Automated Bulgarian Hyphenation

Anton Zinoviev

21 October 2017

Contents

Principles of the Bulgarian hyphenation

One specificity of the Bulgarian language is that the average length of the words is greater than in English. When typesetting a Bulgarian text, hyphenation is more important than when typesetting an English text. Knuth's algorithm for line-breaking is such that in most English paragraphs no hyphenation will be used. With a Bulgarian text, however, even the Knuth's algorithm will use hyphenation in most paragraphs. Hyphenation becomes an absolute necessity if we want to obtain nice, justified paragraphs when using a software with dumb line-breaking algorithm, such as LibreOffice.

According to Decree 936 of the Council of Ministers promulgated on 27 November 1950, the Institute for Bulgarian Language at the Bulgarian Academy of Sciences is authorised to publish the rules of the orthography of the Bulgarian language (within certain limits).

Hyphenation rules between 1945 and 1983

Between 1945 and 1983 Bulgarian used syllable hyphenation with two morphological exceptions: hyphenation is preferred between a prefix and a stem and at the boundary of compound words. The following were the rules governing the hyphenation:

1. One letter does not stay alone. Words of one syllable can not be hyphenated.
2. No hyphenation before or after ь.
3. In a sequence of vowels at least one vowel stays before the hyphen.
4. A single consonant between two vowels links with the second vowel. For example по-ле /po-le/, ра-бо-та /ra-bo-ta/.
5. In a sequence of consonants between two vowels, at least one consonant stays with the second vowel. For example те-сто /te-sto/ or тес-то /tes-to/.[1]
6. In a sequence of consonants between two vowels, if the first consonant is sonorant (й /y/, л /l/, м /m/, н /n/, р /r/), then it stays with the first vowel. For example гер-дан /ger-dan/, сен-ки /sen-ki/.
7. The hyphenation separates two successive equal consonants. For example времен-но /vremen-no/, пролет-та /prolet-ta/.
8. When the letters дж /dzh/ and дз /dz/ denote a single consonant, then they are not separated. For example боя-джия /boya-dzhiya/ but not бояд-жия /boyad-zhiya/. When these letters denote two consonants, then the normal rules apply: над-живявам /nad-zhivyavam/.
9. Word prefixes may not be broken. Compound words are hyphenated either at the boundary of the components or the hyphenation rules are applied to each of the components separately. For example: пред-упреждавам /pred-uprezhdavam/ (not пре-дупреждавам /pre-duprezhdavam/), пред-известие /pred-izvestie/ (not пре-дизвестие /pre-dizvestie/), за-движвам /za-dvizhvam/ (not зад-вижвам /zad-vizhvam/), авто-клуб /avto-klub/ (not автok-луб /avtok-lub/), вакуум-апарат /vakuum-aparat/ (not вакуу-мапарат /vakuu-maparat/).

In some rare cases the proper application of rule 9 depends on the semantics of the word. For example пре-дреша /pre-dresha/ ‘change clothes’ but пред-реша /pred-resha/ ‘predetermine’ or прес-пите /pres-pite/ ‘the snow-drifts’ but пре-спите /pre-spite/ ‘sleep for a while/overnight’.

Hyphenation rules between 1983 and 2012

The Orthographic dictionary published by the Institute for Bulgarian language in 1983 introduced new hyphenation rules. The complexity of the previous rules

[1] In several publications this rule is formulated with the additional restriction that the sequence of consonants begins with an obstruent. I believe this restriction is unintentional. It makes no sense to forbid a hyphenation of the form AB-A but to permit ABB-A (A denotes a vowel and B – a consonant).

was the main reason for the change. The new rules aimed at two objectives: simplicity and unambiguity.

The new rules are:

1. A consonant between two vowels links with the second vowel. For example ви-со-чи-на /vi-so-chi-na/.
2. In a sequence of two or more consonants between two vowels, at least one consonant stays with first vowel and at least one with the second vowel. For example сес-тра /ses-tra/ and сест-ра /sest-ra/.
3. Two equal consonants are separated. For example плен-ник /plen-nik/.
4. In a sequence of two or more vowels, the first vowel stays before the hyphen. For example пре-одолея /pre-odoleya/ and прео-долея /preo-doleya/.
5. In a sequence of three or more vowels, the last vowel stays after the hyphen. For example мао-изъм /mao-izam/ but not маои-зъм /maoi-zam/.
6. The letter й /y/ between a vowel and a consonant stays with the vowel. For example май-ка /may-ka/.
7. When a sequence of two or more consonants follows й /y/ then at least one consonant links with й /y/. For example айс-берг /ays-berg/ (not ай-сберг /ay-sberg/).
8. The letter й /y/ between two vowels links with the second vowel. For example ма-йор /ma-yor/.
9. No hyphenation before or after ь.
10. When the letters дж /dzh/ denote a single consonant, then they are not separated. For example су-джук /su-dzhuk/ (not суд-жук /sud-zhuk/) but над-живея /nad-zhiveya/.
11. There must be at least one vowel before and after the hyphen.
12. One letter does not stay alone.

The total disregard of the morphology by these rules leads to some strange results. For example пре-дизвестие /pre-dizvestie/ is permitted and пред-известие /pred-izvestie/ is forbidden, зад-вижвам /zad-vizhvam/ is permitted and за-движвам /za-dvizhvam/ is forbidden, авток-луб /avtok-lub/ is permitted and авто-клуб /avto-klub/ is forbidden, вакуу-мапарат /vakuu-maparat/ is permitted and вакуум-апарат /vakuum-aparat/ is forbidden. Because of this, the new rules were not universally accepted. The old rules are still mentioned in various places in Internet, they are included even in some grammar books published by the publishing houses of the Ministry of Education and of Sofia University. The software developers, however, soon came into love with the new hyphenation rules.

Hyphenation rules after 2012

In 2012 new rules came into force. There are two differences with respect to the previous rules:

1. Rule 5 of the previous rules is revoked. For example маои-зъм /maoi-zam/ becomes a valid hyphenation.
2. The new rules permit morphologically based hyphenation (however it is not obligatory). For example пред-известие /pred-izvestie/, за-движвам /za-dvizhvam/, авто-клуб /avto-klub/, вакуум-апарат /vakuum-aparat/ are valid hyphenations.

Good hyphenation is a complex matter and it seems the linguists at the Institute for Bulgarian Language have recognised this. They no longer attempt to provide universal rules about everything. Instead, they provide some very permissible rules while the good application of these rules is leaved to the discretion and the experience of the printers and the developers of hyphenation software.

It makes sense to use at least two different sets of hyphenation rules for Bulgarian. In most cases a more restrictive version should be used, one which attempts to eliminate the controversial cases of hyphenation. When typesetting a Bulgarian text in a narrow newspaper column, however, it will be appropriate to use more liberal hyphenation rules. It should be noted that one of the reasons for the hyphenation reform in 1983 was the desire to fix the chaotic hyphenation in the Bulgarian newspapers at that time.

Computer implementations

Mathematical analysis of the Bulgarian hyphenation

The earliest mathematical analysis of the Bulgarian hyphenation rules belongs to Veska Noncheva.[2] In 1988 she proposed a mathematical formalisation of the hyphenation rules in a table with 22 rows.[3]

In the same year Eugene Belogay[4] proposed an alternative formalisation with only 9 rules.[5] Belogay proved that his rules are consistent and that they form a minimal set. The rules of Belogay have negative character – every hyphenation which is not forbidden by a rule is possible hyphenation.

The following are the first 7 rules, as formulated by Belogay:

1. Б-А
2. А-ББ
3. Б-ТТ, ТТ-Б
4. ААА-Б

[2] http://www.researchgate.net/profile/Veska_Noncheva

[3] Нончева В. Алгоритъм за автоматично пренасяне на думи в българския език. Математика и математическо образование. Сб. доклади на 17. ПК на СМБ. С., БАН, 1988, 479-482.

[4] http://www.linkedin.com/in/belogay

[5] Белогай Е. Алгоритъм за автоматично пренасяне на думи. Компютър за вас (1988) 3, 12-14.

5. й-ББ
6. Б-ь
7. д-ж

Here А denotes an arbitrary vowel letter, Б denotes an arbitrary consonant letter (including ь and й), ТТ denotes a sequence of two equal consonant letters and the letters й, ь, д and ж denote themselves. For example the rule "Б-А" says that we are not permitted to separate a consonant letter from immediately following vowel letter.

The eighth rule of Belogay says that hyphenation is forbidden before the first and after the last vowel letter. The ninth rule of Belogay says that hyphenation is forbidden immediately after the first or immediately before the last letter of the word.

Notice that is is very easy to translate the rules of Belogay in the form, required for the hyphenation algorithm of Knuth and Liang used in TeX.[6] Let us remind that this algorithm matches the word with a set of string patterns in which the odd numbers say hyphenation is permitted in this position and even numbers say the hyphenation is forbidden. When two patterns give conflicting numbers for the same position, then the greater number wins.

First, since the rules of Belogay are negative (they say where hyphenation is forbidden, not where it is permitted), we have to permit the hyphenation everywhere:

1. А1
2. Б1

Then, the first seven rules of Belogay obtain the form:

1. Б2А
2. А2ББ
3. Б2ТТ ТТ2Б
4. ААА2Б
5. й2ББ
6. Б2ь
7. д2ж

Since no Bulgarian word starts with more that four consonants and no Bulgarian word ends with more than three consonants, the eighth rule of Belogay can be translated in the following way:

1. .Б2
2. .ББ2
3. .ББББ2
4. 2Б.
5. 2ББ.

[6] Liang, Franklin Mark. Word Hy-phen-a-tion by Com-put-er (Doctoral Dissertation). Stanford University, 1983

The ninth rule of Belogay means that left and right hyphen mins should be set to 2.

The work of Eugene Belogay was not limited to merely a mathematical analysis of the Bulgarian hyphenation rules. In his paper he published a short algorithm in Pascal which implements these rules. It didn't take long for this algorithm to be used in various text processing software. The algorithm of Belogay was famous for many years. Even as late as 1997 in one book about TeX, the author didn't care to give any explanations but simply wrote about "the algorithm of Belogay" as something well known to the reader.[7]

Bulgarian hyphenation in TeX

One unfortunate design decision of Knuth was that the hyphenation algorithm of TeX applied the hyphenation patterns not to the input character codes but to the internal codes of the glyphs in the font. This created a problem for the Cyrillic languages because in TeX the Cyrillic fonts did not have standardised encoding. Perhaps this is one of the reasons why the earliest implementations of the Bulgarian hyphenation in TeX did not rely on the internal hyphenation algorithm of TeX. Instead, external tools were used to insert soft hyphens in all Bulgarian words. For example such a tool would replace the word сричкопренасяне /srichkoprenasyane/ with срич\-коп\-пе\-на\-ся\-не /srich\-kop\-re\-na\-sya\-ne/. The saying "To every disadvantage there is a corresponding advantage" is true – since Cyrillic and Latin letters use different character codes, an external tool could easily insert soft hyphens in all Bulgarian words while leaving the TeX commands intact.

The earliest known attempt to use the hyphenation algorithm of TeX for Bulgarian was made by Ognyan Tonev in 1990.[8] He described his work as "a not very good translation of the rules. I work in this direction. But I don't have a 100% working complect of patterns. So, the copy I send to you[9] is only a beta-version." The hyphenation patterns of Tonev don't work correctly and it seems he never completed his work.

The first usable Bulgarian hyphenation patterns for TeX were developed by Georgi Boshnakov[10] in 1994. In order to solve the encoding problem, Boshnakov had developed TeX fonts supporting the MIK encoding (the prevalent encoding at that time in Bulgaria). This allowed him to introduce a fully working implementation only a few months after LaTeX2e became the official LaTeX version. Later Boshnakov modified his work with the Babel system. The

[7] Василев В. Ултимативният TeX. Удоволствието да правим предпечатна подготовка сами. София, Интела, 1997, 36

[8] The author of this text was unable to find current information about Ognyan Tonev in Internet. Apparently in 1990 he worked in the Center of Informatics and Computer Technology of the Bulgarian Academy of Sciences.

[9] To Yannis Haralambous, http://perso.telecom-bretagne.eu/yannisharalambous

[10] http://www.maths.manchester.ac.uk/~gb/

hyphenation patterns of Boshnakov did their job well enough, so that for almost quarter a century after their initial creation, they remained the only Bulgarian hyphenation patterns in the standard distributions of TeX and CTAN.

There are some similarities between the patterns of Boshnakov and the patterns of Belogay. The following are the main differences.

First, Boshnakov used an ingenious and more compact implementation of the second and the third rule. Instead of {А2ББ, Б2ТТ, ТТ2Б}, or 8×22×22+22×22+22×22=4840 patterns in total, Boshnakov has patterns of the form 2Б3Б2 and 4Т3Т4, or only 22×22=484 in total, with the same effect.

The second main difference between the patterns of Boshnakov and the patterns of Belogay concerns the letter combination дж /dzh/. In Bulgarian this letter combination can denote either a single consonant, or a sequence of two consonants and the hyphenation rules change respectively. Unfortunately, it is impossible to know the meaning of дж /dzh/ without a vocabulary. The solution of Belogay was a cautious one – his rules do the hyphenation in a way which will be correct regardless of whether дж /dzh/ is a single consonant or a sequence of two consonant. On the other hand, the approach of Boshnakov is a bold one – since дж /dzh/ is more often a single consonant, his rules assume that it is always a single consonant. The number of the cases when this decision leads to bad hyphenations is insignificant in comparison with the cases in which we obtain improved hyphenation.

The third main difference between the patterns of Boshnakov and the patterns of Belogay concerns the eighth rule – its implementation in the rules of Boshnakov is rather limited which leads to wrong hyphenations like бри-дж /bri-dzh/. A full implementation of this rule would require 11660 patterns in total and this would be too much for the computers in 1994.

Later developments

In 1995 Atanas Topalov defended a Masters thesis in the Faculty of Mathematics and Informatics at Sofia University titled “Algorithms and software about text processing”.[11] One of the main topics in his thesis was the Bulgarian hyphenation. Topalov criticised vehemently the official hyphenation rules and their total disregard of the morphology. He wrote:

> If we look at the history of the problems of the hyphenation, we will discover something very strange. Instead of the expected involvement with the depths and aspiration for more admissible and satisfactory style, we can find a growing tendency for simplification. One unpleasant discovery is that the development of the hyphenation software stays firmly on the principle “let us do the easiest thing”.

[11] The thesis of Atanas Topalov can be accessed at the author’s website http://www.mind-print.com

> The earliest works which have been studied are from 1978. It turned out that they present the best approach concerning the automated hyphenation. The authors have chosen the most difficult but the most correct (from literary point of view) method for hyphenation, namely the morphological approach.

Topalov proposed his own hyphenation algorithm. The hyphenation it generated was smooth and easy to read. One obvious defect of the algorithm of Topalov was that it contradicted the official hyphenation rules at that time. One can argue, however, that his algorithm is compatible with the current hyphenation rules.

In 1999 Svetla Koeva[12] wrote a paper about the automated Bulgarian hyphenation.[13] At that time she was a junior member of the Department of Computational Linguistics at the Institute for Bulgarian Language but now she is a director of the whole institute. The paper of Koeva contains a list of hyphenation patterns which can be used as a basis of automated hyphenation. In 2004 with the help of Stoyan Mihov[14] the rules of Koeva were formalised with regular relations and rewriting rules. They were implemented in a software product named ItaEst which provided Bulgarian hyphenation and grammar checking for various software products of Microsoft and Apple.

The main differences between the hyphenation of Koeva and the official hyphenation rules effective after 2012 is that the separation of a long sequence of consonants between two vowels is done according to the rules valid before 1983. For example се-стра /se-stra/ and ай-сберг /ay-sberg/ are permitted. The main difference between the hyphenation of Koeva and the official hyphenation rules effective before 1983 is that the rules of Koeva disregard the morphology of the words. The following rule of Koeva is specific: in a sequence of two sonorant consonants between two vowels, we are permitted to separate the first vowel from the first consonant, for example материа-лна /materia-lna/.

In 2000 Anton Zinoviev[15] created new hyphenation patterns for TeX. He didn't know about the previous work of Boshnakov and he didn't bother to make his work available in the various TeX distributions and CTAN. His work was used mostly by the local Linux enthusiasts and the colleagues of Zinoviev. In 2001 Radostin Radnev[16] created a free grammar dictionary of Bulgarian[17] where he used the hyphenation patterns of Zinoviev. From there the work of Zinoviev propagated to OpenOffice, LibreOffice and various online dictionaries, including http://bg.wiktionary.org and http://rechnik.chitanka.info.

The following are the main differences between the hyphenation of Zinoviev and

[12] http://dcl.bas.bg/svetla_koeva/

[13] Коева, Светла. Правила за пренасяне на части от думите на нов ред. Български език. 1999/2000, 1, 84-86

[14] http://lml.bas.bg/~stoyan/

[15] The author of this text.

[16] http://bg.linkedin.com/in/radostinradnev

[17] http://bgoffice.sourceforge.net/

the hyphenation of Boshnakov.

First, the eighth rule of Belogay is fully implemented.

Second, the rules of Zinoviev try to detect when the letters дж /dzh/ (and дз /dz/) denote a single consonant and when they denote a sequence of two consonants. By default, however, Zinoviev (like Boshnakov) assumes that дж /dzh/ is a single consonant and hyphenates accordingly.

Third, the rules of Zinoviev disable some cases of unpleasant hyphenations:

1. In a consonant sequence like тст /tst/, the two equal consonants т /t/ are separated. For example братст-во /bratst-vo/ is forbidden while братс-тво /brats-tvo/ and брат-ство /brat-stvo/ are permitted.
2. The hyphenation is forbidden after a sonorant consonant following an obstruent consonant. For example отм-ра /otm-ra/ is forbidden and от-мра /ot-mra/ is permitted.
3. The hyphenation separates two consecutive kindred voiced/voiceless consonants. For example субп-родукт /subp-roduct/ is forbidden and суб-продукт /sub-product/ is permitted.

At the start of his work on the Bulgarian hyphenation, Zinoviev had the opportunity to discuss the hyphenation with Svetla Koeva. He remembers that some cases of unpleasant hyphenation were suggested to him by Koeva. Unfortunately, he hasn't taken notes so now he doesn't know which cases of unpleasant hyphenation have been suggested to him by Koeva and which are his own findings.

The present work

Motivation

The present work was carried out on the initiative of the leader of the Bulgarian localisation team of Mozilla, who contacted Zinoviev, Boshnakov and the maintainers of the TeX hyphenation patterns.[18] This work pursues the following main objectives:

1. to update the hyphenation patterns in accordance with the current hyphenation rules;
2. to generate the hyphenation patterns by a publicly available script;
3. to make the hyphenation patterns customisable;
4. to provide documentation for the future developers.

The current official hyphenating rules for Bulgarian are rather liberal. Very often, in a long sequence of consonants we are permitted to split the word at any position, for example аген-т-с-т-во /agen-t-s-t-vo/. This is prone to many unusual and unexpected results that interrupt the attention of the reader or

[18] http://hyphenation.org

deceive his expectations during the movement of his eyes to the next line. On the other hand, in order to produce nice justified paragraphs there is no need for so many hyphenation possibilities. It would be sufficient even if only one possible separation between any two syllables was permitted.

Therefore, it makes sense to use a more restrictive version of the Bulgarian hyphenation, one which eliminates the controversial cases of hyphenation. Only when typesetting a Bulgarian text in a very narrow newspaper column it will be appropriate to use a more liberal version. It should be noted that some specialised English dictionaries also separate the word-division positions into two categories – preferred positions and less recommended positions.

There are two methods to determine the optimal division within a sequence of consonants between two vowels:

- we can hyphenate according to the syllables in the word or
- we can hyphenate morphologically.

Hyphenation according to the syllables in the word

Let us look at the properties of the Bulgarian syllables. All syllables have the following structure:

onset - nucleus - code

The nucleus in Bulgarian is always a vowel. Both the onset and the code are (possibly empty) sequences of consonants.

The Bulgarian syllables adhere to the Sonority Sequencing Principle. According to this principle, the consonants within the onset have raising sonority and the consonants within the code have decreasing sonority.

Several grammar books agree that the following sonority scale is valid for Bulgarian:

voiceless obtrusive < voiced obtrusive < sonorant consonant < vowel

According to the investigations of the author, the only exception to this law is due to the letter в /v/ which is a voiced obtrusive but it can be used also as a voiceless obtrusive. This exception is due to a spelling particularity of the Bulgarian language. Whenever the letter в /v/ seemingly violates the Sonority Sequencing Principle, in the spoken language this letter is read as ф /f/, that is as a voiceless obtrusive (for example the word отвсякъде /otvsyakade/ is read as отфсякъде /otfsyakade/).[19]

The author has found that the sonorant consonants in Bulgarian have their own sonority scale:

[19] No Primitive Slavonic word contains the phoneme ф /f/. Therefore, we can safely assume that in the Primitive Slavonic language the consonant ф /f/ was a positional variant of the consonant в /v/.

м /m/ < н /n/ < л /l/ < р /r/ < й /y/

Only a few words such as жанр /zhanr/ and химн /himn/ violate this scale. Such words are always loan-words and their pronunciation is somewhat problematic for the native Bulgarian speakers.

In addition to the Sonority Sequencing Principle, the consonant clusters within the Bulgarian syllable adhere to the following additional principles:

1. Both in the onset and in the code, the labial and dorsal plosives precede the coronal plosives and affricates.
2. If the onset or the code contains two plosives or affricates, then there are no fricatives between them. Few words with the Latin root 'text' are exceptions: контекст /kontekst/.
3. If the onset or the code contains two fricatives other than в /v/, then there are no plosives or affricates between them.
4. If the onset or the code contains two plosives or affricates, then they both have equal sonority (both are voiced, or both are voiceless).
5. If the onset or the code contains two fricatives other than в /v/, then they both have equal sonority (both are voiced, or both are voiceless).
6. Neither the onset, nor the code may contain two labial plosives, or two coronal plosives or affricates or two dorsal plosives.
7. Neither the onset, nor the code may contain two equal consonants with the exception of в /v/ (for example втвърди /vtvardi/).[20]

From all these properties of the Bulgarian syllable we can deduce the following hyphenation rules:

1. In a sequence MK where M is a consonant with higher sonority than K, we are not permitted to hyphenate before M. Exception: when M is в /v/ and K is a voiceless consonant.
2. In a sequence KM where M is a consonant with higher sonority than K, we are not permitted to hyphenate after M.
3. In a sequence KBT where K and T are plosives or affricates and B is fricative, we separate K from T.
4. In a sequence CKB where K is a plosive or affricate and C and B are fricatives other than в /v/, we separate C from B.
5. If in a consonant sequence a coronal plosive or affricate T is followed by a labial or dorsal plosive K, then we separate T from K.
6. If a consonant sequence contains two plosives or affricates, one voiced and one voiceless, then we separate them.
7. If a consonant sequence contains two fricatives other than в /v/, one voiced and one voiceless, then we separate them.
8. If a consonant sequence contains two labial plosives or two coronal plosives or affricates or two dorsal plosives then they are separated.

[20] Actually, the letter в /v/ is not a real exception because in all such cases this letter denotes two different consonants – в /v/ and ф /f/. Only in the Russian loan-word взвод /vzvod/ the two letters в /v/ denote a repeating consonant в /v/.

9. If a consonant sequence contains two equal consonants (not necessarily consecutive), then they are separated.

With so many prohibitive rules, a question arises: if we apply all these rules, aren't we going to eliminate too many hyphenation possibilities? The answer is no. It can be demonstrated that between any two consecutive syllables at least one separation point will be permitted.

Hyphenation according to the morphology

Between 1983 and 2012 the official orthographic rules of the Bulgarian language forbade morphologically based hyphenation. After 2012 such hyphenation is permitted (but not obligatory).

The most important case when it is very desirable to use morphologically based hyphenation is the case of the compound words. Divisions such as автоκ-луб /avtok-lub/ and вакуу-мапарат /vakuu-maparat/ are extremely irritating even if they are formally correct. Unfortunately, we do not have a vocabulary of the compound Bulgarian words that would permit us to produce rules for automated hyphenation. Therefore, the current Bulgarian hyphenation patterns do not attempt to apply morphological hyphenation to such words.

Second in importance (but far more significant in terms of numbers) is the case with the word prefixes. While the eyes of the reader still look at the start of the word, the word is still unknown to him. At this point, it is very important not to deceive his expectations. For example, when the reader sees над- /nad-/ at the end of the line, he will expect that this is the prefix над- /nad-/ with semantics 'attain more than'. This expectation will be fooled if this wasn't really a prefix, but a deceiving (while formally correct) hyphenation of the word надремя /nadremya/ 'have dozed enough' where the real prefix is not над- /nad-/ but на- /na-/ with semantics 'achieve a state after accumulation'. Such hyphenation distracts the reader and makes the reading more difficult.

Third in importance is the case with the word suffixes. With respect to the hyphenation rules we can divide the suffixes into three categories:

1. Suffixes starting with a vowel, for example -ар /-ar/. It is not appropriate to follow the morphology with such suffixes because this will contradict the whole hyphenation tradition of the Bulgarian language. For example крав-ар /krav-ar/ is unwarranted.
2. Suffixes starting with one consonant, for example -ка /-ka/. Usually with such suffixes the syllable boundary in the word coincides with morpheme boundary so no specific cares are necessary, for example кравар-ка /kravar-ka/. The exceptions are rare, for example: обек-тната /obek-tnata/ instead of обект-ната /obekt-nata/.
3. Suffixes starting with more than one consonant (-ски /-ski/, -ство /-stvo/). It is possible to use morphological hyphenation rules with such suffixes.

Even if it is possible to use morphological hyphenation with the suffixes of the third category, it turns out, this is not as useful as it is with the case of the prefixes. When the eyes of the reader have reached this part of the word, the word is already more or less known to the reader. Therefore, at this point the morphological hyphenation does not provide any significant advantages in comparison to the simpler hyphenation based only on the syllables in the word. Consider for example the word геройс-тво /geroys-tvo/ with suffix -ство /-stvo/. When the reader sees геройс- /geroys-/ at the end of the line this will give him an early clue that the suffix of the word is -ство /-stvo/. Such non-morphological hyphenation does not deceive the expectations of the reader. On the contrary, it makes the reading easier because it gives clues to the reader about what follows on the next line.

Because of these considerations, the current Bulgarian hyphenation patterns do not attempt to use morphological hyphenation with respect to the suffixes of the words. Though it would be useful to implement rules about the suffixes of the second cateogory. Hopefully, some future version will have such rules.

Occasionally,[21] a fourth morphological requirement is stated: that hyphenation should conform with the boundary between the word and the definitive articles -та /-ta/ and -те /-te/ (postfixed in Bulgarian). There is no need to pay attention to this rule because it seems to be satisfied by its own nature. The author has searched in a dictionary with over 860000 Bulgarian words for cases when the hyphenation rules would hyphenate badly with respect to the definitive article. He was unable to find even one such case with the hyphenation rules valid after 1983 and only about 10 cases with the rules valid before 1983 (one of them is живопи-ста /zhivopi-sta/ instead of живопис-та /zhivopis-ta/).

One unavoidable characteristic of any morphologically based automated hyphenation is that it can create wrong hyphenations. Because of this, one useful option is to use the morphology in a safe way – to use it in order to forbid bad hyphenations but to create no new hyphenation possibilities solely on the basis of the morphology.

Take for example the word дозрея /dozreya/ 'ripen fully'. According to the phonological rules, we should hyphenate it as доз-рея /doz-reya/. According to the morphology, however, we should hyphenate as до-зрея /do-zreyq/ because this word is formed with the prefix до- /do-/ with semantics 'complete or supplement' and this semantics would be lost if the reader sees доз- /doz-/ at the end of the line. Therefore, there are three methods to hyphenate this word:

1. доз-рея /doz-reya/ when morphology is not used;
2. до-зрея /do-zreya/ when morphology is fully used;
3. дозрея /dozreya/ (no hyphenation) when morphology is used in a safe way.

[21] Правописен и правоговорен наръчник. Състав. Иван Хаджов, Цв. Минков; Ред. Ив. Хаджов и др. София, Бълг. кн., 1945

The option to use the morphology in a safe way is very attractive when the software uses a smart line-breaking algorithm which can produce good results even with less hyphenation possibilities. TeX is one such software. It should be noted that this option does not eliminate too many hyphenation possibilities because the morpheme boundaries most of the time are also syllable boundaries.

The following are results of a statistics about the quality of the morphological rules (the number after the sign ± is the expected standard deviation of our estimations):

With the option `--morphology`:

- in 0.1% ±0.3% of the dictionary words the morphological patterns create very wrong hyphenation;
- in 89.8% ±0.1% of the dictionary words the morphological patterns hyphenate identically with the case when no morphology patterns are used;
- in 0.3% ±0.2% of the dictionary words the morphological patterns hyphenate differently in comparison to the case when no morphology patterns are used and the word is hyphenated in a way which contradicts the morphology;
- in 0.6% ±0.1% of the dictionary words the morphological patterns hyphenate differently in comparison to the case when no morphology patterns are used and there is a possible hyphenation which is compatible with the word morphology but which is nevertheless forbidden by the morphology patterns.

With the option `--safe-morphology`:

- in 0% of the dictionary words the morphological patterns create very wrong hyphenation;
- in 90.0% ±0.1% of the dictionary words the morphological patterns hyphenate identically with the case when no morphology patterns are used;
- in 0.3% ±0.2% of the dictionary words the morphological patterns hyphenate differently in comparison to the case when no morphology patterns are used and the word is hyphenated in a way which contradicts the morphology;
- in 0.6% ±0.1% of the dictionary words the morphological patterns hyphenate differently in comparison to the case when no morphology patterns are used and there is a possible hyphenation which is compatible both with the word morphology and with the syllable boundaries but which is nevertheless forbidden by the morphology patterns.

Notice that the morphological patterns create a different hyphenation only in about 10% of the words. The following explanation can be given for this surprising fact. First, the natural evolution of the human languages tends to simplify the complex sequences of consonants. Therefore, no morpheme contains a complex sequence of consonants. And second, the Bulgarian orthography is morphological. This means that the morphemes are written according to their actual pronunciation, however the simplifications in the spoken languages

which take place at the morpheme boundaries are not taken into account in the orthography. The independent operation of these two factors leads to the result that most of the time the morpheme boundaries coincide with the conventional syllable boundaries. The main exception to this is when a morpheme starts with a vowel, in this case its syllable will include one or more consonants of the preceeding morpheme. The second exception is when a morpheme ends with a vowel and the next morpheme starts with a sequence of two or more consonants.

Usage of the script `hyph-bg.sh`

The `hyph-bg.sh` is all-in-one script which can generate both documentation (this text) and Bulgarian hyphenation patterns. When given the option `--help` the script gives short usage instructions:

```
hyph-bg.sh --help
          Show this info
hyph-bg.sh [--doc-html | --doc-latex | --doc-txt]
          Print documentation in various formats
hyph-bg.sh [other options]
          Generate Bulgarian hyphenation patterns

Options when generating hyphenation patterns:

 --standalone-tex
          Produce hyphenation patterns for TeX with \patterns{ ... }.

 --no-hyphen-mins
          Hyphenation patterns which do not require hyphen mins.
          Otherwise: both left and right hyphen mins should be set to 2.

 --safe-dz
          Do not try to guess whether DZ is a single consonant or not.
          Only use hyphenation which will be correct in both cases.

 --permissible
          Permit any formally correct hyphenation, including unnatural
          divisions, such as studen-tstvo.  Useful for educational tools
          or when typesetting Bulgarian text in a very short column.

 --morphology
          Apply morphology when hyphenating, for example: za-dvizhvam.
          May hyphenate incorrectly in some cases.

 --safe-morphology
          Apply morphology when hyphenating.  Never hyphenates incorrectly
```

```
        but may prohibit some correct hyphenations.

--no-morphology
        Disregard the morphology.  Default.

--1945
        Hyphenate according to the rules effective between 1945 and 1982

--1983
        Hyphenate according to the rules effective between 1983 and 2011

--2012
        Hyphenate according to the rules effective after 2012.  Default.
```

The following are the recommended ways to generate hyphenation patterns by this script:

`hyph-bg.sh --standalone-tex --safe-morphology` For TeX. Apply the morphology in a safe way when the software uses a smart line-breaking algorithm.

`hyph-bg.sh` For most other software.

`hyph-bg.sh --no-hyphen-mins` The current versions of Mozilla (as of 2017) seem to ignore the hyphen mins in words that contain a dash.

`hyph-bg.sh --morphology` For professional typography with human proofreader.

`hyph-bg.sh --permissible` For educational tools and online dictionaries which can show only one kind of hyphenation.

Notice that some specialised English dictionaries separate the word-division positions into two categories – preferred positions and less recommended positions. It would be best if the Bulgarian online dictionaries could do the same. For example hyphen "-" can be used to display the preferred positions and dot "." – the less recommended positions. If a word-division position is permitted only by the patterns of `hyph-bg.sh --permissible`, then this position is less recommended.