

Unicode basic properties and subproperties

Properties are in shaded cells, their subproperties are listed below them.

<code>\p{L*}</code>	<code>\p{letter}</code> Any letter.
<code>\p{Ll}</code>	<code>\p{lowercase_letter}</code>
<code>\p{Lu}</code>	<code>\p{uppercase_letter}</code>
<code>\p{Lt}</code>	<code>\p{titlecase_letter}</code> In some languages, digraphs have a special title-case form. InDesign matches Dz (01F2), Dž (01C5), Lj (01C8), Nj (01CB). Thus, “nj” has the forms <i>nj</i> , <i>Nj</i> , and <i>Nj</i> . InDesign also matches the Ancient Greek letters with “subscript iota”, as they can be written as a separate letter: At, Ht, Ot, and their variants with diacritics.
<code>\p{L&}</code>	Doesn’t work in InDesign. Covers <code>\p{Ll}</code> , <code>\p{Lu}</code> , and <code>\p{Lt}</code> .
<code>\p{Lm}</code>	<code>\p{modifier_letter}</code> Various characters from Spacing modifier letters (02B0-02FF) (seems to miss several).
<code>\p{Lo}</code>	<code>\p{letter_other}</code> Whatever letters not captured by the four <code>\p{L.}</code> classes, i.e. letters without case and that aren’t modifiers: characters from Hebrew, Arabic, the SE-Asian languages, etc.
<code>\p{M*}</code>	<code>\p{mark}</code> Any of the following three types of mark.
<code>\p{Mn}</code>	<code>\p{non_spacing_mark}</code> Including combining diacritical marks and tone marks. Matches characters in a wide variety of ranges.
<code>\p{Mc}</code>	<code>\p{spacing_combining_mark}</code> Vowels in SE-Asian languages.
<code>\p{Me}</code>	<code>\p{enclosing_mark}</code> Circles, squares, keycaps, etc. Found in a variety of Unicode ranges.
<code>\p{Z*}</code>	<code>\p{separator}</code> Spaces, returns, 2028, 2029 (but not hyphens and dashes).
<code>\p{Zs}</code>	<code>\p{space_separator}</code> All spaces except tab and return.
<code>\p{Zl}</code>	<code>\p{line_separator}</code> 2028 is the line-separator character.
<code>\p{Zp}</code>	<code>\p{paragraph_separator}</code> 2029
<code>\p{S*}</code>	<code>\p{symbol}</code> (Math, wingdings) The full form <code>\p{Symbol}</code> works fine, the short form <code>\p{S}</code> matches separators.
<code>\p{Sm}</code>	<code>\p{math_symbol}</code> Math symbols.
<code>\p{Sc}</code>	<code>\p{currency_symbol}</code> All currency symbols.

<code>\p{Sk}</code>	<code>\p{modifier_symbol}</code> Combining characters with their own width, such as the acute 00B4 (not acute 0301).
<code>\p{So}</code>	<code>\p{other_symbol}</code> Wingdings, dingbats, etc. from various ranges.
<code>\p{N*}</code>	<code>\p{number}</code> Any kind of number.
<code>\p{Nd}</code>	<code>\p{decimal_digit_number}</code> The digits 0 to 9.
<code>\p{NI}</code>	<code>\p{letter_number}</code> The Roman upper- and lower-case numerals in Number forms (2150–218F).
<code>\p{No}</code>	<code>\p{other_number}</code> Super- and subscripts, fractions, enclosed numbers in Latin 1, Number forms, and enclosed alphanumerics.
<code>\p{P*}</code>	<code>\p{punctuation}</code> Any punctuation.
<code>\p{Pd}</code>	<code>\p{dash_punctuation}</code> All hyphens and dashes.
<code>\p{Ps}</code>	<code>\p{open_punctuation}</code> Opening brackets, braces, parentheses, and similar, e.g. 2045, FE17, and FF62.
<code>\p{Pe}</code>	<code>\p{close_punctuation}</code> Closing brackets, braces, parentheses, and similar, e.g. 2046, FE18, and FF63.
<code>\p{Pi}</code>	<code>\p{initial_punctuation}</code> All opening quotes.
<code>\p{Pf}</code>	<code>\p{final_punctuation}</code> All closing quotes.
<code>\p{Pc}</code>	<code>\p{connector_punctuation}</code> underscore, 203F, 2040, 2054.
<code>\p{Po}</code>	<code>\p{other_punctuation}</code> All other punctuation: ! " % &, etc.
<code>\p{C*}</code>	<code>\p{other}</code> What it says: ‘other’.
<code>\p{Cc}</code>	<code>\p{control}</code> Control characters in C0 Controls and Basic Latin (0000–0020), such as Tab, Esc, etc.
<code>\p{Cf}</code>	<code>\p{format}</code> Various (non-visible) formatting characters in General Punctuation (2000–206F), such as left-to-right and right-to-left markers, embedding, etc.
<code>\p{Co}</code>	<code>\p{private_use}</code> (E000–F8FF)
<code>\p{Cn}</code>	<code>\p{unassigned}</code> Some of the unassigned unicode ranges (e.g. D7A4–D7FF).

Notes

- For details, see the Boost web site at <http://tinyurl.com/ck9xe5> and <http://tinyurl.com/amenz5>. See also J. Friedl, *Mastering Regular Expressions*, O’Reilly, 2006, pp. 122, 123.
- The first column gives the short forms, the second column, the long forms.
- Use upper-case P for negated classes: `\P{L*}` matches everything that is not a letter.

- Both forms (short and long) are uncharacteristically lenient in that any spacing and capitalisation can be used. `\p{UL}`, `\p{Ul}`, and `\p{ul}` work equally well, as do `\p{uppercase_letter}`, `\p{uppercase letter}`, and `\p{uppercaseletter}` and all case variants.
- InDesign won’t let you use unicode properties in character classes. Unfortunately, constructs such as `[\p{Ps}\p{Pi}]` don’t work. The workaround is to use alternatives: `\p{Ps}\p{Pi}`.