid uneasthetical separations.

For long words, there is another possibility, e.g., *en-so-leillaient* and *es-suyai-ent*. One could place them in the list of private hyphenations: *en-so-leil-leient* and *es-su-yaient*.

2.1.10. The Frisian hyphenation rules (staveringshifker/-ôfbrekking)

Frisian is spoken by a population of 500,000 Frisians in the northern part of the Netherlands, and by a number of Frisians living in the *diaspora*, outside Fryslân. The Frisian language has been influenced by the neighbouring languages, but also by English and the Scandinavian languages, because the trade highway was the Northsea. Hyphenation rules are comparable to the rules for Dutch. However, spelling is not. There is a distinct use of the consonant j, e.g. in *posysje, mâltjirgje, lju* (in English *men*, in Dutch *lui*). Diphthongs differ: *ii* (in tiid), *ea* (in marsjearje), *oa* (in meikoarten (soon)). The closed *i* is written as *y*: *yn-dus-try, yn-keaps-prijs, yn-ke-ping*.

However, the mayor problem of hyphenation is caused by the compounding nature of Frisian: *yn-hel-fer-bod, kopke-tommelje, latte-stek, limoer-hals-kanker, long-ûnt-stekking, etc*.

2.1.11. The Galician hyphenation rules (separazón silábica).

Galician (língua Galega) is spoken in the north-western part of Spain. It is related to Portuguese and Spanish (Castillan). The Galicians feel a strong historical relation with Portuguese, however, the written language is closely related to Spanish, e.g., the ll grapheme is used instead of the Portuguese grapheme lh. Therefore -rr, -ll, and -n are (Água-mariña) comparable to Spanish. Differences in hyphenation are mainly caused by differences in orthography. Typical for Galician is the x alcobexar (-jar), aldraxar (-jar), ambaxes (-ges) the suffix -zón in alegorizazón, alocuzón, animazón, etc. Therefore, the differences between the Spanish and Galician hyphenator are small (see the section: the Spanish hyphenation rules).

2.1.12. The German hyphenation rules (Silbentrennung)

The hyphenation rules of the German language are quite similar to the rules of the Dutch language; they both follow the consonant hyphenation and the vowel hyphenation.

The structure of compound words is comparable to the one in the Dutch, the Danish, the Norwegian or the Swedish languages. Compound elements are not separated from each other by spaces or hyphens, but are unified into one word.

Major elements of the hyphenation rules are listed below. An extensive description of the hyphenation rules can be found in the Duden Rechtschreibung Band

1[18,19a,b,c,d,e].

In general, hyphenation will take place in the following situations:

- The most important rule is that words are divided into their compound elements: *Erholungs-urlaub, Kriegs-erlebnis, Ver-giß-mein-nicht, weg-zu-gehen, sie-ben-hun-dert-tau-send-neun-hundert-acht-und-zwan-zig, hin-ter-ein-an-der*. In case of prefixes, hyphens come after the prefix, e.g., be-span-nen, be-spei-en, ent-hül-len, ent-lang-lau-fen, ü-ber-seh-bar, etc.
- Between two vowels (vowel-groups) immediately following each other: *Steu-er, Stati-on, be-anspruchen, be-aufsichtigen*.
- Before one consonant: *nö-ti-gen, sto-ßen etc*. The ß-character is treated as a single consonant.
- Between two consonants: *Blin-den, fol-gen, nied-lich*.
- When three or more consonants follow each other, as many consonants as possible are transferred to the preceding syllable: *Karp-fen, Bast-ler, Deut-schen, damp-fen*.
- The consonant groups *ph*, *sh*, *th* or *sch* are not divided, except in case of compound words and the bygone ban to hyphenate the *st*: *Gra-phik, Ca-shew-nuss, ka-tho-lisch, Deut-schen, Sys-tem* (not *Sy-stem*).

Pre-1996 Special hyphenations

Before 1996 the character group ck enclosed by vowels was hyphenated as *k-k*, e.g., *dicke* → *dik-ke*. Special hyphenation rules applied to compound words, e.g., *Stilleben* became *Still-leben*, *Brennessel* became *Brenn-nessel* and *Schiffahrt* became *Schiff-fahrt*. Nowaday these orthographic forms are considered as a mistake. These words have to be to be written as Stillleben, Brennnessel, Schifffahrt or an hyphen and a double upper case ought to be used (*Still-Leben, Brenn-Nessel, Schiff-Fahrt*.

Differences between the Swiss German and German language

Most German hyphenation rules equally apply to the Swiss German variant of the German language[20]. However, there are some differences. The ß-character is often replaced by the double s. Words using this replacements are hyphenated between the double s, e.g., *Stras-se, grüs-se* (and not *Stra-sse, grü-sse* !!). In case a double s is followed by a third s, all three are written down, e.g., *Schiessstand, Fusssohle*. The hyphenator support all cases in which ß is replaced by ss, e.g., *Guß-arbeit* → *Guss-arbeit* and **not** *Gus-sarbeit*; *groß-artig* → *gross-artig* and **not** *gros-sartig*!

Differences as du liesest (liest), reisest (reist), hassest (haßt), heißest (heißt), sitzest (sitzt), hexest,(hext), wünschest (wünscht) etc do not disturb the hyphenation.

German reformed

Due to the changes in spelling, some former hyphenation principles can no longer be used[19a,b,c,d,e]. The language module "German reformed" of the hyphenator has been optimized according to all spelling and hyphenation changes. A second style — German reformed conservative — is also available.

The most important changes are:

- The ligature *st* disappears. These consonants should be hyphenated as: *kas-ten, Alabas-ter, illus-ter etc.*

- Instead of hyphenation of *ck*, these dissyllables are kept together, e.g., *Zu-cker, ba-cken, Zwi-ckau*.

- The principle of consonant doubling disappears. The new spelling of *Schiffahrt* is *Schifffahrt*, however, the Hyphenator does detect old spelling variants.

- The eindeutschende principle of hyphenation is introduced. This rule affects compound words which are now treated as individual elements: *an-ei-nan-der, über-ei-nan-der, etc.*

- Greek and Roman prefixes are now hyphenated in a different manner: *Ob-edi-enza* (1980), *O-be-di-enza* (1996) becomes *Ob-e-di-enz* (2006), *Sy-nod* (1980), *Sy-nod* (1996) becomes *Sy-n-od* (2006) and *sy-no-nym* (1980), *sy-no-nym* (1996) becomes *sy-n-o-nym* (2006), etc.

After 2006 different hyphenation principles have become allowable. However, it is recommended to use a consistent style of hyphenation.

2.1.13. The Icelandic hyphenation rules (orðskipihlutinn)

Icelandic today is much as it was when Iceland was first colonized, mainly from Norway, in and after 874. A lot of early history has been compiled in the great poetic and prose Eddas there exists a great number of sagas. Icelandic is one of the oldest Germanic languages of which records exist. These old roots still are understandable to speakers of present-day Icelandic. Icelandic has a rich heritage of many of the old features of the Germanic languages. There is a division in gender, masculine, feminine and neuter and each of these genders has different cases. There are four cases, nominative, accusative, dative and genitive. These cases are endings attached to the nouns. One of the distinguishing features of the Scandinavian languages is the postpositive article, attached to the end of the noun. Again this article has four case-endings[21]. It is this structure that makes hyphenation complex. Moreover, Icelandic largely is blessed with an infinite number of compounded words.
Words are hyphenated as follows:

The most important rule is that words are divided into their compound elements, e.g., *alþingis-maður* (member of the Icelandic legislative assembly), *efnis-hygg-ja* (materialism), *kvæða-bók* (volume of poems), *sím-skeyti* (telegram), *sjálf-boða-vinna* (voluntary work), etc.

- Icelandic hyphenation is strictly morphological, that is, the attached article and the cases are separated from the lemma, e.g., *hest-ur* (horse), *hest-ur-inn, hest-inn, hest-in-um*, but not *hesti-num*.
- Hyphenation of a single vowel at the end the word or lemma is not allowed, *alda-röð* and not *ald-a-röð* (for centuries)
- The diphthongs ei, ey, au are never hyphenated, e.g., *Hauk-ur, laun, laus*.

  Words are hyphenated after a prefix: *af-bragð, eftir-för, ein-angra, for-liður, frá-saga, inn(an)-, ó-, sam-, -út, etc*.
- Words are hyphenated before a suffix, e.g., *-laus, -lega, -legur, etc*. These suffixes behave like compound words.

Icelanders refuse to import Anglicisms and other foreign words. New concepts are put into neologisms, words which sometimes have their root in history. The Icelandic word for telephone is *sími*. For centuries the meaning was thread, wire not exactly like the wire of a telephone line, but figurative.

2.1.14. The Italian hyphenation rules (divisione in sillabe)

The Italian language only has only a few compounds, e.g., *russogiapponese* and *foxterrier* which should be hyphenated as russo-giapponese and *fox-terrier*. On the other hand, the structure of Italian is flavoured with other forms, especially verbs take many forms to express the different times.

Hyphenation is guided by consonant principles[15,16]:

- When a word contains one consonant between two vowels, the hyphen is placed before the consonant, e.g., *a-mo-re, ca-pi-to-lo, de-ri-de-re, piu-mi-no*. Note: When hyphenating e.g., the word *amore*, the single *a* at the end of the line will be omitted.
- When a word contains two or more consonants, the consonants remain with the vowel that follows *ro-vi-sta-re, na-sce-re, e-sclu-de-re, no-stra-le, ve-spro*.

For certain ensembles of consonants the rules are different:

- When the consonant group begins with an l, m, n, or r, the first letter remains with the preceding syllable, e.g., *pro-pul-sio-ne, ram-po-gna, lon-ga-ni-me, pian-ta-re, scor-ta-re*.
- When a consonant group contains a double identical consonant, the first letter remains with the preceding syllable, e.g., *sil-la-ba, pet-ti-ros-so, pic-chio*.
- When the group consists the consonants gm, cm, tm, bn, cn, bs, fg, cq the hyphen is placed between the consonants: *seg-men-to, ac-me, at-mo-sfe-ri-ca, sub-nor-ma-le, a-rac-ni-de, ab-so-le-to, Af-ga-ni-stan, ac-qua*.
- Hyphenation between vowels depends on the pronunciation:
- Diphthongs and triphthongs are not separated, e.g., *cie-lo, pro-fes-sio-ni-sta, miei*. b) Some group of vowels are separately articulated, e.g., *im-pa-u-ri-to, ste-re-o-sco-pio, ne-on, na-ti-o* (tì is accentuated) compared to *ce-no-bio* (no is accentuated and io is a diphthong). In case of prefixes, two distinct vowels exits: *ri-a-ni-ma-re, ri-a-per-tu-ra, an-ti-a-è-re-o*.

Dividing prefixes

There are two rules for the prefixes. They can be considered either as separate syllables, e.g., *dis-interesse* and *in-esatto* or not. In the latter case prefixes should be hyphenated as *di-sin-te-res-se, i-ne-sa-to*.

2.1.15. The Norwegian hyphenation rules (deling av ord)

Two official forms of the North Germanic languages of the Norwegians exist. One ("Bokmål") is derived from Danish and the other ("New Norwegian or Nynorsk") is created c. 1850 from Norwegian dialects. Bokmål is used most frequently in forms of printing. Hyphenation rules of Bokmål and Nynorsk are relatively straight forward, but hyphenation of compound words frequently conflicts with the basic hyphenation rules, and therefore computerprograms without linguistic knowledge fail[22]. Words are hyphenated as follows[23]:

- The first rule is that at least one consonant is transferred to the new line: *gå-te, groms-ke, blomst-rer*. However all endings on *-sjon* are not split but hyphenated as: *sta-sjon, forma-sjon*. Words with *kj* are hyphenated as: *bik-kje, kryk-kje*.

- Compound words are divided into their elements: (Stavings-grense og ord-grense fell saman ved einstavnings-ord) *bleik-sott, lese-bok, gards-drift, tids-alder, heste-drift, bok-stav-tegn, bak-meis, barne-mål, små-gut, stor-etar, stygge-ty, tids-alder*.

  Foreign words are divided into elements: *atmo-sfære, pro-blem, inter-esse*, but it is also possible to apply the rule 1: *dif-tong, prob-lem*. Pronunciation of syllables might bias hyphenation.

- Words with prefixes and suffixes are divided into their elements: *(føre- eller etterstavingar) for-stove, for-mann, mis-tak, mis-nøgd, an-stand, be-stikke, mis-unne, tro-skap, dikt-er, ideal-isme, drøft-ing, rens-else, sølv-aktig. ung-dom, spå-dom, kjær-leik, stor-leik, lyd-nad, fe-nad, spar-semd, tol-semd, bu-skap, doven-skap*. However, rule 1 is also allowed. When the syllable boundary is not felt in speech it is preferable to hyphenate according to rule 1, e.g., *dik-ter, ren-sel-se, san-ge-rin-ne*. Note: do not split single vowel at the end of a word: not *gresk-e*, but *gres-ke*, not *geit-a* but *gei-ta*.

Prefixes are separated from the words, e.g., *ab-, ad-, al-, all-, dess-, efter-, er-, etter-, for-, fore-, fra-, fram-, frem-, ge-, gjen-, opp-, over-, på-, så-, sammen-, til-*

, u-, under-, unn-, ut-, van-, ved-, veder-, vel-, vid-, vidt-, etc, e.g., *alle-helgens-dag, dess-verre, hver-andre, hvor-vidt, med-lidende, så-fremt, opp-gjør, be-drag. be-arbeide, er-fare, for-kjærlighet, for-akt, u-blid, u-oppmerksom, u-hyre, u-drikkelig, u-trolig, i-aktta, i-herdig, i-mellom, i-scenesettelse, i-øynefallende*. Latin prefixes are also separated if their origin is clear: *ab- (ab-ort), ad- (ad-optere), aero-, ag- (ag-gregat), ak- (ak-kusativ), an- (a-) (an-arki, an-neks, a-teist), anti- (anti-pode, anti-krist), apo- (apo-kryf), arki- (arki-tekt), ar- (ar-rest), as- (as-sessor), pro-, re-, sub- super-, syn-, sym-, tele-, trans-, ultra-, etc*.

One should hyphenate before the compound parts as *-graf, -grafi, -gram, -log, -logi, -krat, -man* as should be done the suffixes -aktig, -ferdig, -verdig, -haftig

Geographical suffixes are kept together, e.g., -bu (residence), -berg (mountain), -bygd (hamlet), -dal (valley), -fjell (mountain, rock), -fjord, -foss (waterfall), -haug (hill), -heim (residence), -mark (field), -nes (cape), -rud (red), -skog (forest), -sund (strait), -vik (bay), -ø, -øy (insel).

The Norwegian hyphenator uses the consonant style of hyphenation for non-compound words or elements. Prefixes are separated, but suffixes which are not felt in speech, are hyphenated according to rule 1. These linguistic features of the Norwegian *TALŌ-hyphenator can be applied to Bokmål and to Nynorsk.

Special hyphenations

Like in Swedish and German, compound words with three equal consonants on the compound boundary, lose one of them when hyphenated. However, it is also allowed to insert a hyphen to solve the unclear word boundary of the three consonants. So two spelling forms exist: *trafikk-kultur* or *trafikkultur*. However, the last one needs to be hyphenated as *tra-fikk-kul-tur*. The addition of one consonant is supported by the *TALO hyphenator.

2.1.16. The Portuguese hyphenation rules (regras de hifenizaçao, separaçao sílabas)

Hyphenation for the Portuguese language strictly follows phonological flow of speech, i.e., the pronunciation of the syllables is compelling for hyphenation. Differences between the Iberian and Brazilian Portuguese are only caused by differences in orthography[24].
Words are hyphenated as follows[25]:

- When a word contains one consonant between two vowels, the hyphen is placed before the consonant, e.g., mne-mô-ni-ca, pneu-má-ti-co. Some consonants, the digraphs ch, lh and nh are never separated. They are considered as a single consonant and are hyphenated like *bi-chi-nho, fi-lhi-nho, nhe-nhe-nhém*. An occlusive or stopped consonant (p, b, t, d, hard c and g) plus a liquid consonant belongs to the prefix: *a-blu-çao, a-bra-sar but sub-lin-gual*.

- Two consonants are divided *ab-di-car, nup-ci-al, des-li-gar*.
- All double (two) consonants are divided, *pror-ro-gar, as-sen-tir*.
- The letters s and x remain with the prefix when followed by a consonant *bis-ne-to, ex-tra-ir* but when they are followed by a vowel, they become the first letter of the second syllable: *bi-sa-vô, e-xér-ci-to*.
- Vowels always are separated when there is a hiatus: *fri-ís-si-mo, du-e-lo*.
- Diphthongs and the combinations qu and gu always remain together: *a-ni-mais, e-qui-va-ler, am-bí-guo*.
- Prefixes and suffixes are hyphenated phonologically: *de-sen-vol-ver, su-pe-re-go* and not *des-en-vol-ver, su-per-ego*.

2.1.17. The Spanish hyphenation rules (separación silábica).

Like the other Roman languages, Spanish has adapted phonological syllabification which closely follows the flow of speech. Prefixes and suffixes are frequently used, whereas compounds seldom occur.
Words are hyphenated as follows[26]

- For one consonant between vowels the hyphen is inserted before the single consonant (phonetically ch, rr, and ll are considered as one consonant), e.g., *ma-no, ca-sa, de-lei-ta, ca-ba-llo, fe-rro-ca-rri-le-ro*.
- When a word contains two consonants between vowels, the hyphen comes in between, *bar-ba, tor-ta, con-ser-je, ar-gen-to, plan-cha*.
- When a word contains three consonants, there are two possibilities:
  - hyphen after the first letter: *hom-bre, es-cri-bo*;
  - when the middle one is an s, the hyphen is placed after the second letter: *trans-pa-ren-te, obs-ti-na-to, pers-pi-ca-cia, trans-for-ma-ción*.
- When a word contains four consonants, the hyphen is placed after the second consonant: cons-truc-tor, ins-truc-ción.

Note: When l and r are preceded by b, c, f, g or p (d and t before r) the syllable begins before these consonants *pa-la-bra, ma-dre, es-cla-vo, pos-tre*.

Prefixes and compositions are divided like *des-es-ta-bi-li-sar, en-tre-a-brir, e-qui-án-gu-lo*. They are hyphenated according to etymological rules.

- Syllables containing two or three vowels, like diphthongs (combination of strong and weak vowel) and triphthongs (combination of a strong vowel between two weak ones) are not hyphenated *rei-na, neu-tro, bai-le, cau-sa, fies-ta, de-ges-tión, o-fi-cial, triun-fo, tam-bién, a-ve-ri-güéis, pre-mi-eis, a-guais, a-griéis, a-me-ri-ciáis*.

- Stressed vowels are separated (a, e, o), *ca-er, cal-ma-os, em-ple-ar, ca-os*. When separation leaves a syllable with one vowel only, this one vowel is grouped together with the other syllable *rea-li-dad, de-ceo, pae-lla, poe-sía, a-le-gría, raí-ces*. Note the difference between *a-brí-ais* and *pre-miáis*.

- At the end of a sentence, hyphenation usually does not occur when:
  – one initial letter is left alone (*a-mable*)
  – one final widow-syllable should be suppressed (*decí-a*).

Conjugations

Long groups of vowels occur in conjugated verbs. Unaccentuated vowel groups which sounds like a single vowel are not divided, e.g., *raéis, reío, reía* but accentuated ones are, e.g., *reí-ais* and *roí-ais*.

2.1.18. The Swedish hyphenation rules (avstavning)

There are two principles according to which the Swedish language is hyphenated:

- mekaniska principen: one consonant is transferred to the next syllable, e.g., *sto-lar, sit-ter, stu-dera*; the remaining consonants remain with the current syllable.

- morfologiska principen: words are divided according to their natural border, also known as ordbildningsprincipen, e.g., *stol-ar, sitt-er, etc*.

Newspapers prefer the mekaniska principen. This hyphenation principle was also used in the Swedish *TALO-hyphenator. Hyphenation agrees with the Svenskt avstavningslexikon (Richard Klingspor, 1989)[27], but it is extended to

the hyphenation of compound words. For some cases preferences of the Svenska Dagbladet were used. The consonants ck are hyphenated as c-k (default case), word endings such as ...ring, are hyphenated as ...-ring, and vowel-vowel hyphenation is suppressed except for compound boundaries (*kom-mis-sa-ria-tet* versus *eko-no-mi-ekot*).
Words are hyphenated as follows[27,28]

The S-Laut (German sharp s, or the sje-ljud in Swedish) is not divided: *ma-skin, du-scha, cre-scendo*, but ssj and ssi are hyphenated as *rys-sja, mis-sion*, however, not in compound words like *löss-jord*.

A hyphen nearly always comes after the dissyllable ck, ng and after the x: *äng-ar, nyck-eln, yx-or, reflex-iv* but *tan-gent, ran-gera*. The suffix -ing if preceded by a vowel and a single r is hyphenated as *korruger-ing* (according Avs.Leks.), -ing-en, -ing-arna, etc.

In foreign words with a non-accentuated first syllable as many consonants as possible are transferred to the next syllable: *cy-press, ka-strull, elek-tron, pro-blem, por-trätt, etc*.

However, compound words are always split into their elements: *vit-kalkad, hem-bakad, kyrk-klock, huvud-rätt, etc*. The same occurs with prefixes and suffixes. *vän-skap, för-söka, etc*.

Prefixes are: *an-, av-, be-, bi-, bort-, er-, för-, före-, god-, hem-, här-, in-, miss-, om-, sam-, um-, und-, ur-, ut-, van, veder-, å-, e.g., er-bjuder, er-bjöd, in-be-tala, hem-söka, av-lämna, bort-skämma, av-göra etc*. The word is always hyphenated after the prefix.

A special group of words have double consonants but originally had triple consonants, e.g., *tillåta* is hyphenated as *till-låta*. About 500 cases have been found. In some cases a hyphen is part of the spelling e.g., to differentiate *glas-skål* (a bowl of glas) from *glass-skål* (a bowl containing ice cream). Over 500 of these words have been found. Given 4 declension forms of the Swedish language over 2000 word forms of this type occur. Moreover they will also occur in newly created compounded words.

Dividing compound words

The consonant rules of Swedish hyphenation are simple for non-compound words and can be described with a limited number of rules. However, such a description is not possible when compound words are also considered. Compound words that deviate from the consonant rules are difficult to hyphenate according to most algorithms or dictionary based systems. Difficulties that are correctly recognized by the *TALO-hyphenator are: *fred-uppgörelse, fred-invit, frimurar-orden, frihets-ideal, frihets-straff, frisklufts-intag, från-skjud, full-ändad, etc*. These difficulties are comparable to the ones in the Dutch and German language.

Similar to Dutch and German, the prefix and suffix lists are inadequate for compound words. Even small differences between words influence the hyphenation: *kont-rar, kon-trasten, elekt-riska, elek-troden*.

Words beginning with *an* do not necessarily begin perse with the prefix an, e.g., *a-norna, a-nonyma, a-noden*, but *an-alfabet, an-arkin, an-estesin* are hyphenated after the prefix *an*.

To solve the above difficulties, all available knowledge is gathered in a pattern recognition system containing the most condensed linguistic vision on Swedish hyphenation.

2.1.19. The hyphenation rules of the Greek language

Modern Greek consists of 24 letters expanded with a number of diacritics: the dieresis and the tonus, the acute accent. The traditional diacritic system, the polytonic system, consists of an array of diacritical marks written above and below vowels. The function of the dieresis is to indicate that the two vowels are two separate sounds and therefore could be divided.

- a word break may be placed before a single consonat between vowels (έ-χω)
- a word break may be placed between two vowels only if they represent two separate vowel sounds (να-ός) while the word ναύτης only can be divided as ναύ-της and not να-ύτης.
- a word break may be placed before the first of two consonants between vowels if these consonant can occur together (θάρ-ρος), otherwise the break is placed between the two consonants. This also applies to three or more consonants between vowels.
- The following consonants are examples of combinations that are never divided:
  βγ, βδ, βλ, βρ
  γγ, γδ, γχ, γλ, γν, γρ

δσ
θλ, θν, θρ
χλ, χν, χρ, χτ
μν, μπ
ντ
πλ, πν, πσ, πρ,
σβ, σγ, σφ, σχ, σλ, σμ, σν, σπ, στ, σφ, σχ
τξ, τμ, τρ, τσ
φθ, φχ, φλ, φρ, φτ
χλ, χν, χρ, χτ

The impact of Greek upon the vocabulary of all other languages has been enormous. Many prefixes as poly-, micro-, anti-, auto-, tele, geo-, psycho, or suffixes as -scope, -phone, phobia are pure Greek. These pre- and suffixes determine hyphenation in many languages.

2.1.20. The languages of Eastern Europe and the Balkan

2.1.20.1. Hyphenation of the Slavic languages

Polish

The Polish language is characterized by consonants that are nearly unpronounceable for non-Polish people. These consonants usually appearing in a series of consonants are joined together: zgw in zgwałcić (rape), wzbr in wzbroniony (forbidden) chrz in chrzest (clatter). Special consonants are *ź, ż, ś, ć, ń* and the *ł*. Several consonants belong together and represent a single undividable sound. The *ck* is pronounced as *tsk*, the *szcz* as in fre*sh* or *ch*eddar, the *ś,ć* as the Finnish *ch*eese, *dż* as in *j*am. The consonants *dz* in *pięć-dzie-siąt* and *są-siedz-two*, and the *cz in sfe-rycz-ny* are kept unbroken.

- Apart from these consonants a single consonant between vowels is hyphenated before the consonant: *se-pa-tacja, se-cesja, se-re-na-da, sfo-ra*

- If two consonants between vowels occur the consonants are hyphenated as *sepatac-ja, seces-ja, ser-wat-ka, ser-nik*.

- In Polish certain prefixes are used to add a shade of meaning to the original verb. *do* (to, towards, *w* (in), *nad* (above, near), *przez* (across): *do-brać, do-budować, nad-robić* (and not na-drobić), *nad-skakiwać* (and not nads-kakiwać).

Compounds as in German or Swedish hardly ever occur in Polish, but on the other hand Polish is highly inflected; there are 7 cases in total: nominative, vocative, accusative, genitive, locative, instrumental and dative.

Czech
The Czech language is hyphenated etymologically. A prefix is separated from the stem, and since compound boundaries determine semantic content of words these boundaries are used for hyphenation. In some cases prefixes are no longer felt, so the prefix “roz” is disregarded in the verb ro-zu-mět (to understand).

- a hyphen is placed before a single consonant between vowels: *bá-seň* (poem), *do-nést* (to bring), *do-po-sud* (up to now). The “ch” is considerate as a single consonant.
- 2 consonants between vowels usually are separated: *bás-ník* (poet), *dob-rý* (good), *far-mář* (farmer). For some words the vowels would seem to be missing (*filtr*) but for these cases the r and l serve as vowels: *fil-tr* (filter), *Br-no* (the City Brno), *ře-kl* (conjug. from *říci*). but *za-blounit, za-brat, za-hrada*, keep the "bl, br, hr" together, but *prob-lem* not.
- for more than one consonant hyphenation depends on the consonants that belong together *holič-ství* (hairdresser) *hospodář-ství* (economy, farm)

The Czech language has 7 cases like the Polish language.

Slovak
The Slovak language is closely related to the Czech language and most principles found in Czech apply to Slovak. There are a few spelling differences that influence hyphenation. In Czech *biologie* is pronounced as *bi-yo-lo-gi-ye*, in Slovak the final ie is a single sound as in *blahopria-nie* (pronounced as ....-nye).

- In Slovak a word is divided after a vowel *ry-ba, po-chva-la*, but in writing syllables can be split between consonant clusters: *od-po-vedz-te*.
- Consonant clusters are split following a nasal consonant (*n, ň, m*) or *j*: *slo-ven-ský*
- Diphtongs (combinations of two vowels) are kept together: *pria-teľ, pia-ty, mier-ny, mlie-ko ra-dium, star-šiu.*

In Slovak the major part of the pre- and suffixes behave identical as in Czech. *Ne-* is a prefix in the verb *ne-boj sa*! (don’t be afraid), nearly all verbs can be negated with "ne-", but not all words starting with "ne" are treated equally: it is *nervóz-ny* (nervous).

Slovene

The Slovene language belongs to the South Slavic languages but differs from Czech, Slovak or Croatian. The Slovene alphabet has three consonants with a superscript diacritic: *č, š* and *ž*. The standard language has short or long vowels but in writing no diacritics are used to make these vowels distinct.

- In Slovene the *dž* (džusa), *lj* (Ljubljana) and *nj* (njegov) represent single phonemes which are never divided, they are considered as single consonants like the *ch* in English.

- As in the other Slavonic languages prefixes are frequently used to modify meaning. Hyphenation is allowed on prefix boundaries: *nad-človeški* (nad- super), *pra-zgodovinski* (pra- pre), *pod-žemeljski* (pod- sub), etc For verbs a prefixed perfective exists: *črtati* → *na-črtati, pod-črtati*; *dati* → *do-dati, iz-dati*; *slediti* → *iz-slediti*.

In Slovene words are formed by affixation and various types of compound, as well as deaffixation, where words are truncated. There are native and loan suffixes, and there are consonantal alternations (*drevesce* (small tree) → *dreveščček* (Chrismas tree)). Suffixation and composition mechanisms are diverse and call for a native Slovene hyphenator.

Croatian

The Croatian language is the counter part of Serbian. Together the two varieties are named Serbo-Croatian. In Zagreb and Belgrade the language is known as *srp-sko-hr-vat-ski* or *hr-vat-skp-srp-ski*. For Croatian the Latin alphabet includes the letters *č, ć, đ, š* and *ž*. The Croatian vowels do not have diacritics.

- Croatian does not separate the double consonants *dž, lj* and *nj*. In the Cyrillic script the *dž, lj* and *nj* are represented by a single character.

- The spelling of Serbo-Croatian is phonetic and almost every word is written exactly as it is pronounced. Double vowels as au and ou are pronounced separately and therefore they are hyphenated: *po-u-ka, pa-uk, tri-na-est*, except for some English words *kauč* (couch). The *r* in *hr-vat-ski* serves as a vowel and guides hyphenation.

Serbian

The Serbian language is the Cyrillic counter part of common Serbo-Croatian. Words in the Belgrade variety less frequently use the letter *j* in words as shown in the Latin alphabet *r(ij)eka* (river), *p(j)esma* (song), *dot(j)icati* (to reach), but hyphenation is quite similar for the Cyrillic alphabet. The Croatian *lj, nj* and *dž* are represented by a single character.

Russian
The Russian language is written in the Cyrillic alphabet, but for the purpose of clearness Russian is transcribed here to the Western script.

- The following consonants may not be divided *bl, pl, gl, kl, fl, vl*, and the '*r*' cases: *ru-brika* (ру-брика), *puteshe-stvennik* (путеше-стенник), plus *dv, dr, tv, tr, sk, skv, skr, st, stv, str, zhd, ml*.

- Diphthongs will not be broken too: *oy, ay, ey*.

- Prefixes as *bes, pred, na, pro, za* are separated: *bes-poryadok, pred-lozhit, na-zhat, pro-vesti, za-dat*. (бес.пориадок, пред.лоэхит, на.жат, про.вести, эа.дат).

These principles apply both to Cyrillic and the transliteration form of Russian. Generally these principles apply to Ukrainian and Byelorussian too.

2.1.20.2. Hyphenation of the Hungarian language

The Hungarian language has been isolated from its family members of the Finno-Ugric branch. The Hungarian alphabet is extended with the vowels á, é, ó, ö, ő, ú, ü, and ű. Several consonant clusters represent a single consonant sound cs, dz, dzs, gy, ly, ny, sz, ty, and zs.
Like Finnish the Hungarian language is dominated by vowel harmony and like the other members of the Finno-Ugric languages prepositions are added to the end of words. "To Amsterdam" becomes "Amszterdam-ban"

- In Hungarian a lot of compounded words exists: the stem *ipar* (handiwork) is combined with *engedély, hatóság, művész* to form the words *ipar-engedély* (patent), *ipar-hatóság, ipar-művész* etc.

- If consonant clusters are doubled, spelling is affected too. The consonant doubling for *ly, ny, gy, sz* is written as *lly, nny, ggy, ssz*. However, when these consonants need to be hyphenated they are written as *ly-ly, ny-ny, gy-gy, sz-sz*. Since words with these consonants can occur in nearly any compound, special hyphenation should receive a great deal of attention: *asszony* (woman) *államtitkár-asszony → államtitkár-asz-szony, anyám-asszony → anyám-asz-szony, bába-asszony → bába-asz-szony, cigány-asszony → cigány-asz-szony.*

2.1.20.3. Hyphenation of the Baltic languages

Lithuanian

The Lithuanian language is called a conservative language in the sense of keeping many of the Proto-Indo-European characteristics. Lithuanian is highly inflected like Russian, but it is considered as a separate branch of languages with Latvian.
Special characters are *š, ž, č, ė*, and the nasal letters *ą, ę, į, ų*, and the long vowel *ū*.

- The diphthongs *ai, ui, au, ie* and *uo* are never hyphenated: *mui-las* (soap), *duo-ti* (to give).
  Some diphthongs are borrowed from foreign languages, e.g., the *eu* in Europa, the *oi* in boikotas (boycott) and the *ou* in klounas (clown)

- A word is hyphenated after a prefix: *iš-* iš-au-gant, *kontr-* kontr-ar-gu-men-tų, *ne-* ne-a-be-jo-ja, *ne-* plus *iš-* ne-iš-aiš-kin-ti.

- The digraph *dž* is kept together: *laik-ro-džiai*.

Latvian

Special characters in Latvian are the vowels *ā, ē, ī, ū* and the consonants *č, ĝ, ķ, ļ, ņ, š, ž*.

- In hyphenation the Latvian diphthongs are kept together *ei* (teikt), *eu* (sev), *ai* (laiks), *au* (tau-ta), *oi* (boi-kots), *ui* (pui-ka), *iu* (pliuk-šķināt), *ie* (lie-pa), *uo* is written 'o' (o-la).

- Incidental compounds are divided on the compound boundary: *Čecho-slovakijâ, četr-vietîgas* (four-seated) *div-desmit-četrus* (twenty-four)

2.1.20.4. The language of the Dacians

Romanian

An earlier form of the Romanian language was spoken by the ancient tribe of the Dacians. Influenced by the French orthography Romanian was standardized in the early 20th century. Romanian includes a few additional characters *â, ă, î, ş, ţ*. Diphthongs are: *ai, au, ea, ei, eu, ia, ee, ii* (a longer i), *io , iu, oa, oi, ou*. The origin of Romanian determines hyphenation. Hyphenation principles are close to those of the French language too, but they are not identical.

- A few characteristics of Romanian should be mentioned. In conjugation of verbs a circumflex can become a breve (â → ă) as in the verb *a mân-*

ca (to eat, French: manger): *eu mă-nânc* (I eat), *voi mân-ca-ţi* (you eat, pl.).

- A single consonant between vowels goes to the next syllable, two consonants between vowels are hyphenated in between: *ca-să* (house), *gos-po-di-nă* (housewive), *me-na-je-ră* (housekeeper).
- the â and î are similar, but the â is only used within a word and the î as the first or last character *întâi* (first), the *î* however can occur after a prefix *re-în-ca-dră-ri* (re-framing), *re-îm-păr-ţi-re, re-în-fiin-ţa-re, pre-în-tâm-pi-na-te*.
- Romanian is phonetic, and no double consonants are used, except for a few exceptions; these double consonants are hyphenated in between (cc) *ac-celerat, ac-cent, ac-cident*, (nn) *în-nourat* (cloudy), *în-noptat* (to get darker) (în- is a preposition).
- a few double vowels are not broken, the oo in *al-cool*, the ii in *fi-ii* (the sons) stem *fi*. *co-pi-ii* (the kids) stem *copi*.

2.1.20.5. On the border of Europe: Turkish, Azerbaijanian and Kazakh

The Turkish language has some additional vowels: ö and ü and a dotless i. The additional consonants are ş, ç and ĝ. Azerbaijanian adds a mirrored *e* to their Latin alphabet. This mirrored e is written as *ä* in the Latin Kazakh script. These languages have a richly agglutinating morphology and an elaborate case system.

- In general a consonant between vowels is transferred to the next syllable (*ko-nuş, yü-rü*), two consonants between vowels are hyphenated in between (*gün-ler, an-lat*).

Due to the assimilation this principle also applies to the agglutinations. Particles added to words to change meaning follow the above mechanism:

- (Negation) *kalk-ma* (don't stand up), *otur-ma* (don't sit), (infinitiv -mek, -mak) *gel-mek, bak-mak*,
- The letter y can be used as a buffer sound if a verb is ending with an e or a, (the future -acek, -ecak) *anla-yacek, iste-yecek*.
- (Future with negation -me, -ma) *gel-me-yecek, dur-ma-yacak*.

- (Magnitude, having a size of, lik, lük, lık, luk) egemen-lik, kişi-lık, gün-lük

2.1.21. Hyphenation of the Near Eastern languages

2.1.21.1. Hebrew

It is not possible to implement hyphenation algorithms as used in the Roman languages. The commonly used Hebrew type omits most vowels, and division into syllables becomes partly ambiguous and one ends up with double meanings of words each having there own division in syllables. Still newspapers are forced to hyphenate where possible. A few rules are acceptable: writing in narrow columns of a newspaper requires hyphenation and hyphenation after a י and ו is allowed, but יי and וו are not separated. However most words as רכבת can be pronounced as *raχavt* meaning *you have ridden* or as *rakevet* meaning *train*. In case of the meaning *you have ridden* רכ-בת would be an erroneous hyphenation because the graphem group *vt* is not a syllable.

So the following written remark of a big company is wrong:

> *Hyphenation in Hebrew is simpler than in English. Since almost all letters represent a full syllable, words can be hyphenated almost anywhere.*

2.1.21.2. Arabic and the Arabic script

Due to the calligraphic nature of written Arabic, words can not be divided into syllables. A middle letter would become an initial or final letter, which is in contradiction with the scriptwriting rules of the Arabic script.
The syllables would be unrecognizable for any Arabic native speaker.
The same applies to any language written in the Arabic script. The Maltese language is written in the Latin script and can be hyphenated. However, the Indo-European languages Persian and Urdu are not hyphenated!

2.1.22. Hyphenation of the Austronesian languages

The Bahasa Indonesia, Bahasa Melayu and Tagalog (Pilipino) languages belong to a family of languages unrelated to the European languages. The hyphenation of this branch of languages is discussed in Chapter 4.

2.2. Epilogue

Hyphenation principles are different for different languages. However, the number of basic ideas behind hyphenation is limited. Yet they vary considerable. If pronunciation dictates hyphenation, then hyphenation is phonetical (French, Portuguese). Syllables follow the flow of speech. Hyphenation is etymological when the individual parts of words count (Dutch, German, Swedish). Then prefixes, suffixes and compound determine the location of hyphens. Hyphenation is morphological if word endings are taken into consideration too (Icelandic, Faroese, Morph. Norwegian).

In practice the boundaries between principles are not as strict as the names of principles suggest, e.g. morphological hyphenation also considers the etymology of words. For those cases in which the morphological principles do not apply, phonological rules might be used. However there is a strict order in which principles should be applied, i.e. etymological rules take precedence over phonological ones.

2.3. References

1 Schuring, G.K. 1993. Sensusdata oor die tale van Suid-Afrika in 1991. Ongepubliseerde werksdokument. Pretoria: RGN.

2 Suid-Afrikaanse Akademie vir Wetenskap en Kuns, *Afrikaanse woordelys en spelreëls*, Tafelberg, Kaapstad, 2002.

3 Diccionario Eskara-Castellano/Castellano-Euskara, "Bostak Bat" Iantaldea, Bilbao, 1995.

4 King, A.R., and Elordi, B.O., Colloquial Basque, A Complete Language Course, Routledge, London and New York, 1996.

5 Gallinia, A.M., Grammatica della Lingua Catalana, Editorial Barcino, Barcelona,1969.

6 Diccionari ortogràfic i de Pronúncia, Jesús M. Giralt i Radigales (ed.), Enciclopèdia Catalana, Barcelona, 1990.

7 Retskrivningsordbogen, 4. udgave, Dansk Sprognævn, Aschehoug, Copenhagen, 2012.

8 Grafisk Ordbog, Jan Eskilden (red.), AGI, København, 1998.

9a Woordenlijst Nederlandse taal, Nederlandse Taalunie, Sdu/Standaard, Den Haag/Antwerpen, 1995.

9b het Groene Boekje, Woordenlijst Nederlandse taal, Nederlandse Taalunie, Sdu/Standaard, Den Haag/Antwerpen, 2005.

10a Webster's New Twentieth Century Dictionary Unabridged, Second-Edition, Printice Hall Press, New York, 1983.

10b Webster's Third New International Dictionary, Unabridged, Merrian-Webster Inc., Springfield, Massachusetts, USA, 1993.

11 Dictionary of Contemporary English, New edition, Longman Group UK, Harlow Essex, England, 1987.

12 Webster's Third New International Dictionary Unabridged, Marrian-Webster, Springfield Massachusetts, 1993.

13 Holman, E., Handbook of Finnish Verbs, Suomalaisen Kirjallisuuden Seura, Vaasa Oy, 1984.

14 Mdm. Catach, Personal communications, CNRS-HESO, 1990.

15 Dizionario Fondamentale della Lingua Italiana, Istituto Geografico de Agostini, Firenze, 1987

16 Lo Zingarelli 1998, Vocabolario della Lingua Italiana, Bologna, 1998.

18 Die Rechtschreibung, 18. Auflage, Duden Band 1, Dudenverlag, Mannheim/Wien/Zürich, 1980.

19a Die deutsche Rechtschreibung, Die neuen Regeln, 21. Auflage, Duden Band 1, Dudenverlag, Mannheim/Leipzig/Wien/Zürich, 1996.

19b Die deutsche Rechtschreibung, 24. Auflage, Duden Band 1, Dudenverlag, Mannheim/Leipzig/Wien/Zürich, 2006.

19c Die deutsche Rechtschreibung, 25. Auflage, Duden Band 1, Dudenverlag, Mannheim/Leipzig/Wien/Zürich, 2009.

19d Die deutsche Rechtschreibung, 7. Auflage, Wahrig, Bertelsmann, Wahrig, Cornelsen, München, 2009.

19e Die deutsche Rechtschreibung, 26. Auflage, Duden Band 1, Dudenverlag, Mannheim/Leipzig/Wien/Zürich, 2013.

20 Wie sagt man in der Schweiz?, Dudenverlag, Mannheim/Wien/Zürich, 1989.

21 Íslensk/Ensk Orðabók, IÐUNN, Reykjavík, 1989.

23 Store rettskrivningsordbok, bokmål, Tanums, Kunnskapsforlaget, Oslo, 1996.

24 Prontuário Ortográfico e guia da língua portuguesa, Editorial Notícias, Lisboa, 1999.

25 Dicionário Prático Inglês-Português, Melhoramentos, Saõ Paulo, 1987.

26 Gran Diccionario de la Lengua española, SGEL, Madrid, 1985.

27 Avstavningslexikon, Tryckeriförlaget, Solna, 1989.

28 Svenska Skrivregler, Svenska Språknämnden, Almqvist & Wiksell Förlag, Uppsala, 1993.

CHAPTER 3

3. **A heritage of 5000 years**

All spoken languages consist of a flow of rhythmics. These rhythms originate both from the division of sentences into words and from a finer division within the words: the syllables. The nature of syllables is intriguing. The written representation of a language only partially reflects the spoken language. Orthography, the conventional spelling system of a language, often includes elements of historical development. Hyphenation heavily leans on syllabic information which is present in spoken language but this information is not always coded in the alphabetic representations of languages. The complexity of hyphenation, and also of spelling, has consequences for the implementation of hyphenation and orthography in computer systems. As those systems process and produce large amounts of texts, the lack of control over the underlying assumptions of a language corrupt the results of these automated processes.

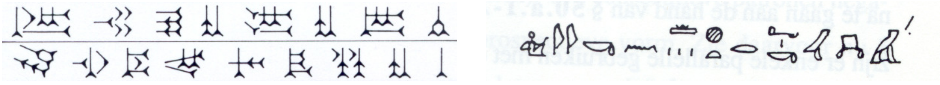


Fig. 3.1: Elamite cuneiform inscriptions[1] and Middle Kingdom hieroglyphs[4]

The written representation of language has a long history which started proximately 3300 B.C. with pictograms (a simple representation of an object) and ideograms (a character symbolizing the idea). These initial writings were the cuneiform (in Latin *cuneas* means corner) in Sumeria and the hieroglyphs in Egypt[†]. In both forms of writing a complete system of phonograms developed. In this system the actual sound of a language was encoded in signs representing, unit by unit, the phonemes as they were pronounced by users of those languages. From the moment the alphabet was born its use extended to both Semitic and non-Semitic languages. Given the nature of the language spoken by the Semites they could not write words like Chinese ideograms, which encode a monosyllabic language with unchanging words. In the Semitic language many words comprise more than one syllable and the change of consonants and vowels (alternations) plays a grammatical role (e.g. they distinguish the singular and plural of nouns, as well as different verb forms). The major part of the basic vocabulary of any Semitic language consists of radicals represented by two, three, or sometimes four or more consonants. Flexions are created by changing the vocalic sound. Take for example the letters *k, t* and *b* in standard written Hebrew. The written word KTB[‡] pronounced as "KaTaBa" means "he has written", but "KiTaB" means "book". Both meanings appear as KTB and the

[†] The original Egypt language belongs to the Afro-Asiatic language family, previously also named Hamito-Semitic. The Akkadian language, the oldest known Semitic language also is Hamito-Semitic language[4].

context alone gives them their meaning (see de Kerckhove, 1988)[3]

The common ancestor of all alphabetic writing systems existing today, is the so-called Proto-Canaanite script. Introduced by the Canaanites the script developed under the influence of the Egyptian uniconsonantal hieroglyphs in the first half of the second millennium B.C. The original Egyptian bi- and triconsonantal pictographic symbols fell in disuse and their scripts became more comparable (alphabetic) to the ones of the Semites. In the middle of the eleventh century B.C. writing became linear, from the right to left: the Phoenician alphabet was born.

The Western theories reached some consensus about the Semitic origin of the Greek alphabet. According to the Greek legend the Phoenicians introduced the Greeks to the alphabet. The names of the letters *alpha, beta, gamma, delta,* etc. have no meaning in Greek, but probably are equivalent to the Semitic equivalents, *alef, bet, gimel, dalet*, which are Semitic words. The archaic Greek Script used the 22 West Semitic letters. Some of these letters were vowels. Gradually five supplementary letters were introduced: Upsilon, then Phi, Chi, Psi and finally Omega. The writing direction was not yet firmly established and changed in a later period from bidirectional to left to right. The archaic Greek alphabet did not consist of one set of letters, but had various local forms (see Naveh, 1988)[3].

In the Semitic family of languages the lexemes — the roots of individual words - consisted of consonants. They were sufficient for the production of meaning. Vocalic sounds played an essential role in verbal or nominal derivation, i.e. to create grammatical rules. Canaanite writings did not reflect complete vocalization, they rather were a way to produce meaning. Even until today Semitic languages rarely adopted the use of vowels. The true Canaanite breakthrough, grafted onto the invention of the consonantic syllables, was its adaptability to every Semitic language. This innovation, in turn, led to a second one: the notation of vowels. As opposed to Semitic languages Greek words beginning with vocalic sounds were common in the Greek language. Therefore the first Greek authors began to adapt the Phoenician alphabet to the needs of their own language.

Syllables in Semitic words consisted of separate letters, in Greek and Latin each character became a phoneme. The latter languages have vocalic alpha-

[‡] In modern Hebrew the orthography of KTB is spelled as Kaf, Tav, Bet כתב. The Webster's Hebrew dictionary[2] gives the following transcriptions: katav (v. wrote), katav/katevet (nmf. reporter), ketav (n. writing, script), all written with the same Hebrew characters. In modern Hebrew the word Book is transcribed as (sefer) Sameth, Feh, Resh ספר.

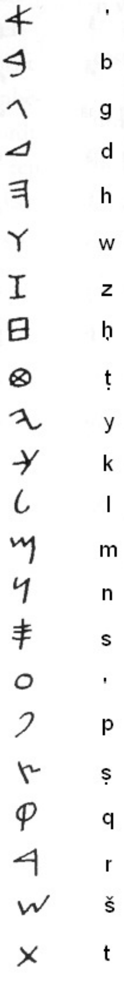


Fig. 3.2: the Phoenician alphabet, 1000 B.C.

bets in which syllables consist of a linear sequence of characters. Word and syllable boundaries were originally omitted in ancient scripts. They became a continuous stream of phonemes.

Under the influence of writing from left to right the form of some letters became more homogenized. That is, the direction of curves of the letters tended to the right, e.g. the letters B, C, D, F, P, R. During Romanization the letter G was introduced and P (rho) became R.

Societies have often changed the alphabets that transcribe their languages. The English language is a system consisting of a complex set of added logographs (characters representing a word or phrase) which are conventionally used for special purposes. There is no consistent mapping between graphemes (the smallest unit in a writing system) and sounds. Symbols that do not represent a sound can alter the sound of another symbol. Consider the difference between us and use (verb) caused by the silent e. The Finnish and Serbo-Croatian languages have a fully phonetic alphabet.

The alphabet is a graphical approximation of the phonemes. Some phonemes are represented by different graphemes and some graphemes have different phonemes. In some languages an alphabet of 26 letters is not sufficient and diacritical marks are used for accentuation of the vowel or a change in sound (the German umlaut). In the Scandinavian languages they represent a separated vowel which in dictionaries follows the Z. Accentuation of vowels effects hyphenation (e.g. French and Spanish). Increasing the number of vowels increases the number of possible syllables and therefore the difficulty of hyphenation. In Swedish a difference should be made between *gra..., grä...,* and *grå...* In the Baltic and Slavic languages some letters have a diacritic to palatalize (to soften) a sound (compare the sound of *keys* with *cheese*). This palatalization feature roughly divides the Indo-European languages into western non-palatalized languages having a 'hard' *g* or *k* and the eastern languages having a *j,z* or *s* in the same place. Through this division in palatalization languages can be classified as *centrum* languages (Latin centrum = hundrud) and those languages without palatalization, the *satem* languages (Sanskrit *satam* 'hundred', Avestan, an Indo Aryan language, *satem* hundred, Lithuanian *šimtas* 'hundred')[6].

3.1. **Today's Languages and Indo-European Ancestry**

The Indo-European family of languages, the current day European languages, are all descended from the prehistoric Proto-Indo-European language, which was spoken in an as yet unidentified area between eastern Europe and the Aral sea (fifth millennium B.C.). Today's European branches consist of the main groups: Balto-Slavic, Germanic, Celtic, Italic, Albanian, Hellenic, Anatolian, Indo-Iranian[5]. Tocharian, an eastern branch with earliest record known from the seventh century A.D., does not exist anymore (last records dated 1000 A.D.). Tocharian has just recently been discovered in manuscripts found in the Tarim Basin's Täklimakan Desert in the north west of China[6]. The Hittite language, spoken by an ancient people in Minor Asia, is extinct too. Hittite belonged to the Indo-Iranian branch, as current day Armenian, Persian (Farsi). Languages make up a family when they share a *common root*. It is generally assumed that somewhere in the past one common language was spoken, it grew apart with time when people spread over the continent. Nevertheless, the root of the Indo-European family of languages is still present. The word **Manu-* has been preserved in the English *man-*, as well as in Dutch, German, and the Scandinavian languages.

Indo-European was a highly inflected language. The Hellenic, Italic and Balto-Slavic group are still heavily inflected.
Except for Icelandic and Faroese, most of the Germanic languages have abandoned many of these inflections. These two languages did not change due to a long history of isolation after the Norwegian settlements around 900.

The only languages in Europe which are not of Indo-European origin are the Finno-Ugric languages *Finnish, Sami* (the language of the Laps), *Estonian*, and *Hungarian*. *Basque* is unrelated to any language in the world. Because of the geographical position in Europe these languages will be discussed in a separate section.

3.1.1. **The Germanic languages**

The Germanic languages consist of 3 groups called the North Germanic, the West Germanic and the East Germanic languages. The language of the Ostrogoths — the only East Germanic language — is Gothic, which no longer exists. Nowadays the word Goth means *a barbarian; an uncouth, uncivilized person*[13].

3.1.1.1. **The Scandinavian languages**

The North Germanic languages consist of the Icelandic, the Faroese, the Norwegian, the Swedish, and the Danish languages. Except for Icelandic and Faroese, these languages developed from Old Scandinavian, via Middle Scandinavian variants to today's languages. Icelandic and Faroese developed out of Old Norse but due to isolation these languages didn't change much.

Iceland is located just below the Arctic circle. It was settled from 870 A.D. onwards by people from Western Norway. Early Icelandic literature has been recorded, however, the stories of Odin and the other Germanic Gods have nearly been lost. There were even settlements in Greenland, which completely vanished around 1600 A.D. Icelandic uses the vowels á, é, í, ó, ú, ý, æ, ö and the consonants þ and ð, e.g., Alþing (the Icelandic National Assembly).

The Faroe Islands are located about 250 miles north of Scotland, midway between Norway and Iceland. They were settled about a thousand years ago by Norwegian Vikings speaking the Old Norse language. Modern Faroese, like Icelandic, strongly resembles Old Norse. It is spoken by most of the islands’ 40,000 inhabitants, although the official language is Danish. The alphabet contains the d (but not the t) of Icelandic, and the r of Danish instead of the Icelandic ö. While the grammatical cases of Old Norse disappeared in most of the North Germanic languages, they were preserved in Iceland and Faroese.

One of the remarkable attributes of the Scandinavian languages is the attached article. Except for the defined article singular, the article is added to the end of a word, e.g., in Swedish: ett besök (a visit), besöket (the visit), besöken (the visits), besöker (visits) en biljett (a ticket), biljetten (the ticket), biljetter (tickets), biljetterna (the tickets). For the genitive case an **s** is added to the end of a word. The attachments of articles differ slightly between the Scandinavian languages. Icelandic and Faroese also add the nominative, dative, and accusative cases to the nouns. Like in German certain prepositions are connected to cases.

Spelling/orthography differs between the languages themselves. Swedish uses å, ä, ö as additional vowels, e.g., lärling (apprentice). Norwegian and Danish use å, æ, ø, e.g., lærling.

The Nordic languages have been influenced by Greek and Latin too. In general the c in words from Latin/French origin is spelled as a k: kombinationen, inskriptionen, stationen, kreativare (Swedish). In French these words are spelled as combination, inscription, station. In Norwegian spelling of ..tion has been changed to ...sjon (versjon, stasjon, etc).

A considerable number of compounds exists in the Scandinavian languages, and new compounds develop daily by combining other words. This mechanism depends on the meaning of words and some combinations are impossible, e.g., in Norwegian (Bokmål) the word *linjeoffiser* exists (an officer on a shipping line) but *linjéroffiser* would be nonsense because it points to an activity of drawing lines on the officer and not to the attribute of being a shipping line. These restrictions in connotation differ from language to language and global rules valid for all languages don’t exist.

Swedish

The Svenska Akademiens Ordlista[7] (a dictionary of the Swedish Academy) was published for the first time in 1874, using spelling of Swedish which was very different from today’s spelling. A text of 1907 highlights the differences in orthography in bold[8].

Laura Fitinghoff: Barnen ifrån Frostmofjället (1907)

> FÖRETAL.
> Vid en diskussion om "Vlra barns nöjesläsning", som jag
> för en del Ir **se’n öfvervar**, fällde en känd talare, en
> s. k." friräsare" yttranden i frlgan, som **gingo** ut pl
> att nämnda art **af** litteratur nu närmast tyckes vara **afsedd**
> för endast den bildade klassens barn. De **äro** hjältarna,
> hjältinnorna i barnböckerna, när dessa senare **äro af** den rätta arten.
> När dessa hjältar och hjältinnor **själfva**, eller genom ädelmodiga
> "upphöjda" anhöriga komma i kontakt med allmogens barn och
> barnen ur de djupa leden blir detta till en rörande handling.

These texts have been made digitally available by the Runeberg Project. The great value of this project is its historical records of texts, which are to the disposal of to everybody.

From 1970 to 1980 it was necessary to include many new words from science and other technical developments in society as noted by the Svenska Akademien. This process is still going on. The same applies to the Norwegian and Danish idiom. Today’s dictionaries have become larger than ever before.

Swedish is spoken by 9 million people. It is the only language in Sweden, but dialects sound different between the north and the south. In the very north, and also in the Nordic regions of Norwegian and Finland, a small population of Laps speak Sami. Sweden belongs to the European Union.

Norwegian

There are 4-5 million people in Norway. Due to the Gulf Stream the climate is temperate, and people have settled along the coast line. Under the foreign rule of the Danes Bokmål (Book language) became the most important language, patterned after Danish for centuries. Nynorsk (New Norwegian) was constructed from the Norwegian dialects and shows characteristics of the original Norse language. Nynorsk is taught in school but scarcely spoken except for the countryside. The attached article differs between Bokmål and Nynorsk[9,10]:

Nynorsk	Bokmål	translated
addisjon [-en,-ar,-ane]	addisjon [-en,-er,-ene]	(addition)
absorbering [-a,-ar,-ane]	absorbering [-a, or -en,-er,-ene:]	(absorption)

vakker, vakrare, vakrast	vakker, vakrere, vakrest	(beautiful)

Some lemmata can be found both in Nynorsk and Bokmål. Other words should not be mixed[11].

Nynorsk	Bokmål	
hermeteikn (-et) gås(e)auge	anførselstegn (Subst)	(quotation mark)
vedkom(a)ande denne den nemnde	angjeldende (adj)	(... in question)

In 2001 Norwegian will have a new orthography. There will be no longer hovedformer (main spelling) and sideformer (spelling variation) in bokmål and there are new words:

> brystvortering *piercing, en ny form for frigjøring — eller en gammel form for slavebinding (Breast nipple piercing, a new form of liberation — or an old form of slave binding)*
>
> dametukler *a lady fiddler ... mange regner med at Bill Clinton vil kjempe for å sette sitt merke i historiebøkene som noe annet enn dametukler (... many people will count on the fight of Bill Clinton to get a nomination as lady fiddler in the history books) (newspaper: Dagbladet 13.2.1999).*

These examples demonstrate the openness and understatements of the Norwegians.

Danish

Danish once was the most influential language of Scandinavia. During the Viking age the Danes conquered a large part of England and established the Danelaw (also spelt as Danelagh). The Danegeld was the money to be paid as tax.

There are 5 million Danes, but their language is also taught in school in Iceland, Greenland, and the Faroe Islands. The southern part, at the German border, is bilingual. Typical for Danish is the swallowing of consonants in speech. The prefix *ind-* is pronounced as *inn-* (*indpakke (to wrap in), indskibning (embarkation), indslusning (gradual absorption), etc.*). In contrast to Danish Norwegian adapted the spelling to *inn-*. The first spelling list, Dansk Haandordbog, was published by Svend Grundtvig in 1872. In 1891 Viggo Saaby's Dansk Retskrivingsordbog was published and new editions were published by Thorsen between 1904 and 1919. In 1955 the first version of the Retskrivningsordbog of the Dansk Sprognævn considered as the official Danish guideline for spelling was published. The most recent edition was published in 2012[12], it includes a series of orthographic changes.

3.1.1.2. **The West Germanic languages**

The West Germanic languages are a subdivision of the Germanic languages and include English, Frisian, Dutch, Flemish, Afrikaans, High German, Low German, and Yiddish. The articles of these languages precede the nouns: the house (English), het huis (Dutch), das Haus (German), die huis (Afrikaans), 't hûs (Frisian). Like the North Germanic languages, West Germanic languages are flavoured with compounds.

English

English is closely related to Dutch and Frisian, e.g., the English word *brother* is *broer* in Dutch and Frisian and brōthar in Old Saxon. Even the Lithuanian *broter-* shows the Indo-European roots. Language is the predominant factor in the social organization of humans and as the organization is changing so does language. In English history there have been periods of sudden change, such as the Norman Conquest (1066-c.1120) and the incursions of the Scandinavian invaders (850-1042). Many Old English words are still recognizable: *cynn* → *kin*, *munuc* → *monk*, *gēar* → *year*. Much more impressive were the Latin/French influences from the Norman conquerors, e.g., *air, bacon, bucket, fry, heritage, honour, noble*. Another category of influences came from the church: *attorney, forfeit, pillory, parson*, and *penance*. From medical discipline came words like *anatomy, poison*, and *stomach*[13]. Modern written English, despite its different pronunciation, has a lot in common with the French language:

abatement	abattement
abater	abatteur
abbey	abbaye
abdication	abdication
abduction	abduction
aberration	aberration

English is not the only language heavily influenced by Latin and French. All European languages were affected by the classic languages and their culture and the church strongly expedited this process.

The Latin/French influence is superficial. The compound mechanism is pure Germanic. It is *schoolmaster* and not *master of the school* (maître d'école) The verb structure is Germanic too:

spring	sprang	sprung	
spring	sprong	gesprongen	(Dutch)
spring	sprang	gesprungen	(German)
springe	sprang	sprunget	(Norwegian)

Spelling achieved its modern form around 1650. New words came from the English colonies. After the United States became independent on July 4 1776 the American English language became influential itself. The first colonists adopted Indian words for objects they never had seen before, e.g., *wigwam, moccasins, squaws, raccoon, opossum*. Other sources of foreign words came from French settlers and the American English language adopted new words like *bayou, butte, crevasse, prairie* and *rapids*.

Frisian

On the basis of sound shifts the Frisian language, like the English language, belongs to the Anglo-Frisian sub branch. The Anglo-Frisian dialects were originally spoken along the North Sea coast. These coastal dialects were also called *Ingwinian*. After the Saxons invaded Britain, but not Frisia, the idiom of Frisian began to differ from English.

Compounds: comparable to the other Germanic languages compounds are extensively used. The possibility to form compounds depends on semantics. *Ruterlearzen* is a normal compound but the compound *rutlearzen* would be an impossible compound. *Regearpartij* is correct, but *regearjepartij* is nonsense.

Dutch

Dutch developed from Old Dutch, via Middle Dutch to today's Dutch. The Old Dutch was characterized by strong vowel endings, e.g., *tongo* degraded to *tonge*[14] in Middle Dutch and to *tong* in today's Dutch.

The Dutch spelling has been modified frequently. A recent spelling reform introduced in 1996 has been modified once more in October 2005 (De Nieuwe Nederlandse Spelling). A former disagreement about spelling details seems to be resolved between "Het Groene Boekje", an official publication of the Dutch Lan-

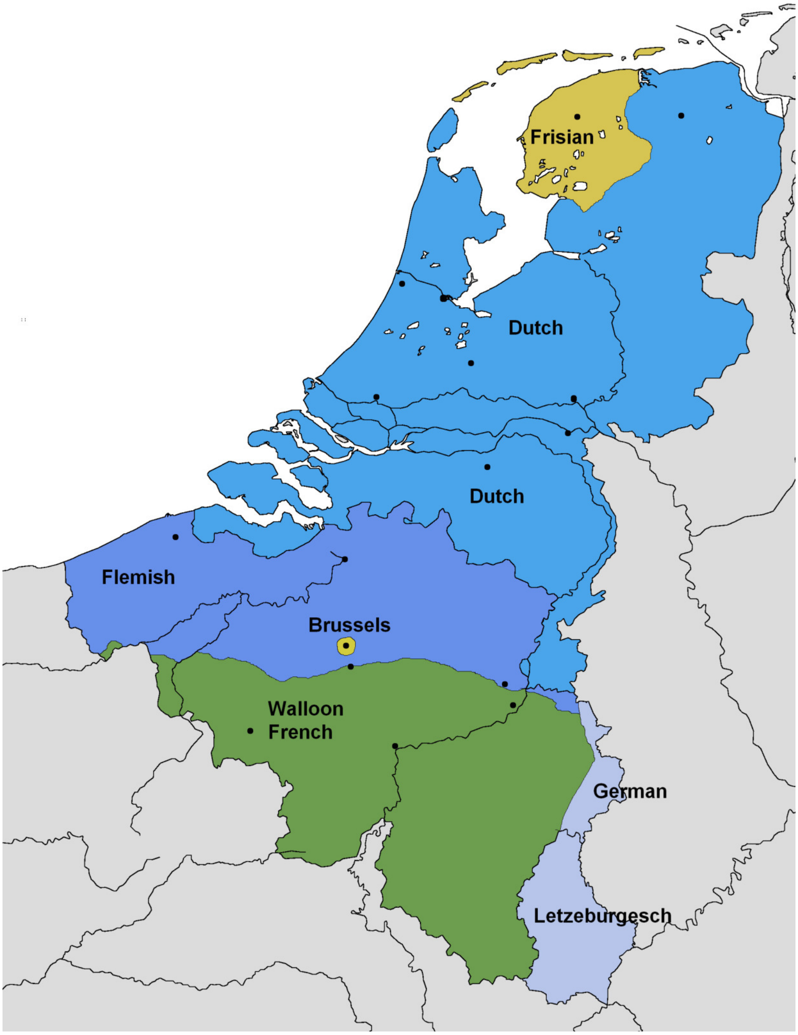
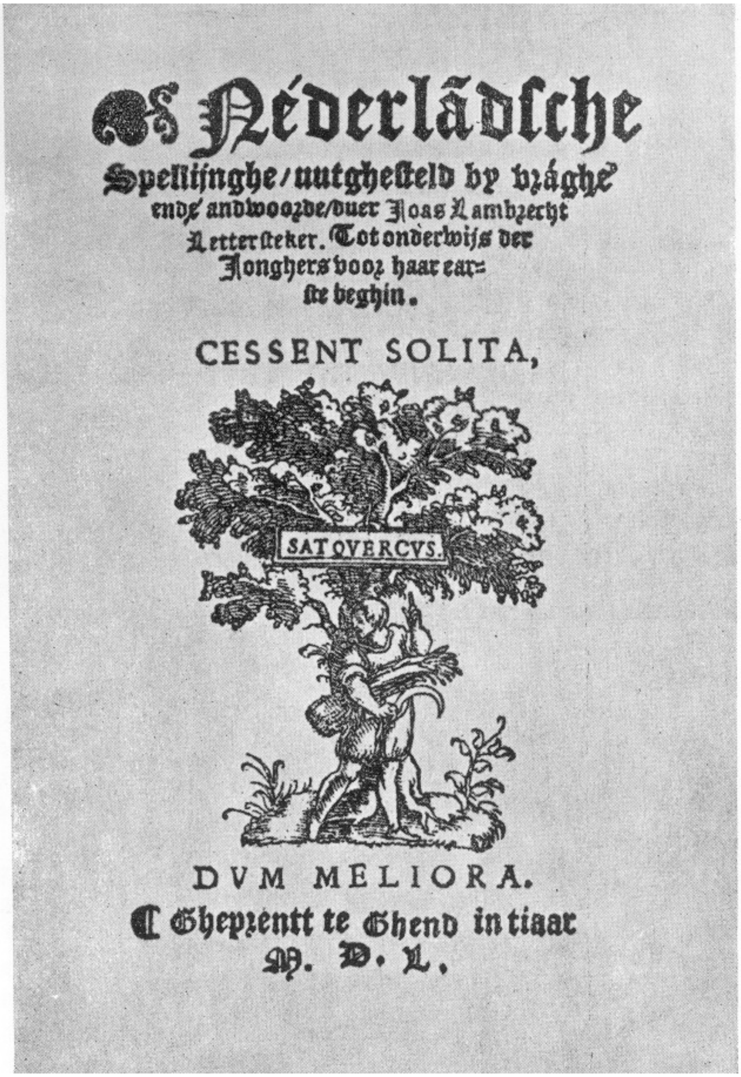


Fig. 3.3.: The languages of the Netherlands and Belgium.
Brussels, the governmental centre of Belgium, is bilingual.

guage Union (de Nederlandse Taal Unie), an official cooperation between the Dutch and Flemish governments, and the huge Dictionary of the Dutch Language (Groot Woordenboek der Nederlandse Taal) of Van Dale Lexicography. These dictionaries will be corresponding from October 2005.

Conflicts in spelling could have been expected since already a few centuries ago, at the time Dutch was first being written and printed, different spelling pro-



De Néderlādſche
Spellijnghe/uutgheſteld by vꝛághe
endē andwooꝛde/duer Joas Lambꝛecht
Letterſteker. Tot onderwijſs der
Jonghers vooꝛ haar ear=
ſte beghin.

CESSENT SOLITA,

SATQVERCVS.

DVM MELIORA.
¶ Ghepꝛentt te Ghend in tiaar
M. D. L.

Fig. 3.4.: The title page of one of the earliest Dutch spelling guides.

posals had been suggested. Joos Lambrecht, one of the first to address spelling issues, wrote — *de Nederlandsche Spellijnghe, Gheprentt te Ghend in tiaar 1550*. This was the first attempt to standardize Dutch Spelling, followed in 1585 by Spieghel's *Twe-spraack vande Nederduitsche Letterkunst door Hendrik Laurensz Spieghel uytghegheven by de Kamer In Liefd Bloeyende t'Amstelredam*. These attempts did not succeed in establishing a standard orthography. The Dutch language from Antwerp as used by Anna Bijns and the language from Amsterdam as used by Roemer Visser were too different. Even nearby cities had different orthographies: *Inden Haegh seytmen ghewassen* (Den Haag) and *t'Amsterdam seytmen ghewossen* (Amsterdam).

Finally, in 1850 a dictionary concept was proposed by De Vries and Te Winkel. Their aim was to reduce the chaos in spelling, but a long, violent struggle followed. In 1934 minister Marchant introduced a simplified spelling in the educational sector. Due to the Second World War his Spelling decision was not effected until 1946. The Word List of the Dutch language (De Woordenlijst der Nederlandse taal) "het Groene Boekje" was published in 1954. A ministerial decision prescribed that the preferential spelling had to be used. The alternative spelling, however, was not considered incorrect! So the conditions for another spelling reform had been created. In 1963 the Commission Pée/Wesselings was requested to reform the spelling again, but the reactions were violent. On the one side there was the "Aksiegroep spellingvereenvouding" (Action group spelling simplification) which preferred a more radical simplification, and on the opposing side were mostly writers and academicians who were against these proposals. This spelling reform was never introduced.

The Society for Scientific Spelling (Vereniging voor wetenschappelijke spelling) founded in 1963 is still alive. This society was founded by the linguist prof. dr. P.C. Paardekoper, not to be reformed in current orthography as Paarde**n**kooper. This society propagated an orthography as *bebie (baby), sjantaazje (chantage), sjurt (shirt), bèzje (beige), jij antwoort (antwoordt, he answers), zij vint (vindt, she finds), de hont (hond, dog) blaft*. These ideas have never been adopted.

In 1990 a new commission was appointed, the spelling commission Geerts. It was decided to accept the preferential orthography as the only spelling, along with some cosmetic adjustments. Differences in the old list as *kopie* next to *fotocopie* (1954) would be removed. However, in 1995 the conflicts between look-alikes were still present, e.g., *foerage, foerageren* against *fourageergebied*. But worse are: *statieportret* (staatsieportret, a official portrait) of Queen Beatrix to analogy *statiegeld* (deposit). Het Groene Boekje and Van Dale used different orthographic rules. In a small category of words Van Dale deviated from the official guideline of the Taal Unie. This resulted in a more simplified spelling: *paardenbloem* instead of *paardebloem*, in analogy with *paardenstaart*. Why should we make a language unnecessarily complex? Because of a difference in rules there were *dienstbodenkamers* (Van Dale) and *dienstbodekamers* (the Groene Boekje). In the past centuries the servant girl's (dienstbode) room (kamer) was somewhere in the attic of the house.

But what can we expect from October 2005 onwards?

Flemish

From the 12th century on, Flemish literature is known. The first *Reinaert*, the mediaeval story about the smart fox Reinaert, was written in East Flanders, the second in West Flanders. Differences between East- and West Flemish still exist, so do differences between the North and the South of Flanders. East Flemish is closer to the dialect in the province of Brabant. The gutturals and labials of the southern regions of Holland used were close to those of the West Flemish dialect. Holland was repeatedly influenced by Flemish merchants from Antwerp resettling in Amsterdam during the 80 years war. Today standardized Dutch is accepted both by Flemish sections of Belgium and the Netherlands. In order to enable a cross-border standardization, the Dutch Taal Unie has been established by the two governments. There are small differences in usage of Dutch idiom, but these differences are not considered to be a form of dialect. Belgian newspapers write and promote standardized Dutch, e.g., in Stijlboek of the Standaard, a leading Belgian newspaper. Ludo Permentier and Ludo van der Eynden stress that they belong to the Dutch speaking community and they avoid using typical regional idiom[15].

Afrikaans

Afrikaans is the youngest Germanic language. It stems from seventeenth-century Dutch, from which it developed more or less independently for three centuries on the African continent. There are less declinations and inflections. The verb structure has been simplified compared to current Dutch.

Afrikaans	Dutch
ek spring	ik spring
hy spring	hij springt
ons spring	wij springen

In spelling lost consonants reappear in plural, e.g. reg, regte (Dutch recht, rechten).

High German

In the Middle Ages such language attributes as an “Gemeinsprache” (a general accepted language) in political borders were not existing. From the North to the South of Germany the language gradually underwent different influences. In the Northern part Low German dialects were spoken, near the Alps, the language of the tribe of the Alemanni. This language was the precursor of Swiss German. The southern part of Germany had been occupied by the Romans, but in the Northern part there had always been the free German homeland. The West bank of the Rhine was Roman territory. Many words came from Latin Pfalz (the Roman Paladium at the river Rhine), pflanzen, Strasse, etc. These Romanizations were quite early and underwent a second change in sound. On the other hand Germanic names were linked to Christianity. From the Germanic goddess Ostara came "Ostern" (Easter). But many other words were introduced by the Church.

The issue of spelling started with the invention of the art of printing.
It is said that printing art came from China, but the information that came along the Silk Route or other routes is scarce and probably reached a few people only.
Laurens Janszoon Coster printed the "Speculum" in 1440 and from 1455 remains a fragment of a German poem about the Apocalypse from Johan Gutenberg. It wasn't until the 16th century before the subject of standardization of orthography was really initiated.
Gradually changes in orthography were accepted. In the 16/17th century the "ß" developed from the voiceless "s".
In the 19th century Jacob Grimm proposed to abolish the Dehnungs-h in those words where no scientific arguments in favour of its use could be found[16]. Today this "h" is still in geographical names as "Berlin-Johanisthal" or "Frankenthal".

Conferences on the orthography took place in Hannover (1854), Leipzig (1857), Stuttgart (1861 and Berlin (1871). But finally in 1880 Konrad Duden published his first orthographic dictionary "Vollständigen Orthographischen Wörterbuch der deutsche Sprache. Nach den neuen preußischen und bayrischen Regeln (Complete Orthographic Dictionary of the German language. According to the new Prussian and Bavarian rules)"[17]. This was the start of standardization of the German language in the German area. Today the Duden has become the authority on German orthography.

In 1901 standardization was proposed again and a discussion on a new orthography resulted in "Regeln für die Deutsche Rechtschreibung nebst Wörterverzeichnis" (Conference Berlin). If the pronunciation of the "c" in foreign words was "k" or if it sounded "z", it was accepted to use these sounds in spelling (akkusative, Porzellan). The other foreign words kept the "c" (Café, Chef). The "th" was replaced by the "t" (thun → tun, Thor → Tor).

The simplification of the Minuscule and Majuscule writing of Oskar Brenner[18] followed in 1902, but after 1933 there were no more reforms.

In 1955 at a conference of ministers of the German federation it was decided to accept the 1901 rules, and in case of doubt the Duden had to be consulted!

Hyphenation in some cases is still confusing: according to etymological rules — the Duden — or the rules of the conference of 1955.

After a long period of preparations starting in the 1970s an interstate declaration about the new German orthography was signed. It was intended that the new rules should reduce the language difficulties. The most important change was the exchange of the ß for ss after short vowels: "daß" became "dass", and "Mißbrauch" became "Missbrauch". Words which were earlier written together were separated: "allzuoft -> allzu oft", "aufeinanderfolgen -> aufeinander folgen". Some of them were separated and the first part of the compound was written in majuscule: "holzverarbeitende -> Holz verarbeitende". Not all case became more simple, like the double forms: "aufwendig" as well as "aufwändig", "Delphin" as well as "Delfin".

The new uncertainties have created substantial conflicts. As well as the Duden[19] the Bertelsmann[20] was introduced. Additionally to the differences of the two large publishing houses, the German Press Agencies tried to eliminate double forms, intended to be used in the domain of the pre-press (newspapers, publishing houses).

A transitional period in which both the old and new orthographies were accepted should have expired from August 1st 2005 on. It was not accepted for federal states of Bayern and North Rhine-Westfalia and the internal opposition against the reform persisted. To solve the conflict the Council for the German Orthography (Rat für deutsche Rechtschreibung[21]) was founded in December 2004. This council consisted of members of the six German countries being language professionals such as scientists in the field of orthography or lexicography. The result they produced was a **second reform** which came in to action on August 1st 2006. It was a rather radical update of the German language which again consisted of a lot of changes. The Duden from 2006 [22] came up with a lot of double orthographical forms such as "*bismarck[i]schen*" next to the upper case variant "*Bismarck'schen*". The orthography of many words was changed too: "Bankrott gehen, zu Eigen geben, jenseits von gut und böse" (1996/2004) became "bankrottgehen, zu eigen geben, jenseits von Gut und Böse" (after 2006), in English going into bankrupcy, giving somebody a gift, on the other side of good and bad. Dictionaries did introduce a new item — recommended orthographics, the most advisable choice between alternatives and the German Press Agencies (Deutsch Presse Agenturen) presented their own preference list in 2007 including cases which deviated from the mainstream dictionaries. The result was certainly not a unification of the German language. Austria and Switzerland accepted the 2006 reform but had their own idiom. Switzerland wrote “ss” instead of “ß” in words as “schiessen” (not “schießen”), in English “to shoot”.

It was not unexpected to see many errors in printed publications due to this variability in orthography, especially the lower and upper case writings (“Mein, Dein, ...” versus “mein, dein, ...”) and in writings of open or closed compounds, e.g. “kleinschreiben” instead of “klein schreiben” (error, to write in lower case), “entgegenlaufen” instead of “entgegen laufen” (error, to come towards s.o.).

However, after 2006 the matter of open or closed compounds became less restricted and a lot of double orthographic forms became allowable, as can be concluded from the latest edition 25. of the Duden “Die deutsche Rechtschreibung”, 2013.

Low German

Nedersaksisch, Neddersassisch, Nedderdüütsch, Plattdüütsch, Plat, Platt, a collection of dialects related to Low Saxon or Low German, a language variation close to Dutch. Low German is related to the dialects in the southern part of Jutland in Danmark. A lot of dictionaries of the dialects have been published in different parts of North Germany, the Eastern parts of the Netherlands, and the South of Jutland, partly due to a revival of the local dialects. However, the Low German language itself is not yet standardized and is not yet a binding factor. The titles presented below show the varieties in dialects:

Wörterbuch des Münsterländer Platt;
Die dithmarsische Mundart;
Kleines plattdeutsches Wörterbuch:
    für den mecklenburgisch-vorpommerschen Sprachraum;
Hochdeutsch-plattdeutsches Wörterbuch:
    auf der Grundlage ostfriesischer Mundart;
Hümmlinger Wörterbuch - auf der Grundlage der Loruper Mundart;
Plattdeutsches Wörterbuch für das Oldenburger Land;
Dictionary of Groningen Low Saxon;
Woordenboek van de Drentse dialecten;
Nieuw Groninger woordenboek

Yiddish

Yiddish is a High German dialect spoken by many European Jews and their descendants on other continents. It is written in characters of the Hebrew alphabet. Yiddish came into being in the early Middle Ages in Germany. It is based on Middle High German and kept old words and meanings which have been lost in High German itself. Even older words from Hebrew and Aramaic origin, specific to Jewish culture. When the German Jews moved they took Yiddish to other countries. Gradually two branches developed: East Yiddish and West Yiddish. East Yiddish contained elements of the Russian and Polish language. West Yiddish disappeared gradually due to the Jewish assimilation and Hitler's holocaust. One of the interesting issues of West Yiddish, especially in Holland, was the disappearing of the cases, "der, dem, die" became "de", and the genitive and the dative were replaced by "von" and "an"[23,24].

3.1.2. **The Italic languages**

The older Italic languages are Latin, Oscan and Umbrian. Oscan was spoken by the Samnites, Campanians, Apulians, Lucanians, and the other people of Central and southern Italy. Umbrian was probably spoken in the north-central part of Italy. The Latin language had a central position in the Roman Empire.

The modern Italic languages developed from Latin and were influenced by certain local languages that do not exist anymore. The languages on the Iberian Peninsula are: Portuguese, Galician, Spanish, Catalan; the Provençal/Occitan language in the southern part of French; French, the standard language of France; standard Italian on the Italian peninsula; in Switzerland Rhaeto-Romance; and in the East up to the Black Sea the Romanian language. The ancient Oscan and Umbrian languages have been extincted.

Latin

The Latin language shows most affinity to Celtic (ancient Gaul and the Britain languages). This might mean that the people's languages which were derived from Latin and Celtic once were in close contact with each other. This idea is supported by the fact that Roman artefacts have been found outside the regions that were occupied by the Roman armies.

The earliest Latin inscriptions date from 1000 B.C., but inscriptions from 200 B.C. are few and short. The earliest literature comes from Titus Maccius Plautus (c.250-184 B.C.). His plays, for example Rudens, are modelled on Greek New Comedy.

The source of our current and Latin scripts is the Western Greek alphabet, used in a part of the Greek main land and in some of the Greek colonies in southern Italy and Sicily. This alphabet was probably passed on via the Etruscan script of north-central Italy (see fig. 3.5).

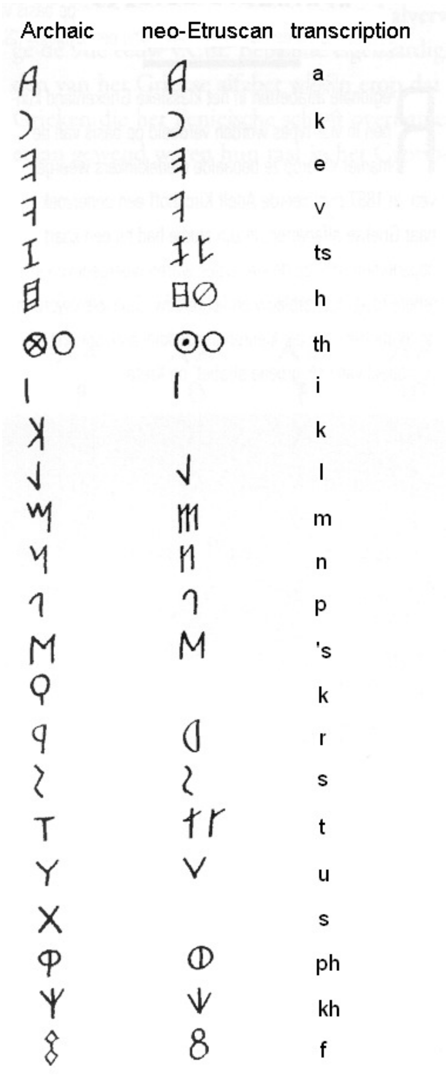

Archaic
neo-Etruscan
transcription
a
k
e
v
ts
h
th
i
k
l
m
n
p
's
k
r
s
t
u
s
ph
kh
f

Fig. 3.5.: the Etruscan alphabet, ca. 750 & 500 B.C.

During the Roman Republic period the Latin alphabet consisted of twenty-one letters:

A B C D E F G H I K L M N O P Q R S T U X

During the reign of Augustus two new letters were introduced for use in words borrowed from Greek; the y (for the Greek υ) and z (for the Greek ξ).
The letter c derived from the Greek γ was at first used for both the sound g and k. As it became to be utilized for the letter k only, a new letter G was formed by adding a small stroke to the C to distinguish the g from k (3th century B.C.).
The letter i became vowel and semivowel as the English y. In mediaeval Latin the j sometime is used (Iulius/Julius, Engl. belonging to July).
The letter u (capital V) is vowel and semivowel like the English letter w, in mediaeval texts sometimes written as v.

After the collapse of the Roman Empire, the light in Europe went out, and probably languages gradually merged. Latin and some Celtic languages disappeared, and new languages were born or melted into other languages. Out of the melting pot came Italian, Portuguese, Spanish, Catalan, Provençal/Occitan, French, Romanian, and Rhaeto-Romance. But there was more, as the meaning of the adjective “Latin” itself was extended, it was linked to different items and, consequently, its meaning changed. It became[12]:

the ancient Latium, or its people;
of the ancient Rome or its people;
of *or* in the language of ancient Latium and ancient Rome;
designating *or* of the languages derived from Latin, the people who speak them, their countries, cultures, etc.
of the Roman Catholic Church

As a noun it became

a native or inhabitant of the ancient Latium or Rome;
a member of one of the modern peoples whose language is derived from Latin;
in Turkey, a person of foreign ancestry belonging to the Roman Catholic Church;
a member of the Roman Catholic Church;
an exercise in schools consisting of translating another language into Latin.

A continent was called Latin America, and their people Latin-Americans. Things could be Latinated (of *or* derived from Latin). There was the Latin Church, the Latin cross, and somebody could become a Latiner, being skilled in Latin, an interpreter.
Also, Paris has its Latin Quarter, and Latin Rites guide the services in the Latin Church.

Catalan (nova ortografia)
Catalan is one of the Italic languages, that is completely independent of the oth-

er ones. It is considered to be a variation of the original provençal dialect. Catalan underwent Gallo-Roman and Ibero-Roman influences. It is spoken from Perpignan in the South of France (Catalunya Nord) up to Alicante in Spain (País Valencià). The Island of Mallorca and L'Alguer on Sardinia also belong to the Catalan region. The "Principat" d'Andorra has chosen in 1993 Catalan as their official language[25].

Catalan clearly became distinct from Provençal/Occitan during the 10th and 11th century. The early documents were from the 12th century, but during the 13th century the Catalan literature (prose) came from Ramon Llull (1233-c.1315) one of the very first writers. The centuries which followed can be qualified as prosperous for the Catalan language and society. During the succession war from 1705-1715 King Philip V suppressed the Catalan language in favour of Castilian (Spanish). Around the fin du siècle a revival of Catalan culture and language took place. Orthographic norms were developed and Catalan dictionaries were published. During the Civil War in the 20th century Catalan was suppressed again. After the death of Franco, Spain was again a democracy and at the end of 20th century the Catalan region became the engine of the economy of Spain. A booming economy led to the revival of the Catalan language. In France Catalan was suppressed since the French Revolution. The French Government quietly discourages the use of any other language then French.

Today Catalunya (the Catalan homeland) has a population of over 6 million Catalans. About 68 % speak Catalan. On the Balearic Islands (Majorca) 66 % of the population speaks Catalan and in the País Valencià 51%.

Typical of the Catalan language is a series of digrams: e.g., *ll* versus *l·l* resp. the Spanish ll and the French ll (village), *ny* (canya, bany), *tg, tj* (homenatge, jutge, viatjar), *tx* (butxaca, cotxe). The diphthongs gu and qu followed by an e or i are written with a diaeresis (ungüent, lingüista, if pronounced as un'gwen or lin'gwiste). If pronounced as a k no diaeresis are used (quietud (kie'tut)).

Like French, Italian and Greek, the Catalan language applies the apostrophe in elisions, e.g., l'arbre, l'himne, l'horitzó, l'hora. *La hora* would mean an orthographic error.

The Catalan language has been reformed October, 2016.

Spanish

Spanish has developed from Old-Castilian, one of the original languages which developed after the defeat of the Romans. The other Roman languages spoken on the Iberian peninsula were: Galician/Portuguese, Asturian-Leonease, Navarro-Aragonese, Catalan and Mozárabe. Mozárabe was the Roman language of the indigenous population under the Arab rule over Southern Spain.

It is difficult to estimate when Castilian became distinct from Latin and the other Roman languages. The year 976 is now considered the official birth year of Spanish, but the very first text was of an earlier date. At that time Latin copyists wrote comments in the margin in their own Roman language, Old-Castilian. The manuscript **Glosas Emilianenses** (in Latin) is from around 975. A series of short remarks are found in the margin. The Latin text *qui .. pauberibibus reddet*, he who gives to the poor, is annotated with

qui dat alosmisquinos

translated into modern Spanish:

quien da a los mezquinos (=pobres).

It is not so much interesting that words are different, but rather that the grammatical structure differs. In Latin the dative case was used (-ibus) and there was no definite article in Latin. These early Castilian fragments, show us the use of definite articles (**the** in English, **el, la, los, las** in Spanish). Moreover the copyist used the preposition **a**, which precedes the noun, and the noun is without a case ending: a los misquinos (to the poor)[26].

A series of sound changes occurred, but from 16th century documents it can be concluded that in those days Spanish got its final form.

Compared to Portuguese, Catalan, French and Italian, Spanish has 5 vowels, *i,e,a, o,u*, only, while the other Roman languages also have closed vowels. This difference might have been caused by the geographical location of the Castilians, close to their neighbours the Basques. The center of the Basque homeland never was occupied by the Romans and the Basque language was not romanized. Contacts between Burgos in Old-Castilia and the Basques lead to the f to h change (*femina*, now *hembra* 'women', fabulare, now hablar 'to speak'). Similar changes in Arogonese and Gascoigne (French side of the Pyrenees) confirm the Basque influence on Old-Castilia.

The Spanish idiom was influenced by the Arab rules. Many words of Arab origin have the first syllable **al**, the Arab article, e.g.,

Spanish	English	Arab
alacena	cupboard	(al-)hazêna
alacrán	scorpion	(al-)^{c}aqrab
alarde	parade	(al-)^{c}ard
alazán	sorrel	(al-)'az^{c}ar
albaricoque	apricot	(al-)birqûq/barqûq

Some words even entered French *amiral* and further north in Dutch *admiraal* (Spanish *admiral*).

With the Reconquista the Arab occupation was gradually pushed back to southern Spain. The old capital Toledo was reconquered in 1058, and the kingdom Castilia and Aragon became the most important powers in Spain and during the Reconquista. In 1492 Granada was conquered, the last city of the Moors/Arabs. Castilian became the language of the administration and the literature.

From 16th century onwards Spanish developed into its current form. The Spanish rule and extension in the New World turned Castilian instead of one of the other Iberian dialects into the national language — Spanish. In all American varieties, the [θ] is missing, and the [s] and [x]are pronounced differently. This has been explained by Spanish migration to the New World. Most people of Spanish origin came from Andalusia and Extramadura. The differences in pronunciation are known from the Andalusian dialects.
Spanish is now spoken in Mexico, Cuba, Dominican Republic, El Salvador, Guatemala, Panama, Costa Rica, Filipina, Nicaragua, Venezuela, Columbia, Ecuador, Bolivia Peru, Chili, Argentina, Uruguay[27]. All these nations have their "Real Academia de Lengua Española" — their national academy for the Spanish language — and all these Academia unify the Spanish language within their possibilities. Of course this does not mean that the differences between the national language varieties are disappearing, but the Academia merely put their efforts into standardizing, limited by financial possibilities in the Latin Americas.

Portuguese

The Portuguese language is of the same origin as the Spanish language. The language has been developed from Old-Galician, one of the Iberic-Roman dialects in the north of Spain. Portugal became a nation in the 11th century under Henry of Bourgondy while Galicia was loyal to Castilia, and Castilians became the rulers of Spain. As Portugal had no obligations to the Spanish kings their language developed from the Old-Galician dialect. Typical for Portuguese are the ending -õa while Spanish has -ón or ión. In Portuguese the -n- en -l- in between vowels disappeared: Portuguese *mão*, Spanish *mano*. There also is quite a difference in pronunciation. Syllables without stress are swallowed. In spite of the common Roman origin the Spaniards can not understand Portuguese. Portuguese is the only Roman language which retained the early Christian days of the week: Monday is *segunda feira*, Tuesday is *tercia feira*, and Friday is *sexta feira*. The Latin letter f persisted: *fabulare* (Latin) became *falar* and not the Spanish *hablar* (to speak) and the Spanish *hacienda* (country estate) turned into *fazenda* in Portuguese[28].

After Portugal's contribution to the Reconquista (Liberation of the Mores), the Islamic Orient (Middle East) still was a barrier between Europe and the East Indies. The kings of Portugal supported the search for another route to reach the East Indies riches. The Portuguese *naos* or galleons discovered Madeira and the Azors, and left their language on the formerly unpopulated Islands. They were the first to round Cape of Good Hope. They discovered Brazil in 1500, a few years later they reached the shores of Japan, leaving the Japanese in bewilderment. The Portuguese brought the words *pagoda, mandarim(n), chá, a hot drink which became the English tee*. From Malaysia they brought lanchara (longboats) to Europe, from India *varandas* which were built in later days by the English to their bungalows in India. Outside Brazil the Portuguese language is not spoken extensively in other colonies.

After the independence of Brazil, Iberian Portuguese and Brazilian became two different varieties of Portuguese in pronunciation and in orthography. The differences are small and in Lisbon they seem unaware of these differences. In Brazil a *u* preceded by a *g* or *q* was written with a diaeresis *agüentar, tranqüilo*. This diaeresis was abolished in Portugal in 1945. The circumflex was also used differently. In Brazil one writes *cênico, anatômico, tônico*, but in Portugal *cénico, anatómico, tónico*[29]. In 1990 a new Orthographical Agreement[30] was entered between the Portuguese-language countries Brazil, Angola, Cape Verde, Guinea-Bissau, Mozambique, Portugal and São Tomé and Príncipe. This agreement was meant to make the orthographies of the Portuguese speaking countries more in accordance to each other. However, as far as possible. Due to differences in pronunciation, small differences in orthografy remained.

Brazil already accepted the Agreement, but it took up to May 2008 before the agreement was approved in the Portuguese Parliament. And from the 1st of January 2009 the Portuguese orthographical reform became valid in Brazil.

There was a lot opposition in Portugal and a "*Manifesto contra o Acordo Ortográfico*", a petition against the reform, was submitted to the President of the Assembly of the Republic. It got support from many well-respected linguists but whatever opposition exists, it is expected that Portugal will plan an introduction date some where in 2009.

Changes of and differences between the "*Acordo ortográfico da Lingua Portuguesa*".
For Brazilian Portuguese the diaeresis in lingüiça and freqüência disappears, as well as the acute accent in *idéia* and *européia*. A category of circumflexes disappears: *enjôo* becomes *enjoo*, *vôo* becomes *voo*. For Iberian Portuguese a lot of silent letters are skipped *acção* becomes *ação*, *accionar* becomes *acionar*. In some cases a word can have a silent letters in Brazil but not in Portugal. Therefore some former orthographies have been kept in Portugal: e.g. *acupun**c**tor, afe**c**ção, amí**g**dala, anticonce**p**cional*, etc. The Brazilian circumflexes as in *cênico, anatômico, tônico*, and the acute accents of Portugal as in *cénico, anatómico, tónico* are kept too.
The new rules also affect the way compounds are written: "*anta-sala*" becomes "*antassala*" and "*contra-regra*" becomes "*contrarregra*", "*microondes*" becomes "*micro-ondes*", the word "*fim-de-semana*" loses its hyphens and becomes "*fim de semana*". There is another unique change: the introduction of the letters k,K; w,W and y,Y in the Portuguese alphabet.

Nevertheless the two variants have become closer to each other than ever before. It will be worth studying the developments closely.

Provençal/Occitan
Provençal/Occitan is the medieval language of the south of France. As cultivated by the troubadours, it was one of the great literary languages of Europe. The invasion of the Germanic tribes are an important factor in development of Provençal/Occitan and French. This influence differs between the north and the south of France. Development went in the direction of standardization of language, i.e., French became dominant.

French
The name for France and for the French language was ironically derived from the name of the German tribe — the Franks, who moved into the regions of the old Gaul. How they took over the Latin-like language is unclear, but their own Germanic language was lost except for some proper names. This heritage is clear in loan words as *fauteuil* coming from *faldestol* (folding chair), and the German *w* became *g* or *gu* as in Guillome (William). Even the very French sounding words *garder, regarder* are derived (except for the Latin prefix re-) from the Old-Germanic language. Other military loanwords are *bannière* (banner), *briser* (to crush), *butin* (booty), *baudrier* (bandolier) and *haubert* (coat of

mail, hauberk). The Germanic idiom belonged to the social relations of the Franks. The feudal *Baron* is of German origin and even entered Latin as *Baro*[28].

French has been influenced by the Old-Norse of the Vikings, but these Vikings did not take women with them and the second generation Vikings had French speaking mothers. In modern French remnants still can be found: *havre* (harbo(u)r) in Le Havre, the quarters of the world *nord, sud, est, ouest* are originally Germanic words.

After the campaign against the Albigensian heretic sect in the 12th-13the centuries, French became the language of the Île-de-France, constituted from the dialects of Roman from Gaul, the Wallonians, the people of Lorraine, Picardy, Normandy, Bourguignon, Franche-Comté, Bourbonnais, Berrichon, Tourangeau, Angevin, Gallo, Poitevin and Saintongeais. This is the French, *la langue d'oïl*, as opposed to *la langue d'oc*, the dialect of the south[30]. The written language is from the 12th century, Modern French vocabulary, e.g., in the metric system, developed after the French Revolution.

During colonization French was exported outside of Europe, e.g., Louisiana, Haïti, Martinique, Maurice, Réunion, Canada (Quebec).

French is one of the official languages in Switzerland. It is spoken in the cantons Le Vaud, Neuchâtel and Genève. Belgium Walonia is French-speaking too.

In 1990 the Conseil supérieur de la langue française has proposed a change in orthography. This change has been presented as a recommendation. According to the new orthography trait d'union (circumflexes) are no longer used on the vowels i and u, e.g., maitre (new) versus maître in the previous orthography. For some plurals in the compound of a verb+noun only the nouns are plural *un cure-dent / des cure-dents*. Accents are more consistent with the general rules. However, the 1990 spelling has not been accepted by the French newspapers, but the Belgian (Wallonian) Catholic education system applies the spelling in all Belgium schools and so all children will forget the old function of the circumflex (maître, 1549-1740 maistre; hôpital, 1549-1718 hospital)[31].

Romanian

The Romanian language survived on an island surrounded by Hungarian and

[†] Dacia is an ancient country of SE Europe in what is now NW Romania. It was annexed by Roman emperor Trajan in 106 A.D. as a province of the Roman empire.

Slavic languages. It is spoken by 20 million Romanians, but how the Romanized language of the Dacians† survived is unclear. An explanation could be their isolation in the Transylvanian mountains.

In the early days Romanian was influenced by Church Slavic and it was written using the Cyrillic alphabet. It took until the 19th century before the Romanians could catalogue and analyse their language to give it, the Romanian, an identity. From this period the standard orthography was shaped after the example of the French language. Words which were necessary in modern times were borrowed from French: e.g., *garaj* and *timbre*.

Rhaeto-Romance
In the valleys of Graubünden, surrounded by the Alp Mountains, people who had lived here for centuries had their own language, Rhaeto-Romance, a Roman language which was named after the Roman province Rhaetia. The status of the language differs from the other Swiss languages, German, French and Italian, in that it is an official language in Graubünden only. Those who speak Rhaeto-Romance are, by necessity, at least bilingual.

3.1.3. **The Balto-Slavic languages**

Today's Baltic languages are Lithuanian and Latvian. Old Prussian was extinct. The Slavic languages are divided in the West, South and East Slavic languages. Sorbian‡, Polish, Slovak and Czech belong to the West Slavic branch. Slovene, Serbo-Croatian, Macedonian and Bulgarian belong to the South Slavic branch. Ukrainian, Byelorussian and Russian form the East Slavic branch of the Slavic languages. Old Church Slavonic has only survived in the Orthodox Church.

3.1.3.1. **The Baltic languages**

Lithuanian and Latvian (and Old Prussian) belong to the Baltic branch of the Balto-Slavic family of languages. The Indo-European ancestry can be seen from Lithuanian words as sūnus, duktė, motina, saulė, mėnuo (son, daughter, mother, sun, moon). Other Lithuanian words show the Balto-Slavic relation (Lith.) galva, (Latvian) galva, Russian golova, Polish głowa. Other words only are distinct between the Baltic and Slavic branch.[32,33,34] The difference between Latvian and Lithuanian is small, and consequently, a Lithuanian will understand Latvian and a Latvian will understand Lithuanian. Russians under-

‡ The Sorbian, the language of the Sorbs, a Slavic language related to Polish and Czech has been revived from near extinction and has around 70,000 speakers. It is also called Wendish or Lusatian.

stand basic Polish too. However, the difference between the Baltic languages and the neighbouring Slavic languages — Russian and Polish — is large and people can not understand each other's language,

Lithuanian
Lithuanian is spoken by roughly 3 million people in Lithuania, which is about 80% of the population (the other 45% of inhabitants are speaking Russian and Polish). Half a million Lithuanians are living abroad — primarily in the Chicago area in the USA.

Latvian
Latvian is spoken by roughly 1.5 million people in Latvia, which is 55% of the population. The remaining inhabitants speak Russian. The oldest texts in Latvian go back to the era of the Reformation; i.e., the sixteenth century. The Latvian language has undergone many secondary developments, while in Lithuanian many archaic forms have been preserved.

3.1.3.2. **The Slavic languages**

The West Slavic Branch — the Polish, Slovak and Czech language — use the Latin script. Whereas some Germanic languages only have 4 cases (nominative, accusative, genitive, dative), the Slavic languages also have vocative, locative, and instrumental cases.

Polish
Every letter in Polish apart from ch, cz, sz, d'z, d.z, rz, is pronounced separately. There are no silent letters. Nasal vowels are marked with an oganek, a mirror image of the cedilla to mark palatalized vowels. Words consist of long groups of consonants which are nearly unpronounceable for non-Polish speakers.

Slovak
The Slovak language is spoken by approximately five million of people in Slovakia, occupying the eastern area of the former Czechoslovakia federation. Slovak has been influenced by Latin, Hungarian, and Czech, but many of the archaic linguistic features have been preserved. There are Western, Central and Eastern dialects, but the standard literary language is based on Central dialects.

Czech

There are 5 vowels in the Czech languages. They can be written with or without accents (a, á, e, é, etc.). In Czech the y and ý are pronounced the same as i and í. The u ring (ů) is used in words as sůl (salt).
Consonants can be divided in labials (p, b, f, v, m), dentals (t, d, c, s, z, n, r, l), palatals (t', d', č, š, ž, ň, ř, j) and velars (k, g, ch, h).

The East Slavic Languages

The East Slavic languages consist of Byelorussian, Russian and Ukrainian. These languages are closely related and could also be considered dialects of one language. Byelorussian, Russian and Ukrainian use the Cyrillic alphabet, named after St. Cyril (Cyrillus) who lived 826-869 A.C.. His holy day in the Eastern Church is May 11. The Cyrillic alphabet was derived from the Greek uncial script, written in a majuscule script between the 4th and 8th centuries.

Russian

The Russian language is spoken in Russia, the largest republic of the Russian Federation. The name *Rus* and the Finnish word *Ruotsi* have a long history. Probably going back to the Swedish Vikings who braved the dangers of the rapids and narrows of the Dnepr to the Black Sea to trade with Byzantium. They founded Kiev. In Old-Russian texts words as *jakor* (anker) and *knut* (knout) exist, in Iceland still the old-Norse word *knutr*.
Russian developed from the dialect of the Principality of Moscow. The independency of Moscow and Kiev ended after the raids of the Mongols (1237). For two centuries Russia was under the brutal rule of the "Golden Horde". The Mongol yoke was lifted in the 15th century, their big catch was Moscow's grand duke's territory. The Turkish-Tartar language of the Mongols hardly influenced Russian: loshad (horse), ishak (easel).
In the 17th century considerable social changes took place in Russia. Slavery was introduced and from the administration and trade sectors a new language arose. In 1708 Peter the Great decreed a new alphabet that was in use up to the fall of the Romanovs (the revolution of 1917).

In many languages articles are used to denote the definite (the) or indefinite (a, an) attributes of a noun. However, there are no articles in the Russian language, and no change takes place in Russian nouns to indicate whether they are used defined or undefined. The particular meaning may be understood from the context. Russian may be described as a language of prefixes, inflexions and suffixes, which give Russians a tool to express their thoughts in all shades of meaning accurately. Many prefixes add meaning to verbs.

в-	inward movement
вс-	upward movement
вы-	outward movement
эа-	beginning, starting
на-	sense of quantity, sufficiency
о-, обо-	sense of movement around or about
от-, ото-	sense of moving away from
пере-	expresses repetition
по-	denotes an action completed
под-	motion under or towards
при-	sense of arriving
про-	through
раэ-, пас-	finality
у-	denotes losing sight of, disappearing

An example for upward movement is in the Russian word for "to boil up" вскипатб (**vs**k’i-patb), or вскрикиватб (**vs**kr’i-ki-vatb) to cry out. These examples show some of the carrier attributes in the Russian language, the upwards boiling of crying, of which the fundaments never could be formulated in English. Apart from the declensions, suffixes are used to denote augmentation and diminution. By adding particles as -ище or -ина nouns get the attribute of enormity or an unreasonable dimension, other particles reduce the object to a minute size in one’s imagination.

There are six cases in Russian:

Nominative, used for the subject of a sentence.
Genitive, showing possession
Dative, used for the indirect object of a sentence
Accusative, used for the object of a sentence
Instrumental, denoting means, instrument or agency
Prepositional, this case is always preceded by a preposition as v (in), ia (on).

Once the Russians discovered this richness and fine shades in their own language, Russian got a stable identity and new concepts of modern time could be included easily. A modern farmer drives a *kombain* or a *traktor*, a secretary uses a *mashinka* and everybody looks to the *televisor*.

Ukrainian and Byelorussian
The Ukrainian literature flourished in the 19th century. Compared to Russian Ukrainian has only a few Old Slavic forms, because it depended largely on oral tradition. Literature flourished in 19th century but in the preceding period Ukrain-

ian identity was suppressed partly by the czar, and partly by the Austrian monarchy. The Byelorussian feared the czar too, and their language suffered. Just at their independency in the beginning of the 20th century their language got its own identity too. There are small differences in Byelorussian and Ukrainian compared to Russian. Byelorussian and Ukrainian use additional Cyrillic characters.

The South Slavic Languages

Today's South Slavic languages are Serbo-Croatian, Slovene, Macedonian, and Bulgarian Old-Church-Slavic has been extincted. Servian, Macedonian and Bulgarian use the Cyrillic alphabet, while the Croats and Slovenes use the Latin alphabet. To meet the typical Slavic sounds diacritics have been added (Central European Latin).

Transliteration of the Cyrillic alphabet

There are many systems for transliterating Russian, the most important one is summarized in table 2.1. The conventions of capitalization in the Cyrillic original are much like those in French. Pronouns, days of week, months, and most proper adjectives are lower cased. Geographical names are capitalized, but references to institutions are lower-cased. Russian punctuation resembles French but no additional spaces are inserted for question marks or exclamations. Citations are put between guillemets (« and »). At some occasions quotations follow the German rules („ and “).

3.1.4. **Hellenic language**

Greek is one of the languages that influenced the European languages strongly. The early Greek alphabet originates from 800 B.C. This alphabet was the vehicle of the *Ilias* and the *Odyssey*, Homer's master piece.

Classical Greek has influenced all European languages. The English words pneuma (that which is breathed or blown), pseudocarp (false + Greek karpos 'fruit'), pseudocyesis (false + Greek kuēsis 'conception'), psyche (via Latin from Greek psukhē 'breath, life, soul'), pteropod (via Latin, from Greek pteron 'wing' + Greek pous, pod- 'foot'). The concept of tekhnē 'art, craft' + logia gives colour to our world of technology. The noun technic(s) (Am. E.) or technique (Br. E.) is derived from the Greek word tekhnēkos. The word atmosphere comes from atmos 'vapour' + sphaira 'ball, globe'.

Recently, a text of Archimedes has been recovered. As parchment was scarce at the time a Greek monk in the 12th century deleted the Greek text and used

Table 2.1 Cyrillic Alphabet and Transliteration[19]

А	а	a	П	п	p
Б	б	b	Р	р	r
В	в	v,w	С	с	s
Г	г	g	Т	т	t
Д	д	d	У	у	u
Е	е	e	Ф	ф	f
Ё	ё	jo, ë	Х	х	ch, h
Ж	ж	sch, ž	Ц	ц	z, c
З	з	z	Ч	ч	tsch, č
И	и	i	Ш	ш	sch, š
Й	й	j	Щ	щ	schtsch, šč
К	к	k	Ъ	ъ	"
Л	л	l	Ы	ы	y
М	м	m	Ь	ь	’
Н	н	n	Э	э	e, è
О	о	o	Ю	ю	ju, û
			Я	я	ja,â

the material for writing a prayers book. Modern technology (tekhnēlogia) has recovered the original Greek texts. The text proofed that Archimedes was using logical assumptions and reasoning to proof with certainty his geometric axiomata[35]. Archimedes was killed by a Roman soldier in 212 B.C., this was the final blow to Greek civilization and modern views on society. In the dark Middle Ages the texts of the Greeks were considered to be of heathen origin and even could be signs of the Antichrist. It was not the monks in the monasteries who kept classical Greek ideas, but the Arabs, who translated the Greek texts.

Greek is spoken by about 13 to 14 million people. It is the official language of the Republic of Greece and it is one of the two languages of the Republic of Cyprus. Standard Greek vocabulary has been derived from two different language sources: the demotic tradition and the *katharevousa* or learned tradition. The majority of demotic Greek words end with an open syllable, i.e. a vowel. The only consonants which may appear in word-final position are /n/ and /s/. In *katharevousa*, words may also have word-final consonant clusters such as /ks/ ’αυθραξ /’anθraks/ (coal).

Modern Greek uses the same twenty-four letters as Classical Greek. In addition a number of diacritics and punctuation marks are used. Diacritics are the diaeresis and the *accent* ('), punctuation marks are the *full stop* (.), *colon* (:), *exclamation mark* (!). Peculiar to Greek is the *raised point* (·), which is used instead of the semicolon. The Greek *question mark* (;) is identical to the Latin semicolon.

Numerals are presented in Latin or in Greek characters. If Greek characters are used the characters are appended with an acute accent. From 1000 the numeral symbols have the a low accent preceding the number. These character symbols came from Old-Greek, but this system had too few characters and therefore it was appended with three new symbols 6, 90 and 900 which came from Ancient Greek[36,37].

α′ β′ γ′	1 2 3
ι′ κ′ λ′	10 20 30
ρ′ σ′ τ′	100 200 300
͵α ͵β ͵γ	1000 2000 3000
ς′/στ′ ϙ′ ϡ′	6 (στίγμα) 90 (κόππα) 900 (σαμπί)

The symbol characters for 90 and 900 are not present in most scripts.

Transcription of Greek

The writing system of Classical Greek consists of an alphabet of 24 letters. Modern Greek uses the same letters plus a number of diacritics and a set of punctuation marks. The Greek alphabet is transliterated to the Latin script in table 2.2.

3.1.5. **The Finno-Ugric languages**

Due to their geographic position in the Baltics Estonian, Finnish and some minor languages (Karelian, Vespian, Votic) belong to the Balto-Finnic Branch of the Finno-Ugric language family. The most northern Finno-Ugric language is Sami, the language of the Laps (northern Finland and Finmark in Norway). Finnish is spoken by ca. 5 million inhabitants, and it is one of the official languages of Finland. Swedish is the other official language of Finland, spoken in the southern coastal areas. The Estonian language is spoken by ca. 1 million people in Estonia. Karelian, by ca. 86,000 people, Vespian and Votic — the Volgaic languages — have suffered from Sovjet suppression.

Hungarian is often assumed to be related the Finno-Ugric languages. It is the

Table 2.2 Greek Alphabet and Transliteration[19,36,37]

Name	Greek	Trans.	Name	Greek	Trans.
Alpha	α	a	Mu	μ	m
Beta	β	b	Nu	ν	n
Gamma	γ	g	Xi	ξ	x
	γγ	ng	Omicron	ο	o
	γκ	nk	Pi	π	p
	γχ	nch	Rho	ρ	r
Delta	δ	d	Sigma	σ,ς	s
Epsilon	ε	e	Tau	τ	t
Zeta	ζ	z	Upsilon	υ	y u
Eta	η	e ē	Phi	φ	ph f
Theta	θ	th	Chi	χ	ch
Iota	ι	i	Psi	ψ	ps
Kappa	κ	k	Omega	ω	o ō
Lambda	λ	l			
			Diphthongs		
	αυ	au			
	ευ	eu		ηυ	eu, ēu
	ου	ou		ωυ	ou, ōu

largest Uralic language spoken by about 14 million people. It has lost many of its Uralic characteristics, and has borrowed numerous words from Turkish and the European languages.

3.1.6. **The Basque languages**

One of Europe’s exotic minority languages, unrelated to any other language in the world, Basque is making its revival. It is spoken by a population known as the mystery people of Europe. Basque is spoken by two-thirds of a population of one million in the Basque Country (Euskara Herri). The official Spanish policy used to be highly unsympathetic to the Basque people. The language went underground during the Franco years (1939-75). Nowadays a Basque homeland government rules, publishing official documents in Basque. A number of newspapers are publishing in Basque.

The Basque language does not possess an alphabet of purely Basque origin, but rather borrowed one from the modern languages spoken in nearby areas. However, the basque language is very old and existed already in the Old Stone Age. All sorts of cutting and hacking tools, such as a hammer, an axe, and a knife, are derived from the lemma *aitz*, which means stone. While material objects are original words, general expressions and abstract ideas come from Latin. The word king (*errege*) comes from Latin *regem* (case ...). The Basque alphabet (agaka) is based on the Roman alphabet. There are twenty-two single letters plus seven compound letters (dd, ll, rr, ts, tt, tx, tz) There are words which start with tt (ttattar (tiny), ttintta (tiny drop), ttonttor (little bump), all related to smallness. Smallness is even seen in -tto, a diminutive suffix[38].

The Basque numeral system is based on the number twenty. There are names for 20, 40, 60, 80, these names literally mean 2 times 20 , 3 times 20 4 times 20[26]. This system preceded the decimal numeral system in Europe. A trace of it is still found in the French quatre-vinght. The Basque numerals

1 = bat	20 = (h)ogei
2 = bi(ga)	40 = berrogei
3 = (h)iru(r)	60 = irurogei
4 = lau(r)	80 = laurogei

3.2. **The Celtic languages**

The Celtic languages once had an important position being spread widely over Europe. The Celtic speaking tribes even threatened the Hellenic civilization when the Galatians in the fifth century B.C. were defeated by the Greek armies. Some of them went over the Bosporus and settled in Turkey. The Greek considered the "Keltoi", as Barbarian, those people who only could say "bar... bar... In 390 B.C. the Celts destroyed the ports of Rome which much *furore*, a military technique of making so much noise to scare and intimidate the enemy. But when the Romans learned to know the Celts they considered this type of intimidation as Celtic bluff. The apostle Paul addressed a letter to them (Epistle to the Galatians).

The Romans adopted Celtic words: carrus, a Celtic luggage wagon, became "car". The Celtic caballus "packhorse" became the French *cheval*, and the horseman *chevalier*.

All geographical names with "dun" or "briga" are Celtic names: Verdun, Bregenz (Austria), Brihuela (Spain), Lugdunum (Lyon, France). Lug was the name of a mighty Celtic God[40].

Today Celtic speaking people are found in Ireland, the very North-West of Scotland, Welsh and Breton which descend from the Cornish language. The name Welsh was deduced from the German word *wealas* which means stranger. There are more variants of names for strangers: the Norman settlers named the French "Valskr", the Slaves named the Romanians "Vlach", and even today the French speaking Belgians are called "Walloons". In German the word "Walachen" is a nick name for the Romanians[19] and in Dutch the name "Walachijer" a stranger, an inhabitant of Walachije exist[41](VD).

The Welsh language has lost the link with the people, in the 20th century a lot of the youngsters who spoke Welsh went to the cities and they neglected their home language. At last count the number of Welsh speaking people was down to half a million. Cornish (Kernow in Cornish) has vanished. The Bretons, who were driven from Cornwall by the Anglo-Saxon invaders in the 5th and 7th Century reintroduced the Celtic language in France. Today Breton is only spoken in the country side and in fishing villages by about 500,000 people. The area in which Irish or Gaelic is spoken is smaller too. This language is now spoken in the west of Ireland, Donegal, Connemara, Mayo, Galway, Kerry and Cork, at the outer edge of the island. The Scottish Gaelic, also known as Scoti, came from a tribe in Ulster. The Scottish Gaelic flourished in the 11th and 12th century. When the Saxon invaders came the Gaelic area was reduced to the Scottish Highlands. Today this language is spoken by only 40,000 people.

3.3. **The Middle Eastern languages**

3.3.1. **Arabic**

The Arabic language belongs to the Semitic family of languages. It is the most important language of the Islam, stretching from the Arab peninsula to Morocco in the West (Maghreb). Spoken Arabic is very different for different regions from the East to the West, Moroccan people do not understand Syrian Arabic. Intra-Arabic exchanges are hardly possible on the local language level. However, written Modern Standard Arabic, based on Classical Arabic, has become the standard way of communication throughout the Arab worlds, unfortunately only accessible for literate people. The Modern Pan-Arabic Standard language is more or less the unifying factor within religion, news and communications between nationalities throughout the Arab world.

Arabic words are derived from 3 or 4 letter roots, which consist of consonants, e.g. KTB (to write). Many words with connotated meaning are derived from these three letter roots.

kitāb, book
kutub, books
kātib, writer
naktubu, we write
yaktubūna, they write

The root MLK can be read as mālik, king. The definite article precedes the noun and no space is inserted: (’a)l-mālik, the king, just like: (’a)l-kitāb, the book.

A sentence as "read and write" becomes "(’i)qra’ wa-ktub", and the Arabic script requires an additional ’alif. On the other hand case endings usually are dropped in the consonantal script. Personal pronouns can be used as suffixes for nouns verbs and prepositions: qatala-nī is he killed me, written as one word.

A wide variety of related verbs exists, they are classified as I to XV, but the verb types II to VIII and X are most important. These classes indicate modifications of the essential meaning: *qatala* means " to kill", but *qattala* is to kill violently, to slaughter. In writings, books, newspapers etc, only consonants are written, however, newspapers frequently use the *shadda* character to indicate a difference in meaning.

All these variations result in a huge lexicon of 5 million words, and theorical considerations suggests that probably more forms might exist. In modern times some of these forms will get a new meaning related to the inherent meaning belonging to the word class.

Many neologisms have been introduced in modern Arabic, just because society needs new idiom to describe current day concepts. The way of building compounds sometimes conflicts with the very nature of building Arabic words, and is not always considered as leading to a beautiful Arabic language[42].

bar-mā’ī, amphibious
kahra-mignātīsī, electromagnetic
mā-faqa-l-banafsajī, ultraviolet

Other words in newspapers are simply phonetically transcribed to the Arabic script, such as personal names, geographical names, brand names, company acronyms, etc.

3.3.2. **Modern Hebrew**

Modern Hebrew or 'Ivrit' is the national spoken and written tongue of the Jewish majority in the state of Israel. It is the latest stage in 3000 years of Hebrew language evolution. Hebrew is a Semitic language based on root or radical words of three or four letters. This root shoresh שורש is a consonantal skeleton upon which various forms of words can be build. Each form has an inherent meaning. The words "wrote" and "dictated" have such an inherent relation in terms of meaning. Verb patterns are somewhat similar to strong verbs in English, write-wrote-written, freaze-froze-frozen. The Hebrew verbs undergo predictable internal changes, mostly in terms of vowels. There exists 7 verb pattern classes in Hebrew. Particles can be added in front or at the word's end[43]. The definite article "ha-" -ה, "ha-rosh" the head, is just one of them, other prefixes and suffixes are:
she-he-, she-ha-rosh that-the-head,
be-, be-yada, her hand
'i, kova'i, my hat
me-ha, me-ha-yeladim, the (my) parents
ve- 'and' is attached to the next noun:
ve-ha-kearot, and the smooth
prepositions be, ke, le, are attached to the noun too
be-ha-bet séfer, in the top (higest qualified) school
This system of modifying vowels, adding particles to the front and end of a word creates an enormous variety of related words. Its outcome is over 5 million different words, and further extensions are possible. Nouns, verbs, and adjectives are derived from a single root pattern, and within each class of words there is a lot of variety.

For spell checking there are two options:
a) an arbitrary word would be related to its root and thereafter the word should match the paradigm of a root. Such a technique would require online processing power and time. It is very likely that Hebrew principles of *conflation*[43] (be-ha-bots → ba-bots; in the mud) would be interpreted wrongly. These principles are unrelated to any indo-european principle.
b) the root paradigms should be unfolded to all possible forms and be present in the lexicons. Such a procedure requires large lexicons of ca. 5 million words, but access of all varieties is nearly instantaneous.

Finally, a language model arrives at linguistically related suggestions.

3.3.3. **Persian**

The Academy of the Persian Language and Literature has stated that the English name of the national language should be **Persian**. They reject any usage of the word "Farsi", which is an Arabic adaptation of the word "Parsi". Instead Persian (English)/Persa (Portuguese)/Persane (French)/Persisch (German/ Dutch) should be used in the Western languages.
Persian is the official language of of modern-day Iran. It is a linguistic continuation of Middle Persian, itself a successor of Old Persian (330 BCE). Persian is an Indo-European language related to the languages of Europe, but written in the Arabic script.

English	Persian
better	behtar
body	badan
dark	tarik
door	dar

In Persian no initial double consonants can exist. So *star* becomes *setāre*, and a*/bus-stop* becomes *bus-estop*.
Comparative and superlative adjectives are formed using suffixes:

large → larger → largest
bozorg → bozorg**tar** → bozorg**tarin**

As in most Indo-European languages a lot of particles can be added to words, and compounds also exist. When writing certain prefixes, suffixes and compound words the natural behavior of joining letters is overruled. A non-joiner character prevents the following letter from joining to prior characters (middle position letters).

incorrect	correct
خا نهها	خا نه‌ها

However word order differs from the word order in English. The sentences *dog eats cat*, becomes in Persian *dog cat eats*.

3.4. **Transliterations of non-European languages**

Hebrew
Hebrew is not an Indo-European language but belongs to the Semitic languages spoken in the Middle East and in the north of Africa. Historically it has played an important role in Western society. The Hebrew characters and their transcription are given in table 2.3.

Hebrew is read from right to left. Ancient Hebrew consists of consonants only.

Table 2.3 Hebrew Alphabet and Transliteration[2]
non-punctuated

Name	Hebrew	Trans.	Name	Hebrew	Trans.
Alef	א	’,a	Lamed	ל	l
Bet	ב	b	Mem	מ, ם	m
Geemel	ג	g	Noon	נ, ן	n
Dalet	ד	d	Samekh	ס	s
Heh	ה	h	’Ayeen	ע	’
Vav	ו	w,v	Peh	פ, ף	p
Zayeen	ז	z	Tsadee	צ, ץ	s,ts
Khet	ח	h	Koof	ק	q
Tet	ט	t	Resh	ר	r
Yod	י	y,e,i	Sheen	ש	sh
Kaf	כ, ך	k	Tav	ת	t

Later on some consonants indicated vowel length and quality. Most commonly the Vav ו was used to indicate ō and ū, the Yod י for é and ī, and the Khet for word-final ā, ē and e, much rarer is the Alef א for word-internal ā. Hebrew has 5 letters called Final Letters that take on a different shape for word endings (ך Khaf Sofeet, ם Mem Sofeet, ן Noon Sofeet, ף Feh Sofeet, ץ Tsadee Sofeet). For learning purposes punctuation marks have been added, but they are not applied in regular texts.

Hebrew numerals are letters preceded by apostrophes. They follow the decimal system, ’א 1, ’י 10, ’ק 100. There is no name for 500, it is represented by 400+100 ת”ק (Hebrew right to left!).

Arabic

Modern Arabic was derived from the Naskhi-script, one of the scripts of the early Islamic period. Like Hebrew it is a consonant script. Typically for Arabic is the position of the characters, having a specific form at the end of a word, in the middle of a word, at the beginning position of a word or as a single character not joined to other characters (see table 2.4). Arabic is a cursive script written from the right to the left.

The Arabic numerals are ٠ ١ ٢ ٣ ٤ ٥ ٦ ٧ ٨ ٩ (left to right 0,...,9). The question mark is reversed (؟). The Arabic numerals reached western Europe (replacing

Table 2.4 Arabic Alphabet and Transliteration

Indep.	midl	transl.	begin	end	name
ا	ـا	ā			alif
ب	ـبـ	b	بـ	ـب	beh
ص	ـصـ	ç	صـ	ـص	sad
ض	ـضـ	d			dad
د	ـد	d			dal
ذ	ـذ	ð			thal
ف	ـفـ	f	فـ	ـف	feh
غ	ـغـ	g	غـ	ـغ	ghain
ه	ـهـ	h	هـ	ـه	heh
ي	ـيـ	y	يـ	ـي	yeh
ج	ـجـ	j	جـ	ـج	jeem
ك	ـكـ	k	كـ	ـك	kaf
ل	ـلـ	l	لـ	ـل	lam
م	ـمـ	m	مـ	ـم	meem
ن	ـنـ	n	نـ	ـن	noon
ع	ـعـ	-	عـ	ـع	ain
ظ	ـظـ	ð	ظـ	ـظ	zah
ق	ـقـ	q	قـ	ـق	qaf
ر	ـر	r			reh
س	ـسـ	s	سـ	ـس	seen
ت	ـتـ	t	تـ	ـت	teh
و	ـو	w			waw
ث	ـثـ	^	ثـ	ـث	theh
ش	ـشـ	š	شـ	ـش	sheen
خ	ـخـ	x	خـ	ـخ	khah
ط	ـطـ	t	طـ	ـط	tah
ز	ـز	z			zain

Roman numerals) through Arabia by about A.D. 1200. These numerals probably have their origin in India.

It is recommended to use a system employing as few diacritics as possible. If the hamza (') and the ^{c}ayan (^{c}) are not available they can be replaced by an apostrophe and a superscript c (see Arabic influences on Spanish).

References to well known persons and places should use transcriptions that can be understood by English readers: Mecca, not Makkah; Faiyum, not Madinat al-Fayyum[39]. The Arab article should be joined to the noun with a hyphen: al-Islam, al-hazêna, al-'az^{c}ar.

3.5. **Epilogue**

Europa was the daughter of the Phoenician king Agenor. She ran away with Zeus who changed into a bull and brought her to Crete. Her name became the toponym for the Middle of Greece, subsequently for the Greek peninsula as a whole, and since 500 B.C. for the continent Europe. The identity of her name became even more powerful, it became Europe's unifying idea of the European Community, a wish to form a federation of all European states.

Europe has been the continent of antitheses, in the era of the Roman empire the Barbarians versus civilization, more recently the free world versus communism. The remains of these antitheses have a long history, and go back to the tribes who settled after a period of long wanderings in different parts of Europe. They exchanged pagan believe for Christianity but with different alphabets, the alphabet of the Latin Church and the alphabet of the Orthodox Church. These many tribes spoke a language which had a mutual ancestor. They diversified and many nationalities developed with their own national language or set of dialects.

Nowadays this diversity is still present. with the Germanic languages in the North, the Roman languages in the South and the Slavic languages in the East. The Celtic languages are nearly extinct. The Baltic languages have lost Old-Prussian. The Finno-Ugric family of languages is reduced to Finnish, Estonian and Hungarian, but the identity of Lapps and their language Sami is treated as an official language by the Norwegian and Finnish governments. It is still in use in the very North of Scandinavia. For some of the small languages a revival is ongoing, often related to a form of nationalism within regions such as the Basque homeland. Other languages have survived due to isolation (Icelandic & Faroese).

In the European Community the eleven member state languages have an equal status, and the multilingual nature of Europe — the old antitheses — is an accepted fact.

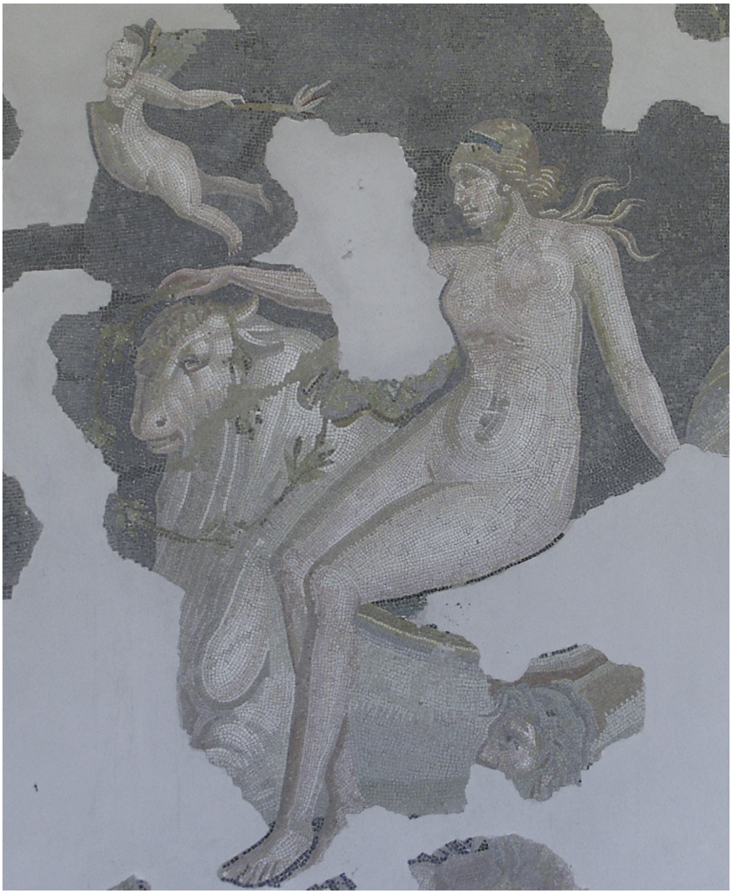


Fig. 3.6: The Roman version of the Abduction of Europa. Jupiter (the Roman Zeus), infatuated with Europa, the beautiful daughter of Agenore, the king of Sidon, changed himself into a bull and brought her to Crete. The Aquileia mosaic is traceable to the beginning of the 1st century B.C. and the theme was suitable to a marital bedroom floor and was well loved in the Roman Aquileia[44].

3.6. **References**

1 Robinson, A., Alfabet, Hiëroglief en pictogram (Alphabet, Hieroglyph and Pictogram, Trion, Baarn, 2001.

2 Webster's New World Hebrew Dictionary, Hayim Baltsan (ed.) Modon Publishing House, Tel Aviv, 1992.

3 The alphabet and the brain (eds. Derrick de Kerckhove, C.J.Lumsden) Springer-Verlag, Berlin, 1988.

4 Borghout, J.F., Egyptisch, Een inleiding in taal en schrift van het Middenrijk (Egyptian, an introduction in language and script of the Middle Egyptian), Peters, Leuven,1993.

5 Canadian Dictionary of the English Language, ITCP Nelson, Toronto, 1998, 1626-1627.

6 Mallory, J.P. & Mair, V.H., Ancient China and the Mustery of the Earliest Peoples from the West, the Tarim Mummies, Thames & Hudson Ltd., London, 2000.

7 Svenska Akademiens Ordlista över svenska språket, Norstedts Förlag, Stockholm, 1986.

8 Project Runiberg, *http://www.lysator.liu.se/runeberg/,* Nordic liturature on Internet, Laura Fitinghoff: Barnen ifrån Frostmofjället, 1907.

9 Bjones, J. & Dalene, H., Nynorsk ordliste for alle, Universitetsforlaget, Oslo, 1996.

10 Store rettskrivningsordbok, bokmål, Tanums, Kunnskapsforlaget, Oslo, 1996.

11 Nils Martin Hole, Kva heiter det (What's the name of that), bokmål nynorsk ordliste, Fagforlaget, Bergen-Sanviken, 1997.

12 Retskrivningsordbogen, 4. udgave, Dansk Sprognævn, Aschehoug, Copenhagen, 2012.

13 Webster's New Twentieth Century Dictionary Unabridged, second edition, New York, 1983.

14 Middelnederlandsch handwoordenboek, (J.Verdam, ed.), Nijhoff, 's-Gravenhage, 1964.

15 Ludo Permentier & Ludo Van Der Eynden, Stijlboek (Style guide), Scoop, Gent, 1997.

16 Grimm, J., Deutsche Grammatik,. Band I, 2. Ausgabe, Göttingen, 1822.

17 Duden, K., Vollständigen Orthographischen Wörterbuch der deutsche Sprache. Nach den neuen preußischen und bayrischen Regeln, Leipzig, 1880.

18 Brenner, Oskar, Die lautlichen und geschichtlichen Grundlagen unserer Rechtschreibung, Halle. 1902.

19 Die deutsche Rechtschreibung (Die neuen Regeln), Duden Band 1, Dudenverlag, Mannheim-Leipzig-Wien-Zürich, 1996, 2006, 2009, 2013.

20 Die neue deutsche Rechtschreibung, Bertelsmann Lexikon Verlag, München, 1996.

21 Zur Neuregelung der Deutschen Rechtschreibung, ab 1. August 2006, Mannheim, 2006.

22 Die deutsche Rechtschreibung, Duden Band 1 (ed. 24.), Dudenverlag, Mannheim-Leipzig-Wien-Zürich, 2006.

23 Beem,H, Resten van een taal (remains of a language), Van Gorcum &

Comp., Assen, 1967.

24 Beem,H, Jerosche, Jiddische spreekwoorden en zegswijzen uit hte Nederlandse taalgebied, Van Gorcum & Comp., Assen, 1970.

25 Gallinia, A.M., Grammatica della Lingua Catalana, Editorial Barcino, Barcelona,1969.

26 Schoten, J., Variaties en grenzen van het Spaans (Variations and boundaries of Spanish), Coutinho, Bussum, 1994.

27 Ortografía de la lengua Española, Edición por las Academias de la Lengua Espaõla, Real Academia Española, Madrid, 1999.

28 Atlas van de Europese Talen (Atlas of the European Languages), Het Spectrum Utrect/Antwerpen, 1984.

29 Prontuário Ortográfico e guia da língua portuguesa, Editorial Notícias, Lisboa, 1999.

30 Le Bon Usage, André Goosse (ed.), De Boeck Duculot, Paris/Louvain-la-Neuve, 1993.

31 Nina Catach, Dictionnaire historique de l'orthographe française (Dictionary of the historical frencch orthography), Larousse, Paris Cedex, 1995.

32 Mathiassen, T., A Short Grammar of Lithuanian, Slavica Publishers, Ohio, 1996.

33 Mathiassen, T., A Short Grammar of Latvian, Slavica Publishers, Ohio, 1997.

34 Eckert, R., Bukevičiūtė, E.-J., and Hinze, F., Die baltischen Sprachen, Eine Einführung, Langenscheidt, Leipzig/Berlin/München, 1998

35 Toebosch, Theo, Archimedes uit de schemer, NRC Handelsblad, Zaterdag 17 maart 2001.

36 Van Dijk-Wittop Koning, A.M., Levend Grieks, een kleine grammatica (Living Greek, a small grammar), Coutinho, Bussum, 1984.

37 Holton,D., Mackridge, P. & Philippaki-Warburton, I., Greek, a comprehensive grammar of the modern language, Routledge, London/New York, 1997.

38 Basque-English Dictionary, Gorka Aulestia, University of Nevada Press, Reno/Las Vegas, 1989.

39 The Chicago Manual of Style, 14th edition, University of Chicago Press, Chicago/London, 1993.

40 James S., Ontdek de wereld van de Kelten (Discover the worlds of the Celts), Areopagus, Vianen.

41 Groot Woordenboek der Nederlandse Taal, Van Dale Lexicografie, Utrecht/Antwerpen, 1995

42 Arie Schippers & Kees Versteegh, Het Arabisch, norm en realiteit (The Arabic language, norm and reality), Coutinho, Muiderberg, 1987

43 Lewis Glinert, Grammar of Modern Hebrew, Cambridge, University Press,

Cambridge, 1989.

44 Vidulli Torlo, Marzia, Aquileia Mosaici, Bruno Fachin Editore, Trieste, 2002.

CHAPTER 4

The Languages of South-East Asia

Bahasa Indonesia
Bahasa Melayu
Thai
Khmer

4. **The languages of South-East Asia**

4.1. **The Austronesian languages**

The Austronesian languages are spoken in Malaysia, Singapore, the Indonesian Archipelago and the Philippines. Malaysia, Singapore, Indonesia and the Philippines have standardized their languages. For Malaysia and Singapore this language is called "Bahasa Melayu" (Standard Malayan) and for Indonesia "Bahasa Indonesia" (Standard Indonesian). These languages are characterized by affix modifiers of words. Some are prefixes or suffixes separately attached to words, other affixes consist of both a prefix and a suffix. Infixes occur too, but are treated as a separate lemma in the dictionaries.

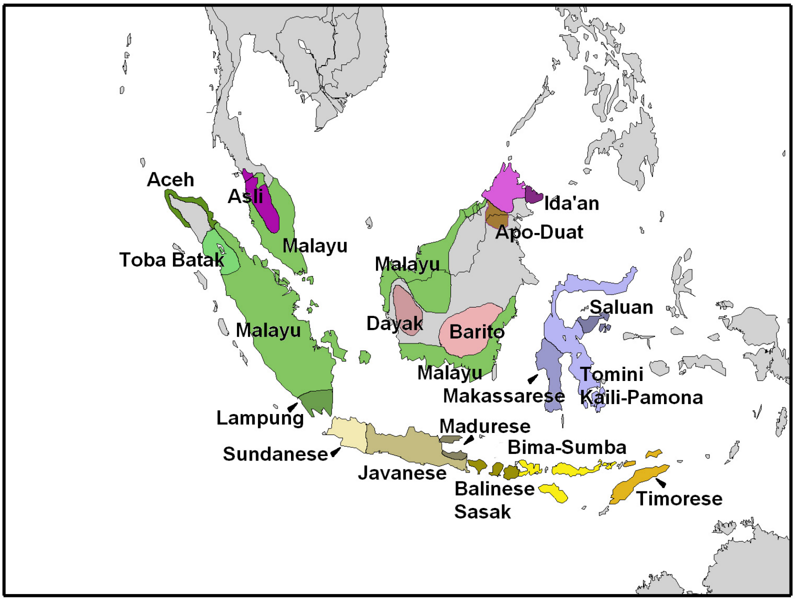


Fig. 4.1: The Austronesian languages spoken in Malaysia, Singapore and Indonesia. The lingua franca, the standardized languages Bahasa Melayu and Bahasa Indonesia, gain influence at the expense of local languages.
[2]Inspired by De Grote Taalatlas (The Atlas of Languages).

In Bahasa Indonesia affix modifiers vary in complexity (see the sample presented below):

-i	me-
ber-	me-....-i
ber-....-an	me-....-kan
keber-....-an	menge-...-kan
se-....-an	berse-...-an
menye-....-i	menye-...-kan

These prefixes are clearly existent in Bahasa Indonesia, but also in Bahasa Melayu, and in the local languages. In the Balinese language the Bahasa Indonesian word *bangun* (rise) is identical to the Balinese word, but the affixes attached to the word are written differently, e.g. *berbangun* (rising up) becomes *ba-bangun-an* in Balinese. Despite the different forms of affixes, the principles of the word modifiers are the same. In Tagalog, one of the main languages spoken in the Philippines, the -an affix modifies *ibigy* to *ibigy-an* (shall give to somebody to be named thereafter).

Another characteristic of the Austronesian languages is the reduplication of words.

Bahasa Indonesia
Nouns: kupu-kupu (butterflies),
Pronouns: saya-saya (the reduplication of I),
Adjectives: bagus-bagus (beautiful),
Verbs: duduk-duduk (sit about),
Numbers: satu-satu (one)

Affixes in reduplication encapsulate the reduplication: *besar* in *membesarkan* is reduplicated to *membesar-besarkan*. Affixation is applied to the unit having a distinct meaning. An artificial mechanism which would generate words as *membesarkan-membesarkan*. Such a mechanism would erroneously accept combinations that don't exist in real language. In other words non-sense would be approved. Some spellcheckers use these artificial mechanisms to camouflage their inability to build proper dictionaries.

The nature of affixing makes the Austronesian languages very distinct from the Indo-European languages and neighbouring Austroasian languages. The European languages French, Spanish, Italian, and Greek also use prefixes but typical of these languages are the modulation of the verb ending to express time in an ongoing stream of the syllables. Contrary to this ongoing stream of syllables the Austronesian affixes are invariable and additional affixes can be added to further change the semantics of a lemma. Usually lemmas are short; one or two syllables. These short lemmas expanded by affixes create a rather irregular structure of syllables.

4.1.1. **Hyphenation**

As discussed earlier, American hyphenator designs often use a theory that syllables are equally distributed throughout the word. This is valid for the French word *fu~tu~ro~lo~gue* but not not for the Bahasa Indonesian word *peng~a~dil-~an* (the court) or *se~per~ang~kat~an* (a complete couple). It is just on the boundaries of the affixes that the linear hyphenation model is applied falsely. The effect of falsely applying these concepts results in an increase of errors at the affix boundaries. There are a few ambiguities too: *meng~u~kur* or *me~ngu-~kur*, *ter~a~ngan(-a~ngan)* and *(ber~te~rang-)te~rang~an*, etc. The hyphenator does not hyphenate ambiguous syllables.

The *TALO model is not linear but starts with a language model that describes the characteristics of a language, in this case, the characteristics of the Austronesian languages, in particular Bahasa Indonesia and Bahasa Melayu. This model reduces the complexity by choosing linguistic units that belong to the languages themselves, instead of using incorrect assumptions about the language's nature.

Bahasa Indonesia and Bahasa Melayu are intimately related to each other. Governmental commissions have even accepted standards between the two languages (e.g. the conference with Brunei, Malaysia and Indonesia, 1991). They use the same orthography.

Therefore a unified hyphenator structure is feasible; just one hyphenator engine which is highly accurate, even usable for most of the local languages.

4.1.2. **Spelling**

For Bahasa Indonesia and Bahasa Melayu the linguistic structures are similar, but spelling focusses on the differences between the neighbours. The inner mechanisms of the speller engine do not need to be different, but the dictionaries are specific for each language. We have built a first series of dictionaries for Bahasa Indonesia and Bahasa Melayu that will confirm the orthography of the main Bahasa Indonesia and Bahasa Melayu dictionaries[3,4]. They will continue to expand as time goes by. Re-spelling capabilities finally will instantaneously know erroneously spelled words and apply correction automatically.

The Bahasa Indonesia and Bahasa Melayu orthographies have standardized spelling of European words[7]. Still English or Dutch orthography quietly enters documents: *expansive* instead of *ekspansif*, or *stabiel* instead of *stabil*. These are the cases to be corrected automatically. During the development of the lexicons this type of erroneous usage has been put into data bases. Spelling errors which cross the word boundary can be included too:

> *penanggungjawab* (should be *penanggung jawab*)
> *pertanggung jawaban* (should be *pertanggungjawaban*)
> *proklamasi republik Indonesia* (should be *Proklamasi Republik Indonesia*)
> *24,500 orang* (should be *24.500 orang*)

Spelling documents include punctuation checks too. Here, preferences between Basaha Indonesia and Bahasa Melayu may differ. The Republic of Indonesia has inherited many of the Dutch regulations while the Republic of Malaysia was strongly influenced by the British empire. Both nations use a different decimal system, respectively the European continental and the Anglo-Saxon system (Rp2**.**477**.**946,00 or Rp2,477,946**.**00).

Broadening our view we named the hyphenator "Euro Asia Hyphenator" and the speller "Euro Asia Speller". For Quark XPress it became the Hyphenator XT and the Speller XT, for Adobe's InDesign it became Smart Hyphen and Smart Speller.

4.1.3. **Acknowledgement**

We thank Mr. Lim Bun Chai of the Kompas Daily for the fruitful discussions and support which have been stimulating the development of the Bahasa Indonesia hyphenator.

4.2. **The Thai language**

The Thai language belongs to the Tai family of languages. All members of this group are located in South-East Asia. Lao, of Laos, and the Chang language of northern Burma also belong to the Tai family of languages.
Thai is spoken by about 40 million people. Thai has its own script, introduced by king Ramkamhaeng in 1283. The script has its origin in India. The Thai alphabet consists of more sounds than European languages. There are 44 consonants and most vowels are not represented by an individual letter but by a mark written above, below, before, or after a consonant, pretty much creating a syllable script[8,9,10]. The Thai language is a tone language with the diacritics of the script indicating middle, low, falling, high or rising tone marks. The script runs from left to right; majuscules don't exist nor do punctuation characters. However, the main difference between Thai and most other languages is the way sentences are written: that is WITHOUT spaces between the words, but a space is left where a comma would be written in English[11]. Together with compounding this is one of the major obstacles in printing.

4.2.1. **Compounding**

In the Thai language compounding is a principle to create new words having a different meaning than the meaning of their original components[10].

> "to understand เข้าใจ" is derived from "to enter เข้า" and "heart/spirit/mind ใจ"
> "train รถไฟ" is derived from "vehicle/car รถ" and "fire ไฟ"
> "electricity ไฟฟ้า" is derived from "fire ไฟ" and "sky ฟ้า"
> "lightning ฟ้าแลบ" is derived from "sky ฟ้า" and "pain แลบ"

The meaning of a compound always implies more than just the combination of the meanings of its components. A married couple for instance is more than a man plus a woman

4.2.2. **Sentences, Words, and Syllables**

A Thai sentence is a single unit, words are not separated from each other by blanks. To divide Thai sentences into words is not fundamentally different from hyphenating European words into syllables.

An English sentence written as a Thai-like sentence would appear as:

"theflowersofthefinestgreenhousesarenotwasted"

When this sentence is divided into individual words the context will be the decisive factor. A "green house" (i.e. a house painted green) is quite different from a "greenhouse" (i.e. a glass building for growing vegetables or flowers), but the context makes clear that this sentence is about greenhouses.
Other sentences could be rather conflicting: "Godisnowhere" could be divided into "God is nowhere" or "God is now here". This phenomenon is not very different from pecularities found in European languages, e.g. the English language: a "rec-ord" (i.e. report, document) or a "re-cord" (i.e. maximum achievement)!
The Thai people read words in context.

A last example of a transcribed Thai sentence and a word-by-word translation (actually the Thai write sentences without spaces)[12]:

mi:	sa:mi:	phanraja:	ramruaj	khu:	nung	maj	mi:	lu:k
is	husband	wife	rich	couple	one	not	have	child

If the word 'ramruaj' ร่ำรวย is divided into 'ram' ร่ำ and 'ruaj' รวย the word 'ram' could be placed at the end of the sentence:

mi: sa:mi: phanraja: ram

However, that would change the meaning of the sentence. 'Ram' means "to scent" (spraying perfume), so the meaning of this text line would be changed to: "is husband wife spraying perfume".
This is absolutely not allowed.

4.2.3. **A second layer for in-word hyphenation for newspapers**

Newspapers require narrow columns. Therefore newspapers do wish to hyphenate Thai words, but only at boundaries that can not be misinterpreted. The Thai Hyphenator consists of two layers, the first layer divides sentences into words, the second layer divides words into syllables. The division of words is a separate procedure distinguishable from the optional division in syllables of individual words.
The Thai word for "chairman", a compound with the *Sanskrit* prefix *pra*, can be divided as: ประ^ธานี.
The word date (day of the month or year) วันที่ can not be split into วัน and ที่, because the meaning would become day followed by 1) place, 2) in , those, 3) that, plus the other words in the sentence. Despite this limitation a high density of hyphenation can be realized thanks to *TALŌ's two-layer technology of the Thai

language model.

4.2.4. **Acknowledgement**

We thank Ms. Pariya Huntjes-Suwannaphoom for her successful efforts in building a learning corpus of the Thai language, we also thank her for the fruitful discussions and support which stimulated the development of the Thai sentence segmentation program.

4.3. **The Khmer language**

The Khmer language belongs to the Austro-Asiatic family of languages. Khmer is spoken by about 10 million people, and is the official language of Cambodia. Khmer differs from the neighbouring languages such as Thai, Lao and Vietnamese. It is NOT a tonal language.
The Khmer script has much in common with the Thai and Lao scripts. It has its origin in India too. The evolusion of the Khmer script goes back to 6th century. Current day Khmer consists of 35 consonants, 15 independent vowels, and a lot of vowel signs. A group of signs exists of two parts written around a consonant letter.

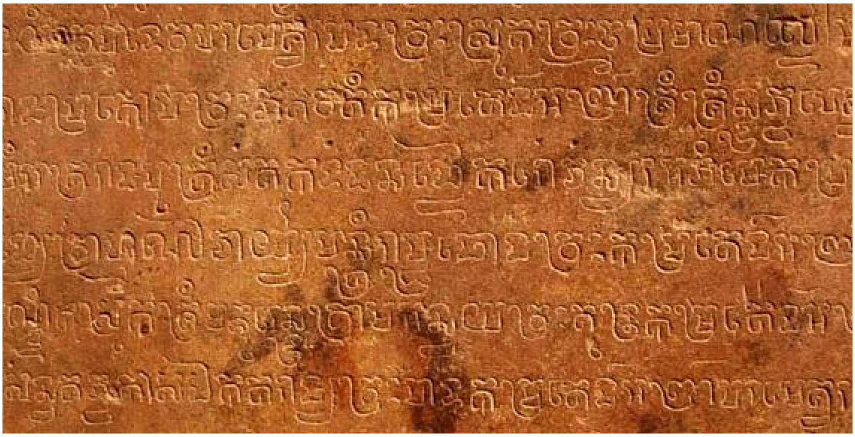
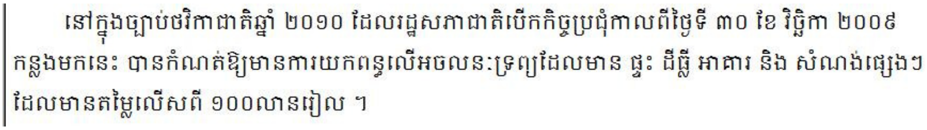


Fig. 4.2: A Khmer inscription, which shows words NOT separated by spaces.

The main difference between Khmer and most other languages is the way sentences are written: that is WITHOUT spaces between words, but a space is left where a comma would be used in English11. At the end of the sentence the Khmer sign Khan is used instead of a period. Together with compounding this is a major obstacle in printing.



នៅក្នុងច្បាប់ថវិកាជាតិឆ្នាំ ២០១០ ដែលរដ្ឋសភាជាតិបើកកិច្ចប្រជុំកាលពីថ្ងៃទី ៣០ ខែ វិច្ឆិកា ២០០៩
កន្លងមកនេះ បានកំណត់ឱ្យមានការយកពន្ធលើអចលនៈទ្រព្យដែលមាន ផ្ទះ ដីធ្លី អាគារ និង សំណង់ផ្សេងៗ
ដែលមានតម្លៃលើសពី ១០០លានរៀល ។

Fig. 4.3: A Khmer sentence broken by Khmer numbers representing commas.

4.3.1. Sentences, Compounds and the Unknown

Since there is no obvious boundary between words many words can be interpreted in different ways13. Basic words with elementary meanings can form a new word and therefore a sentences can be segmented in various ways. If there would be no context uncertainty would remain.
Another difficulty in segmentation comes from unknown words. Unknown words do have a context but this context is unknown to the segmentation program. Unknown words might also be error words, abbreviations, proper names, derived words, compounds, and numerals, but in all cases the sentence continues. If a word by word procedure would be used to step through a sentence we would lose the next word boundary. In other words the system would end in the middle of *nowhere* instead of giving the answer *now* | *here*.
Yet native speakers do recognize such a boundary instanteneously. A language model of the Khmer language helps to recognize such a boundary. It assumes a proces of learning, similar as learning did shape word recognition in native speakers. The pattern recognition mechanisms also are instanteneous and guaranty a fast response of the segmentation software program.

4.3.2. Segmentation and Spell Checking

If word boundaries are recognized correctly and if the next word boundary following the unknown is synchronized by the beginning of the next word (and its context), the nature of the unknown has to be solved by spell check procedures. These procedures are similar to those of other languages. Spell checking should be independent of segmentation.
If it is not a compound an unknown word can be automatically splitted in two words, or it can be exchanged by a different word. Two words in context can be jointed and exchanged by a single word too. These actions have nothing to do with segmentation. If necessary neighbouring content can be entered in the speller's history as a collocation pair (correct versus incorrect), similar to an English example: *In the middle of now here -> In the middle of nowhere.* The clue to a solution based on collocations is meaning.

4.4. **References**

1 Barber, C.C. Dictionary of Balinese-English (Vol. 1 & 2), Aberdeen University, Occasional Publications, Aberdeen, 1979.

2 De Grote Taalatlas (the atlas of the languages), Schuyt & Co, Haarlem, 1998.

3 Kamus besar Bahasa Indonesia, Departemen Pendidikan dan Kebudayaan, balai Pustaka, Jakarta, 1999 & 2001.

4 Kamus Malaysiana, Ensimal (m) Sdn. Bhd. Kuala Lumpur, Malaysia, Edisi Pertama, 1994.

5 Sneddon, J.N., Indonesian, a comprehensive grammar, Routledge, London and New York, 1996.

6 Teeuw. A., Indonesisch-Nederlands woordenboek, KITLV Publisher, Leiden, 1996.

7 KAMA/zz., A(C)-Pedoman Umum Istilah B.M.-May 2000(anum).

8 Woordenboek Thai-Nederlands, L.J.M. van Moergestel, Nangsue, Zaandam, 1995.

9 Woordenboek Nederlands-Thai, L.J.M. van Moergestel, Nangsue, Zaandam, 1995.

10 Thai-English Student's Dictionary, Mary R.Haas, Stanford University Press, Stanford, California, 1964.

11 Writing Systems of the World, Akira Nakanish, Charles E. Tuttle Company, Rutland, Vermont, Tokyo, 1980.

12 The structure of Thai Narrative. Somsonge Burusphat, the University of Texas, Arlington, 1991.

13 Word Bigram Vs Orthographic Syllable Bigram in Khmer Word Segmentation, Chea Sok Huor, et al, PAN L T, Cambodia.

GLOSSARY

5. Glossary

ablaut, the alternation in the vowels of related words, especially found in the Germanic strong verbs (e.g. in sing, sang, sung).

Afrikaans, is one of the youngest West Germanic languages but evolved independently. It was brought to the Cape by the Dutch Protestant settlers in the 17th century and is spoken by ca. 6 million people in the Republic of South Africa.

Albanian, the language of Albania and Serbia (Kosovo). It forms a separate branch on the Indo-European family tree.

Alþing, the Icelandic National Assembly, dating back to ca. 900 A.D.

Anatolian, an extinct group of languages constituting a branch of the Indo-European family such as Hittite, Luwian, Lydian, and Lycian once spoken in the western peninsula of Asia.

Analytic language, is a language tending not to alter the form of its words but to use word order to express gramatical structure. English, Dutch and Chinese are analytic languages.

Arabic, is a Semitic language, which was spoken during the Arabic rule of Spain and which influenced the Spanish idiom and the idiom of other languages. Knowledge of the Ancient Greeks was kept in Arabic translations of the Greek authors.

Arabic numeral, any of the numerals 0, 1, 2, 3, 4, 5, 6, 7, 8, and 9. The Arabic numerals reached Europe around 1200 A.D. replacing the Roman numerals.

Armenian, an Indo-European language written in a distinctive alphabet of thirty-eight letters. It belongs to the Indo-Iranian branch.

Austronesian, a branch of languages spread throughout Madagascar, South-East Asia, Indonesia and the Pacific Islands, also called Malayo-Polynesian.

Avestan, is the ancient Iryan language, closely related to Vedic Sanskrit.

Azerbaijanian, an Altaic language related to the Turkish language

Bahasa Indonesia, the standard language of the republic of Indonesia, closely related to Bahasa Melayu.

Bahasa Melayu, the standard language of Malaysia, closely related to Bahasa Indonesia.

Baltic languages, form a branch of the Balto-Slavic family of languages. The Baltic languages consist of Latvian, Lithuanian, and the extinct Prussian language.

Balto-Slavic languages, a family of languages consisting of Lithuanian, Latvian, and the Slavic languages, such as Russian and Polish.

Basque language, is an independent language of unknown origin. It is spoken in the northern provinces of Spain, around the gulf of Biscay.

bi- and triconsonantal, consisting of two or three consonants, e.g. represented in early Egyptian pictographic symbols.

Bokmål, see Norwegian, Bokmål.

Breton, is a Celtic language spoken in France in Brittany.

Byelorussian, one of the East Slavic languages, today the language of the republic of Byelorussia.

Canaanite, the language of an area of ancient Palestine west of the river Jordan, during the 2nd millennium B.C.

Catalan, is a Romance language related to Spanish and Provençal/Occitan, widely spoken in Catalonia, Andorra and in the southern part of France near the eastern Pyrenees.

Celtic languages, are a branch of the Indo-European family of languages, and include Irish, Scottish Gaelic, Welsh, Breton, manx, Cornish, and several extinct pre-Roman languages such as Gaulish.

consonant, a basic sound in speech in which the breath is at least partly obstructed and which can be combined with vowels to form a syllable. Consonants are the letters b, c, d, f, g, h, j, k, l, m, n, p, q, r, s, t, v, w, x, z.

Cyrillic alphabet, derived from Greek uncials, it is now used for Russian, Bulgarian, Macedonian, Serbian, Ukrainian, and Byelorussian, and some other languages which were influenced by the former Soviet

Union.

cuneiform, (adj. or noun) related to the wedge-shaped characters used in the ancient writing system of Mesopotamia, Persia, Elam, and Ugarit. The characters on clay tablets are called *cuneiform inscription*. Comes from Latin cuneiformis, derived from Latin cuneus 'wedge'.

Croatian, the counterpart of Serbian that uses a Latin script (see also Serbo-Croatian).

Czech, one of the west Slavic languages, the official language of the Czech Republic.

Dacians, the inhabitants of Dacia, an ancient country of SE Europe in what is now NW Romania. It was annexed by Trajan in AD 106 as a a province of the Roman Empire.

Danish, is a North Germanic language derived from the Old Norse, it is spoken in Denmark, and is the official language of Greenland and the Faroe Islands too.

declination, a tendency in certain languages or kind of utterance for pitch to fall near the end of a phrase, clause, or sentence (also named downdrift).

Diphthong, a sound formed by the combination of two vowels in a single syllable (as in *house* or *coin*).

Dutch, a West Germanic language, which is distinct from Low German. Old Dutch underwent Frankish and Ingwinian influences.

English, one of the West Germanic languages. It is the principal language of Great Britain, the US, Ireland, Australia, New Zealand and many other countries.

English, Old, The English of the Anglo-Saxons until about 1150 A.D. It is an inflected language with Germanic vocabulary. The 8th century epic poem Beowulf is written in Old English.

Estonian, is the official language of Estonia. It belongs to the Finno-Ugric family of languages and is closely related to Finnish.

etymological, is the relation of the origin of words and their historical development of meaning.

Euskal Herri, is the Basque homeland.

Faroese, the language of the Faroe Islands, closely related to Icelandic and the Old Norse.

Farsi, see Persian.

Finnish, the language of the Finns, spoken by 4.6 million people in Finland, and also in parts of Russia and Sweden. It is one of the Finno-Ugric languages, related to Estonian, Hungarian and several north central Asian languages.

Finno-Ugric family of languages, is a major group of Uralic languages, which includes, Finnish, Estonian, Hungarian, and several north central Asian languages.

Flemish, is a variant of Dutch spoken in western parts of Belgium.

Frank, a member of a Germanic people that conquerred Gaul in the 6th century and controlled much of western Europe for several centuries.

Frankish, of the Franksi or related to their language, see Franks.

French, the language of France, used in parts of Belgium, Switzerland, and Canada, in several countries of northern and western Africa and the Caribbean. It is one of the Latin languages which developed mainly from Vulgar Latin spoken in Gaul, but was also influenced by Celtic residuals and the language of the Germanic tribe of the Franks, who gave France their name.

Frisian, belongs to the Anglo-Frisian sub branch of the West Germanic languages. Modern Frisian is spoken in Frisia or Friesland, a province in the Netherlands.

Gaelic, is a Celtic language spoken in the western part of Ireland.

German, German is a West Germanic language belonging to the Indo-European family of languages. It is divided into Low and High German. High German is the current language of Germany, the official language of Austria and the German speaking parts of Switzerland.

German, High, is the official language of Germany. A typical difference between Low and High German is the ß as in *waßer* which is in Low

German *water*.

German, Low, a group of dialects spoken in the northern part of Germany. See also High German.

German, Swiss, is a variety or dialect of German spoken in northern parts of Switzerland

Germanic, East, Gothic is an extinct language spoken in the 4th to 6th centuries A.D.

Germanic, North, see the Scandinavian languages.

Germanic, West, the western branch of the Germanic languages consists of English, Frisian, Dutch, Flemish and German.

Gothic, is the extinct East Germanic language of the Ostrogoths, with earliest manuscripts ca. 4th and 6th century A.D.

Greek, ancient, the people or the language, ancient Greek was spoken in the Southern Balkan peninsula from the 2nd millennium B.C. The dialect of classical Athens formed the basis of the standard Greek dialect from 300 B.C.. It remained as a literary language during the Byzantine Empire and Turkish rule. It is the only representive of the Hellenic branch of the Indo-European family of languages.

Greek, modern, is the sole descendant of Ancient Greek and as such is a member of the Indo-European group of languages.

Germanic languages, those languages that developed from Proto-Germanic ca. 3rd millennium B.C., consisting of the East Germanic, North Germanic and West Germanic languages.

Hebrew, a member of the ancient people of Israel and Palestine, and the Semite language of this people since ca. 1300 B.C.

Hellenic languages, a branch of the Indo-European family, including the ancient Greek dialects.

Hebrew, modern, today's language of the state of Israel. It is derived from ancient Hebrew.

hieroglyph,a stylized picture of an object representing a word, a syllable, or sound, as found in the ancient Egyptian writing system. Comes from Greek hieros 'sacred' and gluphē the language of the Hittites, an ancient people who established an empire in Asia Minor and Syria that flourished between ca. 1700 to ca. 1200 B.C. The Hittite language belongs to the Indo-European family of languages.

Hungarian, Is the largest Uralic language, related to the Finno-Ugric family of languages. It has been influenced by Turkish and the European languages.

hyphenation, the process of dividing words in syllables to break words at the end of a line. Comes via Latin from Greek huphen 'together', hupo 'under' plus hen 'one'

Icelandic, the official language of Iceland, which has remained closely similar to Old Norse, due partly to the geographical isolation of Iceland and due to the policy of avoiding loanwords from other languages.

ideogram, a character symbolizing the meaning of something without indicating the sound used to pronounce it (e.g. the numerals and Chinese characters).

Indo-Iranian, a branch of Indo-European languages spoken in northern India and Iran. It is divided in an Indic group and an Iranian group.

Inflection, a change in the form of a word to express a grammatical function, such as tense, mood, person, number, case and gender (e.g. in French *la femme, les femmes*, in German *des Hauses*).

Ingwinian, The West Germanic coastal dialect, from Gaul up to Denmark. The term Ingwinians, the Germanic people at the "Ocean", came from the Roman writer *Tacitus*.

Irish, see Gaelic.

Italic languages, a branch of the Indo-European family of languages, that includes Oscan, Umbrian, and the Romance languages.

Kazakh, is a Turkic language related to Turkish and Azerbaijanian. Kazakh is written in the Cyrillic, Arabic and Latin script. Within 10 to 15 years the Latin script probably becomes the official script in the Republic of Kazakhstan.

Khmer, is an Austro-Asiatic language spoken in Cambodia. The Khmer script goes back to the 6th century.

Latin, the ancient language of the Roman Empire from which the Latin languages are derived.

Latvian, is one of the Baltic languages spoken in Latvia.

Lithuanian, is one of the Baltic languages spoken in Lithuania.

morphological, relating to forms of words, in particular inflected forms.

Norse, Old, the medieval language of Norway, Iceland, Denmark and Sweden, from which the modern Scandinavian languages were derived.

Norwegian, bokmål, one of the official languages of Norway, shaped during the Danish rule.

Norwegian, Nynorsk, the language constructed from the Norwegian dialects which has characteristics if the original Old Norse.

Nynorsk, see Norwegian Nynorsk.

orthography, The conventional spelling system of a language, how to use and combine letters to represent sounds and forms of words. Comes from Greek orthos 'correct' and graphia 'writing'.

Oscan, an extinct Italic language spoken in Italy in the 1st millennium B.C., before the emergence of Latin as standard language.

Ostrogoths, see Gothic.

palatalization, the process of making (a speech sound) palatal, especially by changing a velar to a palatal by moving the point of contact between the tongue and the palate further forward in the mouth.

palate, the roof of the mouth, from Latin palatium 'palace'.

Persian, the language of modern Iran. It is also called Farsi. Persian or Farsi is an Indo-European language and belongs to the Indo-Iranian branch. Earlier forms of this language were spoken in ancient or medieval Persia.

Phoenicia, an ancient country on the shores of the eastern Mediterranean, corresponding to modern Lebanon and the coastal plains of Syria. It flourished during the early part of the 1st millennium B.C.

Phoenician, the language or a member of the Semitic people inhabiting

Phoenicia and its colonies. Phoenician was written in an alphabet that was the ancestor of the Greek and Roman alphabets.

phoneme, any of the perceptually distinct units of sound in a specific language, e.g. the *b* and *p* in the English words *pat* and *bat*.

phonogram, the character or symbol used to represent a word, syllable, or phoneme, it can represent a succession of several letters such as ight in light, bright, etc.

phonetic(al), having a direct correspondence with symbols and sounds.

Pilipino, the national language of the Philippines since 1939; see also Tagalog

Polish, a west Slavic language

Portuguese, is a Romance language that derived from Latin and is now spoken by 160 million people in Portugal and Brazil.

Proto-Canaanite, a language or related to its writing, a lost language from which Canaanite was derived.

Proto-Indo-European, the lost language from which all Indo-European languages are derived.

Provençal/Occitan, is a medieval Romance language, spoken in the South of France.

quark, originally the sound the *seaswans* made in James Joyce's Finnegans Wake (1939), thereafter, any of a number of subatomic particles carrying a fractional electric charge. In the graphical industry it is used as the trademark Quark XPress; being the smallest carrier of transfer, the analogy between physics and language extents quarks to the smallest parts of meaning, quarks of meaning.

Rhaeto Romance, is the Roman language spoken in the Canton of Grisons by fewer than 30,000 people. There are several dialects, and it is an official language of Switzerland.

Romanian, is a Romance language, spoken in Romania by 23 million people. It has been influenced by the neighbouring Slavic languages.

Romanization, the historical process of bringing a region or people under

Roman influence or authority.

Roman numeral, any of the numerals I = 1, V = 5, X = 10, L = 50, C = 100, D = 500, M = 1,000. In this system a letter placed after another of greater value adds (thus XVII or xvii is 17).

Sámi, the language and people of northern Scandinavia, also called Lapps. The Sámi language belongs to the Finno-Ugric family of languages and is closely related to Finnish.

Scandinavian languages, the northern branch of the Germanic languages, comprising Danish, Norwegian, Swedish, Icelandic, and Faroese,

Scottish Gaelic, is a Celtic language spoken in the very north of Scotland.

Semitic languages, a family of languages that includes Hebrew, Arabic, Aramaic, and certain ancient languages such as Phoenician and Akkadian, constituting the main subgroup of the Afro-Asiatic family of languages situated in the Middle East and Northern Africa.

Serbian, the counterpart of Croatian that uses the Cyrillic alphabet (see also Serbo-Croatian).

Serbo-Croatian, the Southern Slavic language spoken in Serbia, Croatia and elsewhere in the former Yugoslavia. The Serbs use the Cyrillic alphabet, the Croats use the Roman alphabet.

Slavic languages, is branch of the Indo-European family of languages. It is divided in a Western, Eastern and Southern branch.

Slavic languages, East, is a Slavic branch of languages, consisting of Russian, Byelorussian and Ukrainian.

Slavic languages, South, is a Slavic branch of languages, consisting of Serbo-Croatian, Slovene, Macedonian and Bulgarian.

Slavic languages, West, is a Slavic branch of languages, consisting of Polish, Czech, Slovak, and the nearly extinct Sorbian.

Slavonic, the collective of the Slavic languages.

Spanish, is the official language of Spain, and most if the South American countries. It is a Romance language that developed after the collapse of the Roman Empire.

Sumeria, or Sumer the ancient region in Southwest Asia in present-day Iraq, comprising the southern part of Mesopotamia. From the 4th millennium B.C. it was the site of city states which became part of the ancient Babylonia.

Swedish, is a North Germanic language derived from the Old Norse. Swedish is spoken in Sweden and in parts of Finland.

Tagalog, one of the major languages in the Philipines, also locally called Pilipino. It is an Austronesian language.

Thai belongs to the Tai family of languages. It is a tone language and has its own alphabet. Sentences are written without spaces between the words.

Tocharian, extinct language, (Tocharian A and Tocharian B) spoken by the Germanic people who inhabited the Tarim Basin in the 1st millennium A.D. Their language shows a curious affinity to the Celtic and Italic languages.

toponym, a place name, especially one derived from a topographical feature. (from Greek *topos* 'place' plus *onuma* 'a name')

transliteration, a writing system of using the closest corresponding letters of a different alphabet.

Triphthong a sound formed by the combination of three vowels in a single syllable (as in French word *beau* or *foie-de-boeuf*).

Turkish, the official language of Turkey. It is an Altaic language with a rich agglutinating morphology and a rich case system. Turkish is written in the Latin script.

Ukrainian, an East Slavic language written in the Cyrillic alphabet.

Umbrian, an extinct Italic language spoken in Italy in the 1st millennium B.C., before the emergence of Latin as standard language.

umlaut, a mark used over a vowel, especially in German, to indicate a different vowel quality, such as in Land, Länder.

uncial, of or written in a majuscule script with rounded letters found in European manuscripts of the 4th-8th centuries and from which modern capital letters are derived.

uniconsonantal, consisting of only consonants.

Uralic languages, is a group of languages spoken from nothern Scandinavia to western Siberia, comprising the Finno-Ugric and Samoyedic branches.

velar, of, or relating to, the veil or velum (the soft palate), e.g. a velar sound.

Volgaic, referring to the River Volga and a group of Finnic languages such as Vespian and Votic.

vowel, a speech sound which is produced by comparatively open configuration of the vocal tract, with vibration of the vocal cords but without audible friction. It is a unit of the sound system of a language that forms the nucleus of the syllable. The letters representing such a sound such as a, e, i, o, u.

vowel gradation, another term for ablaut.

vowel harmony, a phenomenon found in certain languages such as Turkish and Finnish, that all vowels in a word are members of the same subclass.

Welsh, is a Celtic language.

Yiddish, is a High German dialect, spoken by many European Jews. The western branch is extinct.