Perfect Bulgarian Hyphenation, or how not to stutter at end-of-line

Anton Zinoviev
Sofia University "St. Clement of Ochrid"
Institute of Information and Communication Technologies
at Bulgarian Academy of Sciences,
anton@lml.bas.bg

Abstract

What is Perfect Bulgarian Hyphenation? We know that it has to be based somehow on the syllables and on the morphology but considering that these two factors often contradict each other, how exactly are we going to combine them? And speaking about syllables, what are they and how are we going to determine them? Also, how are we going to find the morphemes in the words? Don't we have to develop an electronic derivational dictionary of the Bulgarian language? Isn't all this going to be forbiddingly difficult?

1. Foreword

What heartless man is not going to sympathise with an intelligent speaker whose stuttering distracts the listeners, the thoughts behind his words remaining unheard? Demosthenes had to train in a cave with pebbles in his mouth and a sword over his shoulder. He had to make his speeches more apprehensible, this was a matter of life and death. In this paper we shall see that in the written language, too, there can be distracting things. In the spoken language the speech therapists fight the stuttering and in the written language the professional printers do the same. For example, they know that it is preferable not to use boldface. They also know that irregular white space is distracting, so it has to be eliminated by proper hyphenation. Likewise, the hyphenation should be done in such a way that while the eyes of the reader are moving from a line to the next one, his expectations about what follows are not deceived.

Indeed, it is this striving for clarity what has made the English hyphenation so complex. One peculiarity of the English language is that the pronunciation of a vowel depends on whether in its morpheme it belongs to an open or to a closed syllable. For example, the vowel *e* of the morpheme *hyphen* of the word *hyphenation* is part of the closed syllable *phen*. Consequently, the vowel *e* is pronounced as if it is in a closed syllable even though its syllable in the word *hyphenation* is the open syllable *phe*. If we hyphenate this word as *hyphe-nate* we will confuse the reader because while his eyes are still moving from *hyphe-* to the next line, he will expect that *e* is part of an open syllable with pronunciation as in *me* or *bee*. Therefore, the English hyphenation prefers not to change the apparent closeness of the syllables. This explains cases like *collect-ing*, *mod-el*, *sec-ond*, *trav-el*. It also explains the hyphenation of homonyms with equal spelling but different pronunciation, such as *prog-ress* and *pro-gress*, *rec-ord* and *re-cord*, *eve-ning* and *even-ing*.

2. Bulgarian Affairs

As for the Bulgarian hyphenation, it has always been governed by the same two factors as the hyphenation of most other languages—the syllables and the morphology. As in most languages, the case when a syllable boundary coincides with a morpheme boundary is clear. Uncertainties arise only when the syllables and the morphology specify different positions for word division. For a relatively long period the Bulgarian hyphenation was done intuitively and according to the existing tradition. Formal rules existed only about the most important cases. It seems the earliest attempts to formulate extensive exact rules about the Bulgarian hyphenation were from 1945 (Andreychin, 1945; Hadzhov and Minkov, 1945).

Keywords: hyphenation, morphology, syllables, phonology, Bulgarian

Unsurprisingly, these rules turned out to be complex. Due to this complexity, many mistakes were made. Especially the hyphenation in the newspapers was more or less arbitrary. So, instead of fixing the newspapers, someone decided that it would be easier to fix the hyphenation rules. In result, in 1983 the Institute for Bulgarian language published new hyphenation rules (Georgieva and others, 1983). These rules broke completely with the existing tradition. The morphology was no longer a factor and the syllables were proclaimed as the main ruling factor. However, the syllables were not to be determined according to the most convenient pronunciation, but rather by some non deterministic mechanistic rules.

Because the morphology was ignored, there were some absurd hyphenations, such as авток-луб (*avtok-lub* 'moto c-lub') and вакуу-мапарат (*vacuu-maparat* 'vacuu-m apparatus'). In many cases the mechanistic hyphenation rules were too permissive. For example for the word агентство (*agentstvo* 'agency') we could have аген-тство (*agen-tstvo*), агент-ство (*agent-stvo*), агентс-тво (*agents-tvo*), агентст-во (*agentst-vo*), all at once. This was so despite that the pronunciation of the parts агентст- (*agentst-*) and -тство (*-tstvo*) was clearly impossible. In other cases the rules were too restrictive. For example the hyphenation на-дро-бя (*na-dro-bya* 'crumble') was forbidden despite that it seems that this is the most natural syllable division (and in addition, it is in agreement with the morphology).

However bad the new rules were in some aspects, they were good about the following: they were exact, unambiguous and they were easy to implement in software. The earliest mathematical analysis of the new Bulgarian hyphenation rules was by Noncheva (1988). She proposed a mathematical formalisation of the hyphenation rules in a table of 22 rows. In the same year, Belogay (1988) proposed an alternative formalisation with only 9 rules. Belogay proved that his rules were consistent and that they formed a minimal set. The rules of Belogay had a negative character—every hyphenation which was not forbidden by a rule was permitted. The work of 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 implemented 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 the author of one book about TEX (Vasilev, 1997) didn't care to give any explanations but simply wrote about "the algorithm of Belogay" as something well known to the reader.

The earliest implementations of the Bulgarian hyphenation in TEX did not rely on the internal hyphenation algorithm of TEX. Instead, an external tool implementing the algorithm of Belogay was used to insert soft hyphens in all Bulgarian words. The first usable Bulgarian hyphenation patterns for TEX were developed by Georgi Boshnakov 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 LATEX 2_ϵ became the official LATEX version. Later Boshnakov modified his work with the Babel system. The hyphenation patterns of Boshnakov did their job well enough, so for almost quarter a century after their initial creation, they remained the only Bulgarian hyphenation patterns in the standard distributions of TEX and CTAN.

The algorithm of Belogay and the hyphenation patterns of Boshnakov adhered to the official hyphenation rules of 1983. Nevertheless, the new rules were not universally accepted. Even today, the traditional rules by Andreychin (1945) are mentioned in various places in Internet. They are also included in some grammar books (Pashov, 1989; Stoyanov, 1993).

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" (Topalov, 1995). 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". The earliest works which have been studied are from 1978. It turned out that they present the best approach concerning the automated hyphenation.

In 1999 in a paper about the automated Bulgarian hyphenation, Koeva (1999) published a list of

hyphenation patterns that could be used as a basis for automated hyphenation. In 2004 with the help of Stoyan Mihov she formalised these rules 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 Corp. and Apple Inc. The hyphenation rules of Koeva were more permissive than the official rules. For example they permitted cases such as се-стра (*se-stra* 'sister'), ай-сберг (*ay-sberg*) 'iceberg') and материа-лна (*materia-lna* 'physical').

In 2000 Anton Zinoviev created new hyphenation patterns for TEX. In 2001 Radostin Radnev used the hyphenation patterns of Zinoviev in his free grammar dictionary of Bulgarian. From there the work of Zinoviev propagated to OpenOffice, LibreOffice and various online dictionaries, including `bg.wiktionary.org` and `rechnik.chitanka.info`. However Zinoviev didn't bother to make his work officially available in the various TEX distributions and CTAN.

The hyphenation patterns of Zinoviev were more restrictive than the official rules. For example in consonant sequences like тст (*tst*) in братство (*bratstvo* 'brotherhood'), the two equal consonants т (*t*) were always separated, so that братст-во (*bratst-vo*) was forbidden. The hyphenation was forbidden after a sonorant consonant following an obstruent consonant. For example отм-ра (*otm-ra* 'die out') was forbidden but от-мра (*ot-mra*) was permitted. Also, the hyphenation separated a pair of two kindred, one voiced and one voiceless consonants. For example субп-родукт (*subp-rodukt* 'subproduct') was forbidden and суб-продукт (*sub-produkt*) was permitted.

Eventually, in 2012 the Institute for Bulgarian language published revised hyphenation rules (Murdarov and others, 2012). The new rules are even more liberal than the rules of 1983. While absurdities such as авток-луб (*avtok-lub*) and вакуу-мапарат (*vakuu-maparat*) remain valid, the main advantage of the new rules is that the natural hyphenations авто-клуб (*avto-klub*) and вакуум-апарат (*vakuum-aparat*) are now also valid. It seems that the linguists at the Institute for Bulgarian Language have recognised that good hyphenation is a complex matter. They no longer attempt to invent universal rules about everything. Instead, they provide some very permissive 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.

The present work was carried out on the initiative of the leader of Bulgarian localisation team of the browsers Mozilla and Firefox. In 2017 he contacted me with an inquiry about the best automated Bulgarian hyphenation. Since the new official hyphenation rules were so permissive, I told myself: "Great, it seems I will be free to implement the hyphenation in any way I see fit or I deem appropriate. Good or bad, there will be fair chances that my implementation will be in compliance with the official rules. If I want to make a computer implementation of the *Perfect Bulgarian Hyphenation*, my hands will be untied and I will be free to act." So, to act I decided.

3. Plan

Evidently, we are coming to the real question: what is *Perfect Bulgarian Hyphenation*? We know that it has to be based somehow on the syllables and on the morphology but considering that these two factors often contradict each other, how exactly are we going to combine them? And speaking about syllables, what are they and how are we going to determine them? Also, how are we going to find the morphemes in the words? Don't we have to develop an electronic derivational dictionary of the Bulgarian language? Isn't all this going to be forbiddingly difficult? Let us delay no more and move right to the answers.

4. Combining Syllables with Morphology

Let us recall that when the English printers decided to hyphenate *beam-ing*, *draw-ing*, *stew-ing*, etc., they did so in order to make the reading easier. Then some grammarian noticed these specific cases and proclaimed the general rule that in all present participles we have to hyphenate before the ending *-ing*. This was a generalisation made by someone who admired general grammatical rules but didn't really understand the real objective of the hyphenation. In result, now we have deceiving cases, such as *hat-ing*.

Therefore, whatever rules about the hyphenation we invent, we should never lose sight of its main objective. And this objective is to make the reading smoother and easier. When our eyes are moving from

one line to the next, we know only the first part of the hyphenated word and we have only expectations about what follows. Any unnecessary ambiguity is bad. Creating wrong expectations and fooling them is worse. Bad hyphenation certainly is capable to confuse and to disturb the readers.

Ignoring completely the morphology is one good way to deal with the conflict between syllables and morphology. We the people are adaptive creatures. We can get used to any rules as long as they are not too unreasonable. And hyphenating according to the syllables of the word is certainly not a totally unreasonable way to hyphenate.

Nevertheless, there are cases when hyphenation according to the morphology creates less confusion. We already saw one such case: we have to respect the constituents of a compound word. Divisions such as авток-луб (*avtok-lub*) and вакуу-мапарат (*vakuu-maparat*) are extremely irritating.

Second in severity (but far more numerous, so also important) is the case with the word prefixes. While the eyes of the reader still look at the first part of the word, the rest is unknown. At this point, it is very important not to deceive the expectations. For example, when the reader sees на- (*na-*) at the end of the line, he will expect that this is the prefix на- (*na-*) with semantics 'achieve a state after accumulation'. This expectation will be fooled if this wasn't really a prefix, but a deceiving hyphenation of the word на-диграя (*na-digraya* 'outplay') where the real prefix is not на- (*na-*) but над- (*nad-*) with semantics 'attain more than'. Even more confusing is the case when we see над- (*nad-*) at the end of line and this wasn't really the prefix над- (*nad-*) but a deceiving hyphenation of the word над-ремя (*nad-remya* 'have dozed enough') where the real prefix is not над- (*nad-*) but на- (*na-*). Such hyphenations distract the reader and make the reading more difficult.

The traditional Bulgarian hyphenation rules (Andreychin, 1945; Hadzhov and Minkov, 1945) prescribed that in all cases the prefixes should be respected. In some cases the hyphenation was able to differentiate between two homonyms. 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 overnight'). On the other hand, the requirement to respect the suffixes was significantly relaxed: they should be preserved only when this doesn't create impossible syllables. Indeed, a hyphenation хлеб-ар (*hleb-ar* 'baker') is completely unwarranted despite that хлеб (*hleb*) is the root, -ар (*-ar*) is the suffix and both morphemes are productive in the modern language.

How can we explain the different treatment of the prefixes and the suffixes? If we try to shout rhythmically, syllable by syllable the word хлебар (*hlebar*), then the shouting хлеб-ар (*hleb-ar*) will be very unnatural and strained. However, if we do the same experiment with the prefixes, we will find something unexpected: it is quite possible (even if somewhat inconvenient) to rhythmically shout над-и-гра-я (*nad-i-gra-ya*), под-у-ча (*pod-u-cha*), etc. Despite that in normal speech these are not the syllables in these words, it is possible, nevertheless, to divide the words in this way. Clearly, the different treatment of the prefixes and the suffixes in the traditional Bulgarian hyphenation was not an arbitrary decision but it had to do with something different about the phonology of the prefixes and the suffixes.

The glottal stop (ʔ) is this different thing. Many languages use the glottal stop as a regular consonant. In the Bulgarian language it is not phoneme,[1] however it is readily inserted at the beginning of words starting with a vowel. For example, if we try to pronounce the word уча (*ucha* 'learn'), then there are good chances that in reality we will pronounce ʔуча (*ʔucha*). Now, notice that there are words starting with a prefix or with a root, but there are no words starting with a suffix. Therefore, there are plenty of cases when Bulgarians will add a glottal stop in front of a prefix or in front of a root starting with a vowel. So the Bulgarians are used to treat the prefixes об- (*ob-*) and ʔоб- (*ʔob-*) as identical, the roots уча (*ucha*) and ʔуча (*ʔucha*) as identical and so on. In fact, untrained Bulgarians won't even notice the difference. On the contrary, since a word never starts with a suffix, there are no cases when Bulgarians would insert a glottal stop before a suffix. If someone decides to pronounce a glottal stop before a suffix, this will be something very noticeable to any Bulgarian.

Therefore, we can formulate the following rule: in the traditional Bulgarian hyphenation, the morphology was subordinated to the syllables. We would never divide a word according to the morphology

[1] One curiosity is the negative particle ʔъ-ʔъ (*ʔă-ʔă*) which is the only Bulgarian word using the glottal stop as a phoneme. Despite that this word is very common, I have no idea how to write it with Cyrillic letters.

if this creates impossible syllables. On the other hand, we were not insisting about using the most natural syllable division. If it was possible to preserve the morphology by insertion of a glottal stop, then we would do so in order to preserve the morphology. For example divisions such as над-играя (*nad-igraya*) and под-уча (*pod-ucha*) in reality were над-ʔиграя (*nad-ʔigraya*) and под-ʔуча (*pod-ʔucha*).

Should we respect the prefixes in the *Perfect Bulgarian Hyphenation*? Unfortunately, it seems impossible to formulate a clearly cut rule about this. Remember that our goal should be to make the reading easier by creating the right expectations in the reader while his eyes are moving from one line to the next. Cases as над-играя (*nad-igraya* 'outplay') and под-уча (*pod-ucha* 'to prompt') seem most clear because the most natural syllable divisions на-диграя (*na-digraya*) or по-дуча (*po-ducha*) create deceiving impression about the prefix. Somewhat less clear are cases as раз-ора (*raz-ora* 'plough up') whose derivation is productive in the modern language. To a more literate reader a prefix раз- (*raz-*) will provide more useful information than a meaningless syllable ра- (*ra-*). Therefore, to such a reader this hyphenation will be helpful. Is this going to be so with a less literate reader? I don't know.[2] And perhaps there is no need to preserve the prefix in cases when the derivation is from an obscure root (под-ема (*pod-ema* 'take on')) or when the root is clear but nevertheless the prefix is not felt as a prefix (раз-ум (*raz-um* 'intelligence')).

One clear counterexample is the word отявлен (*otyavlen* 'downright') where the morphology boundary and the syllable boundary coincide. Therefore, the preferable hyphenation of this word is от-явлен (*ot-yavlen*) even though the Cyrillic letter я (*ya*) becomes merged with the following vowel а (*a*) and creates the false impression of incorrect syllable division. Indeed, the letter я (*ya*) in this word signifies the semivowel й (*y*). If, on the other hand, the syllable boundary were о-тявлен (*o-tyavlen*), then the letter я (*ya*) would no longer signify a semivowel but a palatalisation of the preceeding sound т (*t*).

The traditional Bulgarian hyphenation tried to respect the suffixes but only when this would create no conflict with the syllables. Should we do the same in the *Perfect Bulgarian Hyphenation*? There are three cases to consider.

First, it is not appropriate to follow the morphology when the suffix starts with a vowel. This would contradict the whole Bulgarian hyphenation tradition where the morphology has a subordinate role with respect to the syllables. For example хлеб-ар (*hleb-ar*) is unwarranted.[3]

Second, when a suffix starts with one consonant, for example -ка (*-ka*), then the morpheme boundary is a possible syllable boundary. Therefore, even if we disregard the morphology, we are not going to divide the suffix. The only thing we should watch out is not to divide the morpheme preceding the suffix. There is no need to have too many hyphenation possibilities in order to obtain good looking printed document. Therefore, since обект-ната (*obekt-nata* 'object (adjective)') is permitted both according to the morphology and to the syllables, then there is no need to use обек-тната (*obek-tnata*), especially considering that обект- (*obekt-*) at the end of line provides the reader with more useful information than обек- (*obek-*). Similarly, since the division агент-ка (*agent-ka* 'agent woman') is permitted both by the morphology and by the syllables, then there is no need to use аген-тка (*agen-tka*), especially considering that агент- (*agent-*) provides the reader with more useful information than аген- (*agen-*).

And third, there are suffixes starting with more than one consonant (-ски (*-ski*), -ство (*-stvo*)). The traditional Bulgarian hyphenation did not allow such suffixes to be divided.[4] Nevertheless, I assert that in the *Perfect Bulgarian Hyphenation* it is permissible to divide these suffixes. In fact, it is not just permissible to do so, but it is also preferable to do so. When the eyes of the reader reach the end of line and he sees there, say, братс- (*brats-*), then he will know that there are very good chances that this is one of the words братски (*bratski* 'brotherly') or братство (*bratstvo* 'brotherhood') and the suffix is -ски (*-ski*) or -ство (*-stvo*). If, on the other hand, the reader sees at the end of line брат- (*brat-*) then he will know that брат (*brat*) is the root of the word, but there will be too many other possibilities for the word besides братски (*bratski*) and братство (*bratstvo*). While the hyphenation братс-ки (*brats-ki*) is not morphological, it does not deceive the expectations of the reader and it makes the reading easier because

[2] Even the illiterate people feel the prefixes intuitively. The current official hyphenation rules leave to the discretion of the writer whether to respect the prefix or not. I think this is the best possible decision about this issue.

[3] Remember that we are not permitted to insert a glottal stop before the suffix -ар (*-ar*).

[4] It seems before 1945 this was a mandatory rule and after 1945—only a recommendation.

it gives more clues to the reader about what follows on the next line.

With this I can conclude this section of the article. We saw that the case about the suffixes was clear and unambiguous, even if somewhat complex. Some things about the prefixes were more ambiguous and depended on the personal preferences. Fortunately, there is software using a smart line-breaking algorithm which is able to produce good results even when only few hyphenation possibilities are available. One such software is TEX. With such software perfect results can be achieved when the hyphenation rules permit a word division only when it is compatible *both* with the prefix morphology and with the syllables. Therefore, when we use such software, both раз-ора (*raz-ora*) and ра-зора (*ra-zora*) should be forbidden and the software will still be able to produce a good printed document.

5. The Bulgarian Syllables

Many things about how our brain processes the speech are still unknown. It seems that the audio signal is processed as a hierarchy of carefully arranged segments. For example the intonation helps us to divide the signal into sentences. The stress helps us to divide the sentence into words. That's why in so many languages the stress has a fixed position in the word (penultimate, ultimate, or word-initial). One interesting thing about the stress is that it does not exist as such in the audio. Instead, it is an illusion, created by our perception. Several different factors, such as the tone, the rhythm within the sentence, the loudness and the reduction of the vowels, are used together in a complex way in order to determine where the stress is. Even when the information provided by these factors is inconclusive, we may still perceive the stress in its most probable position.

On a lower level, just as the stress helps divide the sentence into words, the sonority within the audio signal helps us to divide the words into syllables. The real nature of the sonority is still unknown. When the sonority reaches a peak above a certain threshold (which depends on the language), then we perceive a syllable. The peak of the sonority is exactly at the *nucleus* of the syllable. The part of the syllable which is before the nucleus is called *onset* and the part after the nucleus is called *code*.

When the sonority reaches a peak which is below the threshold, then such a peak does not signal the existence of a syllable (Zec, 1995). Such peaks make the speech perception more difficult. That's why the languages try to eliminate such false peaks. This is known as the *sonority sequencing principle*. It says that the sounds within the onset have raising sonority while the sounds within the code have decreasing sonority.

It is said that the syllables are abstract phonological constituents without clear phonetic correlates (Ladefoged and Maddieson, 1996). Nevertheless, the syllables are not just fanciful artificial creations whose only purpose is to amuse the linguists. They correspond to the way our brain processes the speech. Because of this they can be the base of a very natural system for hyphenation. That's why they are used for this purpose in many languages, Bulgarian included.

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

$$\text{voiceless obtrusive} < \text{voiced obtrusive} < \text{sonorant consonant} < \text{vowel}$$

According to my investigations, it seems that Bulgarian respects the sonority sequencing principle more accurately than most other languages. The only exception to the above scale in the written language is due to the letter в (*v*) which is a voiced obtrusive but it can be used as if it were 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 the letter в (*v*) is read as ф (*f*) (which is a voiceless obtrusive). For example the word отвсякъде (*otvsyakăde* 'all round') is read as отфсякъде (*otfsyakăde*).[5]

I have found that the sonorant consonants in Bulgarian have their own sonority scale: м < н < л < р < й (*m* < *n* < *l* < *r* < *y*). Only a few words such as жанр (*zhanr* 'genre') and химн (*himn* 'anthem') violate this scale. Such words are always loan-words and their pronunciation is somewhat problematic for the native Bulgarian speakers.

[5]Since no Primitive Slavonic word contained the phoneme ф (*f*), we can hypothesize that in the Primitive Slavonic language the consonants ф (*f*) and в (*v*) were two positional variants of a single phoneme.

From the sonority sequencing principle we can deduce the following two hyphenation rules. First, in a sequence MK where M is a consonant with higher sonority than K, we are not permitted to hyphenate before M (except when M is *v* and K is a voiceless consonant). And second, in a sequence KM where M is a consonant with higher sonority than K, we are not permitted to hyphenate after M.

In addition to the Sonority Sequencing Principle, the consonant clusters within the Bulgarian syllable adhere to the following 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* 'context').
3. If the onset or the code contains two fricatives other than в (*v*), then there can be 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* 'indurate')).[6]

From these seven properties we can deduce corresponding hyphenation rules. For example from the first property we deduce that in a consonant sequence where a coronal plosive or affricate T is followed by a labial or dorsal plosive K, we separate T from K. From the second property we deduce that in a sequence KBT where K and T are plosives or affricates and B is fricative, we separate K from T. Etc.

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. All that these rules do is helping the software to determine more accurately the exact boundary between the syllables. It can be demonstrated that between any two consecutive syllables at least one separation point will be permitted.

6. Finding the Morphemes

How a computer can find the morphemes in a word? It turns out, there is no need to do this. At least not too often. We saw already one reason for this—there are cases when we have to ignore the morphology (remember хлеб-ар (*hleb-ar*)). And the second reason is the following: when the second morpheme starts with a consonant, then the morpheme boundary coincides with the syllable boundary. So we only have to discover the syllable boundary. In the previous section we saw how we can do this with sufficient precision.

The reason the morphology so often does not contradict the syllables is the following. First, every language has a tendency for simplification during its natural evolution. When a particular simplification concerns a single morpheme then it is easier for this simplification to actually happen in the language. However, when a simplification concerns the contact between two separate morphemes, then this simplification is more difficult and can actually happen only in the following two cases: 1. when it concerns unproductive and obscure morphemes or, 2. when it is a result of a regular phonological law in the

[6] 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*). In the word втвърди (*vtvardi*) the first в (*v*) is pronounced as ф (*f*). Only in the Russian loan-word взвод (*vzvod* 'platoon') the two letters в (*v*) denote a repeating consonant в (*v*).

language. The case 1 should not concern us. As for case 2, it is rare because the Bulgarian orthography is largely morphological. This means that morphemes are written according to their pronunciation. However, Bulgarian orthography usually ignores the phonological changes that happen in the spoken language at the contact of two morphemes. In other words, case 2 of the above two happens rarely in the written Bulgarian language. Because of this, when the contact of two morphemes is at a consonant cluster, then the place of greatest complexity within the cluster is boundary both of the syllables and of the morphemes.

In order to discover the morphological hyphenation rules, first select an arbitrary morpheme. Then try to predict when it will have the potential to generate different hyphenation with respect to the hyphenation based solely on the syllables. For example when a prefix ends with a consonants, then we will be interested by cases when this prefix is followed by a vowel, like in the word раз-ора (*raz-ora*). This is so because in such cases the hyphen determined by the morphology will differ from the hyphen determined by the syllables. You will find that such cases are not numerous. It is possible without too much efforts to manually observe all potential cases and write hyphenation rules for each prefix.

A somewhat more complex is the situation with prefixes like по- (*po-*) and под- (*pod-*). It can be summarized by the following rules:

1. When a word starts with по- (*po-*) and the next letter is not д (*d*), then по- (*po-*) is likely a prefix.
2. When a word starts with под- (*pod-*) and the next letter is a consonant, then под- (*pod-*) is most likely a prefix.
3. When a word starts with под- (*pod-*) and the next letter is not a consonant, then по- (*po-*) is most likely a prefix.

We only need to describe the exceptions to the above rules and such exceptions are not numerous.

Let me give a complete example. For the prefixes о- (*o-*), об- (*ob-*) and от- (*ot-*) I have found the following rules:

prefix о- **(*o-*)** when the following letter is not б (*b*) nor т (*t*)[7]

Exceptions: оазис (*oazis*), овц (*ovc*), овч (*ovch*), огн (*ogn*), окси (*oksi*), окт (*okt*), олтар (*oltar*), омлет (*omlet*), омни (*omnni*), онбаш (*onbash*), ондул (*ondul*), онзи (*onzi*), онко (*onko*), онлайн (*onlayn*), онто (*onto*), опт (*opt*), опци (*opci*), опб (*opb*), орг (*org*), орд (*ord*), орк (*ork*), орл (*orl*), орн (*orn*), орт (*ort*), орф (*orf*), орх (*orh*), осман (*osman*), осмин (*osmin*), осмиц (*osmic*), осмич (*osmich*), осмо (*osmo*), осте (*oste*), остро (*ostro*), осци (*osci*), охва (*ohva*), охка (*ohka*), охна (*ohna*).

prefix об- **(*ob-*)** when it is followed by a consonant

Exceptions when this is not об- (*ob-*) but о- (*o-*): облаго (*oblago*), облаж (*oblazh*), обрем (*obrem*), обрул (*obrul*), обръс (*obrăs*), овдов (*ovdov*), овлад (*ovlad*). Exception when this is neither об- (*ob-*), nor о- (*o-*): общн (*obshtn*). Cases of об- (*ob-*) followed by a vowel: обагн (*obagn*), обигр (*obigr*), обясн (*obyasn*), ообощ (*obobsht*), обозн (*obozn*), обозр (*obozr*), обосн (*obosn*), обособ (*obosob*), обузд (*obuzd*), обусл (*obusl*).

prefix от- **(*ot-*)** when it is followed by a consonant

Cases of от- (*ot-*) followed by a vowel: отив (*otiv*), отид (*otid*), отуч (*otuch*).

7. Implementation

The author has implemented a shell script[8] which generates Bulgarian hyphenation patterns in the form expected by TEX. The output of this script is about to be used by TEX, LibreOffice, OpenOffice and the browser Mozilla. The script is configurable—the user choses whether or not to use the morphology,

[7] Despite that the Bulgarian hyphenation rules do not permit lone letters, we have to discover this prefix anyway. This is so because we have to ensure the root is not divided in the vicinity of the prefix.

[8] `https://sourceforge.net/p/bgoffice/code/HEAD/tree/trunk/hyph-bg/hyph-bg.sh`

whether or not to use only good syllable divisions and which version of the hyphenation rules published by the Institute of Bulgarian language to use (1945, 1983 or 2012).

Some statistics follow. When the script is used to generate patterns strictly adhering to the rules published by the Institute for Bulgarian Language (no detection of the syllables and no morphology), then it will output 1676 patterns. If we chose to detect the syllables, then we will have 5798 patterns. And if, in addition, we chose to use also the mophology, then we will have 6886 patterns.

Any computer implementation of a morphology based hyphenation will make mistakes. According to a test with 303 randomly chosen words we have the following figures (after the sign $\pm$ is the expected deviation of the estimation): in $3.3\% \pm 1.0$ of the words there is a hyphenation which contradicts the prefix morphology, such as несп-равям (*nesp-ravyam* 'failure to do s.th.') and $2.9\% \pm 1.0$ of the words are badly hyphenated compound words, such as самок-ритичен (*samok-ritichen* 'critical to o.s.').

No words were found where the hyphenation contradicts both the morphology and the syllables. In order to find a more precise estimate of this worst case, an additional test with 122 words was run, using only words whose hyphenation with and without the morphology is not identical (there are about $10.2\% \pm 0.1$ such words). Again no such cases were found. Based on both tests, we can expect that only 0.06% of the words are hyphenated in a really bad way.

8. Debate

Now, some people might ask: is the hyphenation that important to justify all the big efforts to implement them in software? And to this I give the following response: but are the efforts really that big? We have to develop good hyphenation rules only once and then thousands can use them for years to come. The rules I have developed are part of the standard distributions of TEX, LibreOffice and Firefox, so all users will benefit for free and with no efforts.

Others will ask: OK, maybe it is not difficult to implement the rules in software. But clearly, these rules are too complex to be used by people. Well, when people use computers, how often do they hyphenate the words by themselves? Seldom. We live in the 21^{th} century! In these days hyphenation is done by computers, not by people. Only in handwriting people still hyphenate themselves. Are the people going to use simplified hyphenation in their handwriting? Of course. Is there a problem?

But the Institute for Bulgarian Language has published *the official* hyphenation rules. Shouldn't we follow these rules exactly instead of inventing our own? Well, we do follow the official rules. But do we have to hyphenate изг-рев (*izg-rev* 'sunrise') only because the official rules say this is OK? No, because the official rules say из-грев (*iz-grev*) is also OK and we like the second option more. Thanks be to God for after 2012 abominations like селскос-топански (*selskos-topanski* 'agric-ultural') are no longer compulsory.

9. Conclusion

Since I don't have a list of the Bulgarian compound words, the morphological hyphenation rules I have developed are not concerned with the morphology of such words. Because of this, the Bulgarian computer users are still coerced to accept crazy things like селскос-топански (*selskos-topanski*). Let us hope that in the future some good person with love for the Bulgarian language will make such a list. Then we all will benefit.

Fortunately, it wasn't that difficult to develop morphological rules about the prefixes. These rules make very few errors. I haven't started with the suffixes yet but I hope they won't be difficult either.

The development of the rules about the Bulgarian syllables has given me so much fun! I had to dive deep into the wonders of the Bulgarian phonology. So many questions and kindling curiosity! Why the coronal consonants follow the labial and the dorsal consonants? Does this happen only by accident or there is a more significant reason?

In the battle between the syllables and the morphology, each pushing its own principles, we found that the syllables were victorious. But they were a generous victor who leaved to the morphology quite a lot of governing rights. To save ourselves from problems we will have to reckon with both of them.

Gratitude is due to all who have worked before me in the area of hyphenation.

References

Andreychin, L. (1945). *Pravopisen rechnik na bulgarskiya knizhoven ezik*. Sofia, Hemus.

Belogay, E. (1988). Algoritam za avtomatichno prenasyane na dumi. *Kompyutar za vas*, 3:12–14.

Georgieva, E. et al. (1983). *Pravopisen rechnik na savremenniya bulgarski knizhoven ezik*. Sofia, BAN.

Hadzhov, I. and Minkov, T. (1945). *Pravopisen i pravogovoren narachnik*. Sofia, Bulgarska kniga.

Koeva, S. (1999). Pravila za prenasyane na chasti na dumata na nov red. *Bulgarski ezik*, 1:84–86.

Ladefoged, P. and Maddieson, I. (1996). *The sounds of the world's languages*. Oxford, Blackwell.

Liang, F. M. (1983). *Word Hy-phen-a-tion by Com-put-er (Hyphenation, Computer)*. Ph.D. thesis, Stanford University, Stanford, CA, USA. AAI8329742.

Murdarov, V. et al. (2012). *Oficialen pravopisen rechnik na bulgarskiya ezik*. Sofia, Prosveta.

Noncheva, V. (1988). Algoritam za avtomatichno prenasyane na dumi v balgarskiya ezik. In *Matematika i matematichesko obrazovanie. Sb. dokladi na 17 PK na SMB*, pages 479–482. Sofia, Izd. na BAN.

Pashov, P. (1989). *Prakticheska bulgarska gramatika*. Sofia, Narodna prosveta.

Stoyanov, S. (1993). *Gramatika na bulgarskiya knizhoven ezik*. Sofia, Universitetsko izdatelstvo "Sv. Kliment Ohridski".

Topalov, A. (1995). *Algoritmi i programi v tekstoobrabotkata*. Master's thesis, Sofia University, Faculty of Mathematics and Informatics. `http://www.mind-print.com/diploma/`.

Vasilev, V. (1997). *Ultimativniyat TeX. Udovolstvieto da pravim predpechatna podgotovka sami*. Sofia, Intela.

Zec, D. (1995). Sonority constraints on syllable structure. *Phonology*, 12:85–129.