CommonRuby - Feature #10085

Add non-ASCII case conversion to String#upcase/downcase/swapcase/capitalize

07/23/2014 11:04 AM - duerst (Martin Dürst)

Status:	Closed
Priority:	Normal
Assignee:	duerst (Martin Dürst)
Target version:	

Description

Case conversion functions are currently limited to ASCII characters. When used with formal languages, that may be appropriate, but it is often not appropriate for applications.

In order to avoid backwards-compatibility problems and to make sure that the various variants of case conversion (e.g. language-dependent) can be selected, we propose to add an optional parameter to the case conversion functions.

Our current design idea is as follows:

ASCII-only if no parameter:
'Türkiye'.upcase # => 'TüRKIYE', note lower-case ü

Parameter triggers (general) Unicode conversion:
'Türkiye'.upcase 'en' # => 'TÜRKIYE', note upper-case Ü

The parameter is actually a BCP 47 (http://tools.ietf.org/html/bcp47) language tag.
This means that for languages with special case conversion rules, such as Turkish, this works as follows:
'Türkiye'.upcase 'tr' # => 'TÜRKİYE', note upper-case İ (with dot!)

In the second example, we used 'en', but most other languages would work, too, because a single case conversion works for most languages. Turkic languages are the biggest exception.

The Unicode standard also defines various cases of "case-folding", which usually is lossy, e.g. mapping German ß to ss and so on. It should be possible to include this functionality in this proposal, e.g. by using :symbols or CONSTANTs for the few specific foldings. It may also be possible to define a reversible variant of case conversion in particular for use with swapcase.

In the long term, instead of a direct BCP 47 string, we could create a Locale class that would incorporate language-specific facilities, but this may need more detailed considerations.

The idea of using an additional parameter to indicate language-dependent or other processing variants should be extensible to areas such as number-to-string conversion and date formation. While this proposal is only about case conversion, we should check that there is a good chance to use similar parameter conventions for such extensions.

[This proposal is based on research done together with my student Kimihito Matsui.]

Related issues:

Related to Ruby trunk - Bug #3376: russian support	**Closed**	**06/01/2010**
Related to Ruby trunk - Feature #2034: Consider the ICU Library for Improving...	**Rejected**	
Related to Ruby trunk - Feature #10002: String swapcase	**Closed**	
Related to Ruby trunk - Feature #10152: String#strip doesn't remove non-break...	**Open**	**08/19/2014**
Related to Ruby trunk - Bug #10550: Resolv::DNS.getaddresses returns no IPs w...	**Closed**	**11/26/2014**
Has duplicate Ruby trunk - Bug #11284: String#upcase and String#downcase don'...	**Rejected**	

History

#1 - 07/23/2014 11:06 AM - duerst (Martin Dürst)

- Related to Bug #3376: russian support added

#2 - 07/23/2014 11:06 AM - duerst (Martin Dürst)

- Related to Feature #2034: Consider the ICU Library for Improving and Expanding Unicode Support added

#3 - 07/23/2014 11:07 AM - duerst (Martin Dürst)

- Related to Feature #10002: String swapcase added

#4 - 07/23/2014 11:10 AM - duerst (Martin Dürst)

- File CaseConversion.pdf added

#5 - 07/26/2014 08:47 AM - matz (Yukihiro Matsumoto)

- Assignee set to duerst (Martin Dürst)

I want default case conversion should be Unicode aware (when encoding is Unicode).
The previous behavior can be done by str.downcase(:ascii).

Non unicode encoding (e.g. Latin-1) can support non ASCII case conversion, but not mandatory.

Matz.

#6 - 08/20/2014 01:39 AM - matz (Yukihiro Matsumoto)

- Related to Feature #10152: String#strip doesn't remove non-breaking space added

#7 - 09/20/2014 05:44 AM - duerst (Martin Dürst)

- Target version set to Ruby 2.3.0

#8 - 12/31/2014 06:26 AM - duerst (Martin Dürst)

- Related to Bug #10550: Resolv::DNS.getaddresses returns no IPs when nameserver returns in differing case than query added

#9 - 12/31/2014 09:19 AM - akr (Akira Tanaka)

The related issue, [Bug #10550] Resolv::DNS.getaddresses, needs ASCII-only case conversion.
Unicode aware case conversion is not suitable for the issue.
See RFC 4343.

#10 - 06/19/2015 08:10 AM - duerst (Martin Dürst)

- Has duplicate Bug #11284: String#upcase and String#downcase don't work for accented characters added

#11 - 10/15/2017 11:21 PM - duerst (Martin Dürst)

- Status changed from Open to Closed

Close way overdue, should have happened somewhere around r55281.

Files

CaseConversion.pdf	340 KB	07/23/2014	duerst (Martin Dürst)