Internationalized semantic punctuation framework

115:2025, Version 1

115:2025, Version 1

Ribose

Contents

Warning for Drafts

This document is not a Ribose Standard. It is distributed for review and comment, and is subject to change without notice and may not be referred to as a Standard. Recipients of this draft are invited to submit, with their comments, notification of any relevant patent rights of which they are aware and to provide supporting documentation.

Ribose Asia Limited

Suite 1, 8/F, New Henry House
10 Ice House Street
Central
Hong Kong

copyright@ribose.com
www.ribose.com

Foreword

This document is part of the Metanorma specifications series that defines requirements for internationalized document rendering, specifically addressing punctuation usage across multiple languages and writing systems.

Introduction

Metanorma is a semantic authoring system that supports multiple languages and the publication of multi-lingual output.

A major component of Metanorma architecture is dedicated to providing presentation rendering for the semantic content it encodes. This means that Metanorma often composes punctuated text to represent semantic text, and that punctuation needs to follow the norms of a target language. Metanorma requires a semantic framework to capture the various types of punctuation it will add to the text it composes, and to provide well-governed rendering of that punctuation in supported languages.

This document defines a framework for semantic punctuation usage across multiple languages and writing systems, to be drawn from in the generation of punctuated text based on semanticly marked up text. This framework is not intended to supplant the punctuation provided by authors in originally authored texts.

The document covers the following languages and writing systems:

- CJK (East Asian ideographic writing systems)
  - Traditional Chinese (Taiwan/Hong Kong conventions)
  - Simplified Chinese (Mainland China conventions)
  - Japanese (including vertical text layout)
  - Korean (Hangul-specific rules)
- Latin-script languages (English, Spanish, French)
- Cyrillic-script languages (Russian)

The document defines punctuation marks through primary usage categories based on the semantic function of the punctuation marks.

This document addresses critical challenges including:

Internationalization of auto-numbered elements	Applying appropriate punctuation to document elements, such as: phrases, lists, figures, tables and references use. Enabling definition of Label Auto-assignment Definition Language profiles (LADL) (MN 112).
Internationalization of bibliographic citations	Applying appropriate punctuation to in-text citations and reference lists, allowing for citation style definition and language-specific variations.
Automatic spacing	Applying appropriate whitespace around punctuation marks based on language-specific rules.
Rendering in vertical text layouts	Proper punctuation positioning and transformation in vertical text modes.

1. Scope

This document defines a framework for semantic punctuation usage across multiple languages and writing systems. It also outlines how that framework is realised in Metanorma implementation, and how its configuration can be updated.

Examples and discussion about specific writing systems are as of this writing limited to CJK, Latin and Cyrillic; but the document is intended to be generally applicable.

Formal notation systems that use punctuation in ways divergent from normal language, such as mathematical notation, chemical notation, and computer programming, are out of scope of this document. Numeric punctuation is in scope, but it is not discussed in detail, as it is handled in Metanorma by distinct mechanisms.

NOTE Following linguistic practice, and to avoid confusion with quotations, individual characters are cited in angle brackets, and where necessary followed by Unicode codepoint, and preceded by Unicode name; e.g.

- < , >
- < , > (U+FF0C)
- FULLWIDTH COMMA < , > (U+FF0C).

2. Normative references

The following documents are referred to in the text in such a way that some or all of their content constitutes requirements of this document. For dated references, only the edition cited applies. For undated references, the latest edition of the referenced document (including any amendments) applies.

MN 112, MN 112: *Label Auto-assignment Definition Language (LADL) specification*

ISO/IEC 10646:2020, International Organization for Standardization (committee). *Information technology — Universal coded character set (UCS)*. Sixth edition. 2020. Geneva: International Organization for Standardization and International Electrotechnical Commission. https://www.iso.org/standard/76835.html.

Unicode 15.1, *Unicode* Edition 15.1. https://www.unicode.org

3. Terms and definitions

For the purposes of this document, the following terms and definitions apply.

3.1. script PREFERRED

set of symbols representing a particular language

3.2. writing system PREFERRED

script, and the rules whereby it represents a particular language

3.3. locale PREFERRED

regional variant of a language and/or script, specific to a country or region

3.4. morpheme PREFERRED

semantic component of a language

3.5. ideograph PREFERRED

written character that represents a morpheme, rather than a sound in a language

3.6. CJK PREFERRED

Chinese, Japanese, Korean, and other writing systems based on East Asian ideographs

3.7. internationalisation PREFERRED
i18n ADMITTED

support in software for input and output in more than one language

3.8. punctuation mark PREFERRED

symbol used in written language to clarify meaning, indicate text structure, separate elements, or convey syntactic relationships

3.9. usage category PREFERRED

functional classification of punctuation marks based on their primary purpose in text structure

EXAMPLE: Sentence delimiters, pause indicators, grouping markers.

3.10. full-width character PREFERRED

character that occupies the width of an ideographic character, commonly used in CJK typography

3.11. half-width character PREFERRED

character that occupies standard Latin character width, used in Western typography and mixed CJK text

3.12. directionality PREFERRED

text arrangement of characters, including left-to-right (LTR), used in Latin and Cyrillic, right-to-left (RTL), used in Hebrew and Arabic, horizontal (identical to LTR), usually used in CJK, and vertical (top-to-bottom in columns arranged from right to left), traditionally used in CJK

3.13. spacing rule PREFERRED

specification for whitespace insertion around punctuation marks that varies by language, locale, and context

3.14. phrase PREFERRED

meaningful grouping of words, that is smaller than a sentence

4. Framework

4.1. General

This framework provides guidelines for the consistent application of punctuation marks across different languages and writing systems. It aims to ensure that punctuation usage aligns with semantic meaning and cultural conventions.

Punctuation profile

A set of rules and mappings that define how punctuation marks are applied in a specific language, locale or document context. Even in the same language and locale, different publishers may require different punctuation styles.

EXAMPLE 1: In Japanese, the 「公用文作成の考え方」（建議）（付）「公用文作成の考え方（文化審議会建議）」解説 of the Agency for Cultural Affairs (2022) specifies different punctuation rules from JIS X 8301:2019, the Japanese Industrial Standard stating requirements for punctuation in Japanese standards

- sentence pausal mark: the former requires use of the *ten*, IDEOGRAPHIC COMMA < 、> (U+3001), but the latter requires the use of the full-width Latin comma, FULLWIDTH COMMA < ，> (U +FF0C).

Punctuation mark

A symbol used in writing to clarify meaning, indicate text structure, separate elements, or convey syntactic relationships. Each punctuation mark can have several semantic functions, and therefore belong to multiple usage categories; these may vary across languages. Not all punctuation marks exist in all languages. A mark is commonly represented by one or more glyphs.

Semantic function

The purpose or role of a punctuation mark in text, which may include indicating sentence and phrase boundaries, grouping, emphasis, indicating relationships, and semantic categorisation such as for numerals. Semantic functions are grouped as usage categories. The correspondence of semantic function to punctuation mark is many-to-many.

EXAMPLE 2: The semantic function of “minor phrase separator within a sentence” is part of the “phrase stop” usage category, and is represented by different punctuation marks in different languages:

- Japanese: IDEOGRAPHIC COMMA < 、> (U+3001)
- Chinese: FULLWIDTH COMMA < ，> (U+FF0C)
- English: COMMA < , > (U+002C)

Not all semantic functions have distinct punctuation marks expressing them in all languages.

EXAMPLE 3: The semantic function of "enumeration delimiter", used in Traditional Chinese and Simplified Chinese, does not exist in English or Japanese as a separate punctuation mark. It is instead conflated with the punctuation mark for minor phrase separator, the comma.

The correspondence of semantic function to punctuation mark is many-to-many.

EXAMPLE 4: In English, a period can be used as a declarative sentence stop (5.1.2), but also as an abbreviation mark (7.3.2) (*Bros.*, *p.m.*). So English period is a punctuation mark fulfilling two semantic functions. In Hebrew, there are two abbreviation marks, *geresh* <#> (U+05F3) for a single word (ר׳ "r." = רבי rabbi), and *gershayim* <#> (U+05F4) for multi-word phrases (ארה״ב "U.S." = ארצות הברית "United States"); these are distinct from the sentence delimiter:

> ר׳ יעקב גר בארה״ב. וגם אני, "R(abbi) Yaakov lives in the U.S. And so do I".

So the abbreviation mark function is conflated with the sentence delimiter function into the period punctuation mark in English, whereas it is split into a single-word and a multi-word punctuation mark in Hebrew.

The *geresh* also has other semantic functions, such as indicating numerals, or as a diacritic for non-Hebrew sounds (גם /ɡam/ "also" vs ג׳ים /dʒam/ "jam"). As a diacritic, the *geresh* is no longer punctuation, but represents phonological sounds.

Punctuation rule

A guideline that specifies how a punctuation mark should be used in a particular language or context, including placement, spacing, and variations. Some punctuation rules are completely predictable, and can be automated. Other punctuation rules are not predictable, and need to be applied manually and on an idiosyncratic basis.

EXAMPLE 5: The semantic function of range in Chinese differentiates between word ranges, which use EN DASH <–> (U+2013) (1月—7月 "January to July", literally "Month 1–Month 7"), and numeric ranges, which use WAVE DASH <〜> (U+301C), (5〜20個字 "5 to 20 words"). This punctuation rule is predictable: wave dash is used between two numerals (including Chinese numerals).

EXAMPLE 6: In Hebrew, *geresh* follows the initial letters of a word as an abbreviation mark in Hebrew, but the abbreviation may involve just the first letter (ר׳ "r." = רבי "rabbi"), or more than one letter (גב׳ "Mrs, Ms" = גברת "lady"). It is not predictable how many letters appear

in a Hebrew abbreviation: implementing Hebrew abbreviation derived from full words, in a semantic punctuation model, involves either a lookup table of abbreviations, or providing the already abbreviated words as input. (The latter is what is already done by typing גב׳, just as it is in English by typing *Mrs* instead of a directive like `abbreviate("Mistress")`.)

A punctuation profile is defined by:

- A set of semantic functions that apply to semantic elements in text
- Punctuation marks that are associated with each semantic function
- A collection of predictable punctuation rules that govern their usage

Given a punctuation profile, a document rendering with appropriate punctuation that suits the target audience can be achieved—provided that the punctuation rules can be automated. Not all punctuation rules described here can be so automated, and they will be addressed by being applied manually in authored text.

4.2. Punctuation semantic functions

Punctuation marks are assigned semantic functions by their primary functional usage rather than their linguistic origin.

This approach enables consistent implementation across different languages, while accommodating language-specific variations within each usage category.

Under each usage category of semantic functions, each distinct function in the framework is given, along with its typical corresponding marks in Latin (exemplified by English), Cyrillic, and CJK.

Each usage category and each individual function defines one of more of the following:

- Primary function and purpose
- Range of semantic functions covered, including potential ambiguity with other usage categories
- Common punctuation mark variations across languages for each function
- Spacing requirements
- Special handling rules

The semantic functions described here reflect usage, and punctuation usage does not follow rigorously differentiated functions: similar functions are routinely conflated in punctuation marks, and the same function is routinely represented by different punctuation marks, with little rigour. This framework is concerned with contexts where the intended meaning of punctuation can be controlled by the publishing system, in deriving punctuated, template-generated text from underlying semantic markup. Such templates can specify the intended function of punctuation more rigorously than it is reasonable to expect of a human author.

The mapping of semantic functions to punctuation marks is often idiosyncratic, regional, and subject to fashions, as well as to disagreements between authorities as fashions change. This is particularly notable with ongoing changes in English punctuation noted in this document, as well as with the differences between American English and British English practice.

Differences in practice between languages often reflect lags in change; for instance, Greek uses the same ditto mark as Quebec French (8.2.2), not because Greece was in contact with Canada, but because both have preserved older France French practice, which has since been abandoned in France itself. This is also reflected in the formerly more widespread practice of French spacing (5.1.1.3), and in Japanese practice matching Traditional Chinese practice rather than Simplified Chinese practice.

Not all semantic functions described here are expected in the kinds of documents that Metanorma processors, but they are included for completeness. Semantic functions that are not expected to be supported by Metanorma are flagged in the following as "(out of scope)".

5. Functions: phrase-level structure

5.1. Sentence stops

5.1.1. General

5.1.1.1. Primary function and purpose

Sentence stops delimit a full grammatical sentence.

5.1.1.2. Range of semantic functions

Writing systems vary as to whether they also terminate a standalone sentence, and therefore act as a sentence terminator. That is the case in formal English (though often not in texting), but it is not the case in Japanese.

5.1.1.3. Spacing rules

- Latin and Cyrillic require space after sentence stops as delimiters between sentences, as does Korean.
- Writing systems do not require space after the final sentence stop in a paragraph.
- Chinese and Japanese by default do not use space after sentence stops.
- None of the writing systems in scope of this framework has space before sentence stops.
- In French in France, Switzerland and Belgium, some punctuation marks, including some sentence stops, are preceded by a non-breaking thin space (U+202F) ("French

spacing"). In common practice, a full non-breaking space (U+00A0) is used instead. French spacing is not used for most punctuation marks in Canada for French.

5.1.1.4. Special handling

- Korean sentence stops are half-width.
- Chinese and Japanese sentence stops are full-width.
- In text mixed between Latin and Chinese or Japanese, usual practice is to follow the document main language. Fine typography will apply kerning in such contexts so that the switch in width does not look obtrusive.

  EXAMPLE 1 — Script switch with Chinese as main language: 现在很多人都在用 iPhone。(with full-width punctuation in a Chinese text after English words)

  EXAMPLE 2 — Script switch with English as main language: This app 很好用.

- Some desktop publishing applications in Japanese and CJK OpenType fonts allow switching contextually between full-width and half-width punctuation (欧文混在時の約物処理 "treat punctuation when mixed with Western text").

5.1.2. Declarative stop (*period*)

5.1.2.1. Primary function and purpose

Mark that indicates the end of a complete declarative grammatical statement. This function is represented in Metanorma i18n files as `punct.period`.

EXAMPLE 1: English: the period in *This is a sentence.*

EXAMPLE 2: Traditional Chinese: the ideographic period in 這是一句句子。

5.1.2.2. Punctuation mark in scripts, languages, and locales

- Latin, Cyrillic: FULL STOP < . > (U+002E)
- Traditional Chinese: IDEOGRAPHIC FULL STOP < 。> (U+3002)
- Simplified Chinese: IDEOGRAPHIC FULL STOP < 。> (U+3002)
- Japanese: IDEOGRAPHIC FULL STOP < 。> (U+3002)
- Korean: FULL STOP < . > (U+002E)

5.1.2.3. Special handling

- French spacing does NOT apply (5.1.1.3).
- The placement of period in CJK is different by directionality, but this does not involve different Unicode codepoints, so the browser/operating system is assumed to handle positioning appropriately. (While there is a PRESENTATION FORM FOR VERTICAL IDEOGRAPHIC FULL STOP < ﹒> (U+FE12), this is only included in Unicode as a compatibility character.)

— In Japanese and Simplified Chinese horizontal text, the period appears at the bottom right. In Japanese and Simplified Chinese vertical text, the period appears below and to the right of the character.

— In Traditional Chinese, the period appears at mid-height in both horizontal and vertical text orientations.

5.1.3. Interrogative stop (*question mark*)

5.1.3.1. Primary function and purpose

Mark that indicates this grammatical statement is a question. This function is represented in Metanorma i18n files as `punct.question-mark`.

EXAMPLE 1: English: the question mark in *Is this a question?*

EXAMPLE 2: Traditional Chinese: the ideographic question mark in 這是問題嗎？

5.1.3.2. Punctuation mark in scripts, languages, and locales

- Latin, Cyrillic: QUESTION MARK <?> (U+003F)
- Traditional Chinese: FULLWIDTH QUESTION MARK < ? > (U+FF1F)
- Simplified Chinese: FULLWIDTH QUESTION MARK < ? > (U+FF1F)
- Japanese: FULLWIDTH QUESTION MARK < ? > (U+FF1F)
- Korean: QUESTION MARK <?> (U+003F)

5.1.3.3. Spacing rules

- French spacing applies (5.1.1.3).

5.1.3.4. Special handling

- In Spanish and languages under Spanish cultural influence (e.g. Catalan), INVERTED QUESTION MARK <¿> (U+00BF) is used at the start of an interrogative sentence or phrase. If a declarative sentence contains an interrogative phrase, only the interrogative phrase is so delimited:

  EXAMPLE: *Si no puedes ir con ellos, ¿quieres ir con nosotros?* "If you cannot go with them, ¿would you like to go with us?"

  — There is some variation by locale in usage: short questions can omit the inverted question mark in Galician; Catalan in Catalonia does not use inverted question marks; Catalan in Valencia uses them optionally.

> *Metanorma does not currently implement inverted question marks.*

5.1.4. Exclamatory stop (*exclamation mark*) (out of scope)

5.1.4.1. Primary function and purpose

Mark that indicates this grammatical statement is an exclamation. This function is represented in Metanorma i18n files as `punct.exclamation-mark`.

EXAMPLE 1: English: the exclamation mark in *What a great day!*

EXAMPLE 2: Traditional Chinese: the ideographic exclamation mark in 多麼美好的一天！

5.1.4.2. Punctuation mark in scripts, languages, and locales

- Latin, Cyrillic: EXCLAMATION MARK < ! > (U+0021)
- Traditional Chinese: FULLWIDTH EXCLAMATION MARK < ！ > (U+FF01)
- Simplified Chinese: FULLWIDTH EXCLAMATION MARK < ！ > (U+FF01)
- Japanese: FULLWIDTH EXCLAMATION MARK < ！ > (U+FF01)
- Korean: EXCLAMATION MARK < ! > (U+0021)

5.1.4.3. Spacing rules

- French spacing applies (5.1.1.3).

5.1.4.4. Special handling

- In Spanish and languages under Spanish cultural influence (e.g. Catalan), INVERTED EXCLAMATION MARK < ¡ > (U+00A1) is used at the start of an exclamatory sentence or phrase.
- The same locale considerations apply for Catalan as for inverted question mark.

> *Metanorma does not currently implement inverted exclamation marks.*

5.2. Phrase stops

5.2.1. General

5.2.1.1. Primary function and purpose

Phrase stops delimit phrases within a sentence.

5.2.1.2. Special handling

CJK full-width punctuation (5.1.1.4) applies.

5.2.2. Minor phrase separator (*comma*)

5.2.2.1. Primary function and purpose

Mark that delimits a minor break between phrases within a sentence. This function is represented in Metanorma i18n files as `punct.comma`.

EXAMPLE 1: English: the comma in *This is a sentence, with a pause.*

EXAMPLE 2: Traditional Chinese: the full-width comma in 這是一句句子，有停頓。

5.2.2.2. Range of semantic functions

- The contexts in which the minor break is marked between phrases vary significantly by language and by style.
- The types of phrase eligible to be so marked also vary significantly by language and by style; it usually includes both clauses (which express a complete predicate, including both a subject and a verb) and smaller units such as noun phrases, or successive adjectives.

  EXAMPLE: *Ich weiß, dass du lügst* "I know, that you're lying", with comma separating a complement clause from the main clause, is correct punctuation in German, but not English

- In many writing systems, the same mark (comma) is used for minor phrase separators and other separating functions, such as enumerator delimiters (5.3.2), and decimal point (8.1.2).

5.2.2.3. Punctuation mark in scripts, languages, and locales

- Latin, Cyrillic: COMMA < , > (U+002C)
- Traditional Chinese: FULLWIDTH COMMA < ， > (U+FF0C)
- Simplified Chinese: FULLWIDTH COMMA < ， > (U+FF0C)
- Japanese: IDEOGRAPHIC COMMA < 、> (U+3001)
  - In mixed Japanese-Western text, FULLWIDTH COMMA < ， > (U+FF0C) can be used instead to maintain visual consistency
  - IDEOGRAPHIC SPACE <　> (U+3000) can be used in Japanese corresponding to English comma or colon usage in certain contexts.
- Korean: COMMA < , > (U+002C)

5.2.2.4. Special handling

The placement of comma in CJK is different by directionality, but this does not involve different Unicode codepoints, so the browser/operating system is assumed to handle positioning appropriately. (While there is a PRESENTATION FORM FOR VERTICAL COMMA < ﹐> (U+FE10) and a PRESENTATION FORM FOR VERTICAL IDEOGRAPHIC COMMA < ﹑> (U+FE11), these are only included in Unicode as a compatibility character.)

- In Japanese and Simplified Chinese horizontal text, the comma appears at the bottom right. In Japanese and Simplified Chinese vertical text, the comma appears below and to the right of the character.
- In Traditional Chinese, the comma appears at mid-height in both horizontal and vertical text orientations.

5.2.3. Major phrase separator (*semicolon*)

5.2.3.1. Primary function and purpose

Mark that delimits a major break between phrases within a sentence. This function is represented in Metanorma i18n files as `punct.semicolon`.

EXAMPLE 1: English: the semicolon in *This is a sentence; with a pause.*

EXAMPLE 2: Traditional Chinese: the full-width semicolon in 這是一句句子；有停頓。

5.2.3.2. Range of semantic functions

- The contexts in which the major break is marked between phrases varies significantly by language and by style. In many styles, it is avoided as an overly fine distinction from the minor break (comma).
- The phrases separated by a major phrase separator are typically independent clauses grammatically (i.e. complete predicates, with both a subject and a verb).

5.2.3.3. Punctuation mark in scripts, languages, and locales

- Latin, Cyrillic: SEMICOLON < ; > (U+002C)
- Traditional Chinese: FULLWIDTH SEMICOLON < ； > (U+FF1B)
- Simplified Chinese: FULLWIDTH SEMICOLON < ； > (U+FF1B)
- Japanese: FULLWIDTH SEMICOLON < ； > (U+FF1B)
- Korean: SEMICOLON < ; > (U+002C)

5.2.3.4. Spacing rules

- French spacing applies (5.1.1.3).

5.2.3.5. Special handling

The placement of semicolon in CJK is different by directionality, but this does not involve different Unicode codepoints, so the browser/operating system is assumed to handle positioning appropriately. (While there is a PRESENTATION FORM FOR VERTICAL SEMICOLON < ； > (U+FE14), this is only included in Unicode as a compatibility character.)

5.2.4. Introductory phrase separator (*colon*)

5.2.4.1. Primary function and purpose

Mark that delimits phrases within a sentence, where the first phrase introduces the second. This function is represented in Metanorma i18n files as `punct.colon`.

5.2.4.2. Range of semantic functions

- Sometimes punctuation is used to separate a term from a description in a definition list, although the default is to use only indented space. This function can be regarded as an extension of the introductory phrase separator. In Metanorma Presentation XML, this is notated as `<span class="fmt-dt-delim">`.

EXAMPLE

Framework *basic structure underlying a system, concept, or text*

Framework: *basic structure underlying a system, concept, or text*

5.2.4.3. Punctuation mark in scripts, languages, and locales

- Latin: COLON < : > (U+003A)
- Cyrillic: COLON < : > (U+003A)
- Traditional Chinese: FULLWIDTH COLON < ： > (U+FF1A)
- Simplified Chinese: FULLWIDTH COLON < ： > (U+FF1A)
- Japanese: FULLWIDTH COLON < ： > (U+FF1A)
  - IDEOGRAPHIC SPACE <　> (U+3000) can be used in Japanese corresponding to English comma or colon usage in certain contexts.
- Korean: COLON < : > (U+003A)

5.2.4.3.1. Spacing rules

- In French, colon is preceded by a non-breaking space. There is more variation by locale of this spacing rule than for other punctuation marks, for which French spacing applies (5.1.1.3).
  - In Switzerland, it is thin space (U+202F).
  - In Canada, colon is the only punctuation mark where French spacing is expected, and it is a full space (U+00A0).
  - In France, a full space (U+00A0) is usual.

5.2.4.4. Special handling

The placement of colon in CJK is different by directionality, but this does not involve different Unicode codepoints, so the browser/operating system is assumed to handle positioning appropriately. (While there is a PRESENTATION FORM FOR VERTICAL COLON < ： > (U+FE13), this is only included in Unicode as a compatibility character.)

The interpunct (5.3.2.3) is used instead of hyphen, dash, or colon in Japanese vertical text.

5.2.5. Breaking phrase separator (*em-dash*)

5.2.5.1. Primary function and purpose

Mark that delimits phrases within a sentence, where there is a conceptual break of some sort between the first phrase and the second. This function is represented in Metanorma i18n files as `punct.em-dash`.

5.2.5.2. Range of semantic functions

- An interrupted ending to a sentence can be indicated by a breaking phrase separator. Similarly, a sentence start which counts as a resumption from a previous interruption can be indicated by a breaking phrase separator.

  EXAMPLE: *He was the miracle ingredient Z-147. He was—*
  "Crazy!" Clevinger interrupted, shrieking. "That's what you are! Crazy!"
  "—immense. I'm a real, slam-bang, honest-to-goodness, three-fisted humdinger. I'm a bona fide supraman." — Joseph Heller, *Catch-22*

  - Interrupted and resumed sentences are characteristic of literary prose, but not formal prose, such as is in scope of Metanorma.

5.2.5.3. Punctuation mark in scripts, languages, and locales

- Latin, Cyrillic: EM DASH <—> (U+2014)
  - The EN DASH <–> (U+2013) is also used in this function
  - Double and triple hyphens are typewriter approximations of em-dashes, and are commonly used in word processing as easy ways to data enter em-dashes through auto-text; they are not normally expected to display as such in finished documents, although they remain a convention of comic strips.
  - In some style guides (e.g. the *Australian Government Style Manual*), interrupted and resumed sentence breaks are notated with two em-dashes, as a distinct function from the breaking phrase separator.
- Traditional Chinese: TWO EM DASH <⸺> (U+2E3A)
- Simplified Chinese: TWO EM DASH <⸺> (U+2E3A)
- Japanese: TWO EM DASH <⸺> (U+2E3A)
  - Informally in CJK, twice em-dash (U+2014 U+2014) is used instead of the two em-dash.
- Korean: EM DASH <—> (U+2014)

5.2.5.3.1. Spacing rules

- There is variation between languages and locales as to whether em-dash or en-dash is used. Spaces surrounding the dash are required for en-dash, and may or may not

be required for em-dash. The spaces are thin spaces in fine typography, but may be normal spaces in common practice.

5.2.5.4. Special handling

- The placement of em-dash in CJK is different by directionality, but this does not involve different Unicode codepoints, so the browser/operating system is assumed to handle positioning appropriately. (While there is a PRESENTATION FORM FOR VERTICAL EM DASH < | > (U+FE31), this is only included in Unicode as a compatibility character.)
- The interpunct (5.3.2.3) is used instead of hyphen, dash, or colon in Japanese vertical text.
- An interrupted ending to a sentence can be indicated by a breaking phrase separator. In that case, the breaking phrase separator replaces the sentence stop.

  EXAMPLE: *He was the miracle ingredient Z-147. He was—*
  "Crazy!"

5.2.6. Hesitancy phrase separator (out of scope)

5.2.6.1. Primary function and purpose

Mark that delimits phrases within a sentence, where there is some sort of hesitation between them.

5.2.6.2. Range of semantic functions

- The hesitancy phrase separator is closely related to the breaking phrase separator, and in past practice was conflated with it.
- The hesitancy phrase separator is routinely conflated with the missing text mark (8.2.1), but the two interact differently with other punctuation.
- "Hesitancy" is to be broadly understood, and it includes pauses, interruption, speechlessness, deliberate silence, longing, and surprise. Different languages place different expectations on the punctuation mark.
- The hesitancy phrase separator is associated with emotion, and is therefore avoided in formal writing, such as is in scope of Metanorma. The missing text mark, on the other hand, is entirely consistent with formal writing.

5.2.6.3. Punctuation mark in scripts, languages, and locales

- Latin, Cyrillic: in fine typography, a single HORIZONTAL ELLIPSIS <…> (U+2026) is used. In informal practice, three periods in succession are used as well.
  - In some fine typographic practice, the em dash for the breaking phrase separator is used instead.
  - In some fine typographic practice, the three periods are interpolated with non-breaking spaces, or thin non-breaking spaces.

- CJK: the ellipse can be presented as three dots, or six dots.
  - The six dot form is entered as twice the full-width three-dot form.
  - The CJK form is either MIDLINE HORIZONTAL ELLIPSIS <⋯> (U+22EF), as explicit centered markup, or the half-width HORIZONTAL ELLIPSIS (U+2026) where centering is inexplicit, rendered in CJK fonts as full-width.
  - Japanese rarely also uses a two-dot ellipse (*rīdā*).

5.2.6.3.1. Spacing rules

- Languages and style guides vary as to whether they put spaces before or after the hesitancy phrase separator, and between the ellipse dots within the hesitancy phrase separator—and when. For example, the Modern Language Association styleguide has recently changed its recommendation from space before in all contexts, to no space before in the phrase separator function.
- In French, French spacing applies (5.1.1.3).

5.2.6.4. Special handling

- If the hesitation follows the final phrase in a sentence, some practice has the hesitancy mark replace the sentence stop (e.g. common British English practice). Other practice appends the sentence stop to the hesitancy mark, as a fourth dot (e.g. *Chicago Manual of Style*).

  EXAMPLE: *I like traffic lights...*
  I like traffic lights... .

- Russian combines not only declarative but interrogative and exclamatory sentence stops with the hesitancy mark and missing text mark: <? . . >, <! . . >.
- Unlike other stops, the hesitation phrase separator can appear at the start of a sentence.
- In horizontal directonality in Traditional Chinese and Japanese, the ellipse is vertically centered; in vertical directionality, it is horizontally centered.
- In Simplified Chinese, the ellipse is usually aligned to the baseline in horizontal directionality; in vertical directionality, it is still horizontally centered.
- In vertical directionality in CJK, Unicode has a codepoint for VERTICAL ELLIPSIS <#> (U+22EE), and it is not designated as a compatibility character. However as with other punctuation, the browser/operating system is assumed to handle positioning appropriately, so MIDLINE HORIZONTAL ELLIPSIS would still be expected to be entered.

5.2.7. Verse separator (out of scope)

5.2.7.1. Primary function and purpose

Mark that delimits verses in poetic writing.

5.2.7.2. Punctuation mark in scripts, languages, and locales

- In verses presented as poetry, verses are globally separated by line breaks, which are not regarded as punctuation.
- In prose transcription of poetry, verses are often separated by a slash: SOLIDUS </> (U +002F).
  - There is some usage of VERTICAL LINE <|> (U+007C) instead.

EXAMPLE 1 — Poetic typesetting: *To be, or not to be, that is the question:*
Whether 'tis Nobler in the mind to suffer
The Slings and Arrows of outrageous Fortune,
Or to take Arms against a Sea of troubles,
And by opposing end them…

EXAMPLE 2 — Prose typesetting: *To be, or not to be, that is the question: / Whether 'tis nobler in the mind to suffer / The slings and arrows of outrageous Fortune, / Or to take arms against a sea of troubles, / And by opposing end them…*

5.2.7.3. Spacing rule

- When slash is used to separate verses, there is a space either side of the slash.

5.2.7.4. Special handling

In poetic presentation (one verse per line), separate verses must not be right-justified at the line-break.

5.2.8. Section separator (out of scope)

5.2.8.1. Primary function and purpose

The section separator includes a break in a document, at a higher level than a sentence or paragraph. This is distinct from a clause heading, which introduces a new section of text, with a number and/or a title.

EXAMPLE: *…Randy looked at this watch. He was sweating buckets.*
"If the otter police don't get here soon, we're in deep trouble," he said.

James nodded. "I'm almost out of kibble. And the bits are running low too."

* * *

In the blimp floating high over Ferretsburg, Captain Crandle looked down at the unfolding battle with growing annoyance….

— https://greenwalledtreehouse.com/2022/03/05/writing-corner-the-dinkus/

5.2.8.2. Range of semantic functions

- In formal documents of the type considered by Metanorma, the only permitted breaks at a document level higher than a paragraph are clauses, and new clauses are explicitly indicated by clause headings. Section separators are characteristic of extended literary prose, such as novels; even though novels have chapters, they do not have subchapters, and breaks in a chapter are indicated by a section separator instead.
  - In literary use, the section separator is intended to convey a logical or emotional break.
- There is overlap between the section separator and the missing text mark (8.2.1), when the scope of the missing text mark is at the paragraph level.

5.2.8.3. Punctuation mark in scripts, languages, and locales

- The cover term for the traditional punctuation mark for a section separator, used in literary writing, is the dinkus. The usual contemporary form of the dinkus is three asterisks or three bullets in a row, with space between them.
  - Older practice used other decorative symbols, including ASTERISM <⁂> (U+2042) and various fleurons—such as ROTATED FLORAL HEART BULLET <❧> (U+2767) and NORTH WEST POINTING LEAF <🙰> (U+1F650).
- In word processing and HTML documents, this function is conveyed by the horizontal rule (`<hr/>`), which is not regarded as punctuation.
- The Japanese PART ALTERNATION MARK <〽> (U+303D) has narrower application, being used in the Renga genre of poetry to indicate the start of a song. The LEFT CORNER BRACKET <「> (U+300C) may also be used.

5.2.8.4. Spacing rule

- The dinkus typically appears centrally aligned, on a line of its own, with vertical spacing before and after it.

5.2.8.5. Special handling

- In contemporary practice, the dinkus is mutually exclusive with clause headings; in older usage a dinkus can appear before a new clause heading, especially when it conveys an emotional break.

5.3. Listing stops

5.3.1. General

5.3.1.1. Primary function and purpose

Listing stops are used to delimit multiple items in a listing, other than phrases.

5.3.1.2. Range of semantic functions

Because of the conceptual similarity between phrase stops and listing stops, as joining multiple items, punctuation marks for the two are often conflated. In fact, if the definition of "phrase" is liberal enough, listing stops are phrase stops. This framework considers punctuation separating noun or adjective phrases to be listing stops, and not phrase stops.

EXAMPLE: *I went to the market, and I bought some beans*: comma as phrase stop (it is separating two phrases that are complete sentences)
I bought some beans, cheese, and a nice bottle of Chianti: comma as listing stop (it is separating noun phrases, which can include articles, adjectives, and quantifiers)

5.3.2. Enumeration delimiter (*enumeration comma*)

5.3.2.1. Primary function and purpose

The enumeration delimiter separates items in a list. This function is represented in Metanorma i18n files as `punct.enum-comma`, and in Metanorma Presentation XML as `<span class="fmt-enum-comma">`.

In CJK, both comma-like punctuation marks and interpuncts are used as enumeration delimiters. There is no consistent semantic distinction between the two, so no such distinction is made here.

EXAMPLE: Japanese: 小・中学校 or 小、中学校 "elementary, [and] middle school"

5.3.2.2. Range of semantic functions

- There is variation in whether the enumeration delimiter is used in a list of two items.
- There is (linguistic) variation in whether a conjunction is used before the last item in a list, and (punctuation) variation in whether the enumeration delimiter is used before the last item. Thus English allows both *A, B and C* and *A, B, and C* as style variants (the latter is known as the "Oxford comma"). Chinese allows both *A*、*B*及*C* and *A*、*B*、*C*.
- The punctuation mark for the enumeration delimiter is often the same as that for the minor phrase separator (comma).
- The punctuation mark for the enumeration delimiter is often the same as that for other separators; for example Japanese uses the interpunct as an enumeration delimiter, as a decimal point (8.1.2), and to separate professional titles, names, and positions (6.1.5).

5.3.2.3. Punctuation mark in scripts, languages, and locales

- Latin, Cyrillic: COMMA < , > (U+002C)
- Traditional Chinese: IDEOGRAPHIC COMMA < 、> (U+3001)
- Simplified Chinese: IDEOGRAPHIC COMMA < 、> (U+3001)
- Japanese: IDEOGRAPHIC COMMA < 、> (U+3001)
  - Japanese uses the interpunct for short lists instead of comma in some contexts. This is rendered as KATAKANA MIDDLE DOT < ・> (U+30FB)

- In mixed Japanese-Western text, FULLWIDTH COMMA < ， > (U+FF0C) can be used instead to maintain visual consistency

- Korean: COMMA < , > (U+002C)
  - Korean uses the interpunct for short lists instead of comma in some contexts. This is rendered as HANGUL LETTER ARAEA < · > U+318D.

5.3.2.4. Special handling

- When lists contain mixed scripts in CJK, follow the punctuation convention of the list's primary language while maintaining readability.
- Japanese also supports HALFWIDTH KATAKANA MIDDLE DOT <･> (U+FF65) for the interpunct.
- In Japanese vertical text, the interpunct appears centered rather than at a specific corner position.

5.3.3. List itemisation mark

5.3.3.1. Primary function and purpose

List itemisation marks are used at the start of list items, when they are rendered as separate lines in a list paragraph.

List items may be ordered or unordered; if they are ordered, the list itemisation mark is an ordered numeral or letter, optionally followed by a caption number delimiter (6.2.2). In the case of unordered lists, a single typographic symbol is used.

5.3.3.2. Punctuation mark in scripts, languages, and locales

- The default list itemisation mark for unordered list items is BULLET <•> (U+2022). Alternatives include the em dash, en dash, and WHITE BULLET <◦> (U+25E6).

5.3.3.3. Spacing rules

- List items of different levels in a list are typically indented to differing degrees, so as to indicate that level.
- The list itemisation mark is typically separated from the list item content by a tab (a horizontal space of consistent width between different list items).
- List items are typically rendered with hanging indentation, so that their list itemisation marks and content lines up across multiple-line list items.

5.3.3.4. Special handling

- In word processing and HTML, list itemisation marks are not entered directly by the author, but are indicated to be rendered via list markup. The selection of list

itemisation mark in unordered lists is almost never specified in markup, but in the document configuration/document stylesheet.

5.4. Quotation markers

5.4.1. General

5.4.1.1. Primary function and purpose

Quotation markers indicate that a span of text constitutes direct speech, or otherwise attribute it to some third party.

5.4.1.2. Spacing rules

- In French in France, Switzerland and Belgium, for paired quotation delimiters, the opening delimiter (left guillemet, left single guillemet) is followed by a non-breaking thin space (U+202F), as an instance of French spacing (5.1.1.3); the closing delimiter (right guillemet, right single guillemet) is preceded by a non-breaking thin space, as the mirror counterpart of French spacing. As with other French spacing, in common practice, a full non-breaking space (U+00A0) is used instead.

5.4.2. Paired quotation delimiters (*double quotes*)

5.4.2.1. Primary function and purpose

Paired quotation delimiters indicate the start and end of quoted text. This function is represented in Metanorma i18n files as `punct.open-quote` and `punct.close-quote`.

5.4.2.2. Range of semantic functions

- Paired quotation markers are often conflated with title marks (7.3.6).

5.4.2.3. Punctuation mark in scripts, languages, and locales

- Latin, Cyrillic: There is significant variation among Latin and Cyrillic script languages on their choice of paired quotation delimiter and their orientation, and different languages and locales draw from 15 Unicode codepoints.
  - In American English, good typographical practice uses LEFT DOUBLE QUOTATION MARK <“> (U+201C) and RIGHT DOUBLE QUOTATION MARK <”> (U+201D) as opening and closing marks. In British English, some publishers use double quotation marks, and others use single quotation marks, LEFT SINGLE QUOTATION MARK <‘> (U+2018) and RIGHT SINGLE QUOTATION MARK <’> (U+2019).
  - Finnish uses the RIGHT DOUBLE QUOTATION MARK <”> (U+201D) as both opening and closing marks.

- French uses guillemets, LEFT-POINTING DOUBLE ANGLE QUOTATION MARK <«> (U+00AB) and RIGHT-POINTING DOUBLE ANGLE QUOTATION MARK <»> (U+00BB) as opening and closing marks.

- Traditional Chinese: LEFT CORNER BRACKET <「> (U+300C) and RIGHT CORNER BRACKET <」> (U+300D)
- Simplified Chinese: LEFT DOUBLE QUOTATION MARK <“> (U+201C) and RIGHT DOUBLE QUOTATION MARK <”> (U+201D).
  - The corner brackets of Traditional Chinese are also in common use.
- Japanese: LEFT CORNER BRACKET <「> (U+300C) and RIGHT CORNER BRACKET <」> (U+300D).
  - Japanese also uses lenticular brackets, LEFT BLACK LENTICULAR BRACKET <【> (U+3010) and RIGHT BLACK LENTICULAR BRACKET <】> (U+3011).
- Korean: South Korea uses the same punctuation marks as English. North Korea uses full-width guillemets, LEFT DOUBLE ANGLE BRACKET <《> (U+300A) and RIGHT DOUBLE ANGLE BRACKET <》> (U+300B).
  - Corner brackets and white corner brackets are often used in practice.

5.4.2.4. Special handling

- The placement of quotation marks in CJK is different by directionality, but this does not involve different Unicode codepoints, so the browser/operating system is assumed to handle positioning appropriately. (While there is e.g. a PRESENTATION FORM FOR VERTICAL LEFT CORNER BRACKET <﹁> (U+FE41) and PRESENTATION FORM FOR VERTICAL RIGHT CORNER BRACKET <﹂> (U+FE42), these are only included in Unicode as a compatibility character.)
- Quotation marks in English are routinely typed as straight quotes, QUOTATION MARK <"> (U+0022), with the expectation that software will transform them contextually into paired quotation marks (“smart quotes”).

> *Metanorma transforms straight quotes in Asciidoc source into smart quotes following English conventions.*

- Simplified Chinese uses the font to make Western quotation marks (U+201C, U+201D) full-width, instead of deploying distinct codepoints: REVERSED DOUBLE PRIME QUOTATION MARK <〝>, (U+301D) DOUBLE PRIME QUOTATION MARK <〞> (U+301E).
- In vertical directionality, Traditional and Simplified Chinese use presentation variants of the codepoints LEFT WHITE CORNER BRACKET <﹃> (U+300E) and RIGHT WHITE CORNER BRACKET <﹄> (U+300F).

5.4.3. Nested paired quotation delimiters (*single quotes*)

5.4.3.1. Primary function and purpose

Nested paired quotation delimiters indicate the start and end of quoted text within another delimited quoted text, and attributed to a different speaker from that of the nesting quoted text. This function is represented in Metanorma i18n files as `punct.open-nested-quote` and `punct.close-nested-quote`.

EXAMPLE: *"Didn't she say 'I like red best' when I asked her wine preferences?" he asked his guests.*

5.4.3.2. Range of semantic functions

- The English closing nested paired quotation delimiter is often conflated with the elision mark (7.3.3).

5.4.3.3. Punctuation mark in scripts, languages, and locales

- Latin, Cyrillic: There is significant variation among Latin and Cyrillic script languages on their choice of paired quotation delimiter, and different languages and locales draw from 15 Unicode codepoints.
  - In American English, good typographical practice uses LEFT SINGLE QUOTATION MARK <‘> (U+2018) and RIGHT SINGLE QUOTATION MARK <’> (U+2019) as opening and closing marks. In British English, some publishers use single quotation marks, and others use double quotation marks, LEFT DOUBLE QUOTATION MARK <“> (U+201C) and RIGHT DOUBLE QUOTATION MARK <”> (U+201D).
  - Finnish uses the RIGHT SINGLE QUOTATION MARK <’> (U+2019) as both opening and closing marks.
  - French uses single guillemets, LEFT-POINTING SINGLE ANGLE QUOTATION MARK <‹> (U+2039) and RIGHT-POINTING SINGLE ANGLE QUOTATION MARK <›> (U+203A).
- Traditional Chinese: LEFT WHITE CORNER BRACKET <『> (U+300E) and RIGHT WHITE CORNER BRACKET <』> (U+300F).
- Simplified Chinese: LEFT SINGLE QUOTATION MARK <‘> (U+2018) and RIGHT SINGLE QUOTATION MARK <’> (U+2019).
  - The white corner brackets of Traditional Chinese are also in common use.
- Japanese: LEFT WHITE CORNER BRACKET <『> (U+300E) and RIGHT WHITE CORNER BRACKET <』> (U+300F).
- Korean: South Korea uses the same punctuation marks as English. North Korea uses full-width guillemets, LEFT ANGLE BRACKET <〈> (U+3008) and RIGHT ANGLE BRACKET <〉> (U+3009).
  - Corner brackets and white corner brackets are often used in practice.

5.4.3.4. Special handling

- The placement of quotation marks in CJK is different by directionality, but this does not involve different Unicode codepoints, so the browser/operating system is assumed to handle positioning appropriately. (While there is e.g. a PRESENTATION FORM FOR VERTICAL LEFT WHITE CORNER BRACKET <﹗> (U+FE43) and PRESENTATION FORM FOR VERTICAL RIGHT WHITE CORNER BRACKET <﹘> (U+FE44), these are only included in Unicode as a compatibility character.)
- Nested quotation marks in English are routinely typed as straight quotes, APOSTROPHE <'> (U+0027), with the expectation that software will transform them contextually into paired quotation marks ("smart quotes").

> *Metanorma transforms straight quotes in Asciidoc source into smart quotes following English conventions.*

- Simplified Chinese uses the font to make Western quotation marks (U+2018, U+2019) full-width, instead of deploying distinct codepoints
- In vertical directionality, Traditional and Simplified Chinese use presentation variants of the codepoints LEFT CORNER BRACKET <﹁> (U+300C) and RIGHT CORNER BRACKET <﹂> (U+300D).

5.4.4. Single quotation delimiters (*quotation dash*) (out of scope)

5.4.4.1. Primary function and purpose

A single initial marker used to represent one turn in an alternation of direct speech.

EXAMPLE: *— O saints above! Miss Douce said, sighed above her jumping rose. I wished I hadn't laughed so much. I feel all wet.*
— O Miss Douce! Miss Kennedy protested. You horrid thing!

—James Joyce, *Ulysses*

5.4.4.2. Punctuation mark in scripts, languages, and locales

- Latin, Cyrillic: Properly HORIZONTAL BAR <―> (U+2015) is used; in practice EM DASH <—> (U+2014) is usually used.
  - The quotation dash is almost unknown in English, but is commonplace in other European languages, such as French.
- The Japanese PART ALTERNATION MARK <〽> (U+303D) has narrower application, being used in Noh drama to indicate the start of a speaker's or the chorus' part. The opening square quotation mark <﹁> may also be used.

5.4.4.3. Spacing rules

There is a thin space after the quotation dash.

6. Functions: word-level structure

6.1. Conjunctors

6.1.1. General

Connection indicators link related elements, show relationships between parts, or indicate continuation across boundaries.

6.1.2. Conjunctive mark

6.1.2.1. Primary function and purpose

The conjunctive mark presents two or more words or morphemes as both applicable. It is a punctuation counterpart to “and”.

6.1.2.2. Range of semantic functions

- In formal usage, ampersand is restricted to citations of multiple authors, and to disambiguate scope of conjunction (items in a list joined by “and”, some of which are themselves joined by “and”).

  EXAMPLE 1 — Citation of multiple authors: *Jones & Jones (2005)*

  EXAMPLE 2 — Items in a list joined by “and”, some of which are themselves joined by “and”: Rock, pop, rhythm & blues and hip hop

6.1.2.3. Punctuation mark in scripts, languages, and locales

- In Latin script, the default conjunctive mark is AMPERSAND <&> (U+0026).
  - In Irish and Scots Gaelic, TIRONIAN SIGN ET <⁊> (U+204A) is used traditionally.
  - In Swedish, underlined *o* is also used.
  - In informal usage, the plus sign may be used.

6.1.2.4. Spacing rules

- By default the ampersand is consider to stand in for “and” as a word, and is spaced either side as a word. It is printed without spaces when it is part of an acronym, e.g. *R&D* = “Research and Development”.

6.1.3. Disjunctive mark

6.1.3.1. Primary function and purpose

The disjunctive mark presents two or more words or morphemes as alternatives. It is a punctuation counterpart to “or”.

EXAMPLE: *Iran/Persia* (place may be designated as either)
is/are
he/she
s/he (truncated version of *she/he*)

6.1.3.2. Range of semantic functions

- The disjunctive mark can be used to express inclusive "or", or "and"; in that case it overlaps with the conjunctive mark and the range mark.
- English uses both a disjunctive mark and range mark to express points on an itinerary; e.g. *Shanghai/Nanjing/Wuhan/Chongqing* or *Shanghai–Nanjing–Wuhan–Chongqing*

6.1.3.3. Punctuation mark in scripts, languages, and locales

- In English, the expression of the disjunctive mark is the slash, SOLIDUS </> (U+002F).

6.1.3.4. Spacing rules

- Normally there is no space either side of the slash. Some style guides require space when the phrase being joined contains a space:

  EXAMPLE: *and/or*
  New Zealand / Western Australia

6.1.4. Range mark (*en-dash*)

6.1.4.1. Primary function and purpose

A marker between two items, used to convey a range between the two. The items can be verbal values (*Paris–New York*), numerals (*3–5*), dates (*Jan–Oct*), quantities, etc. This function is represented in Metanorma i18n files as `punct.en-dash`.

6.1.4.2. Range of semantic functions

- Latin script conflates the punctuation marks for verbal and numeric ranges. CJK uses distinct punctuation marks for verbal and numeric ranges. The latter function is represented in Metanorma i18n files as `punct.numeric-en-dash`.
- The range mark is also used to contrast two items, or convey a relationship between them: *Mother–daughter relationship*,

6.1.4.3. Punctuation mark in scripts, languages, and locales

- Latin, Cyrillic: In good typography, EN DASH <–> (U+2013) is used. Informal use typically uses the hyphen, and some style guides prescribing formal use prefer the hyphen when conveying a relationship, in its role as a word divider (6.1.8).
  - In French, TILDE <~> (U+007E) is used: *3~5 m*.

- Traditional Chinese, Simplified Chinese, Japanese, Korean: EN DASH <–> (U+2013), WAVE DASH <〜> (U+301C), FULLWIDTH TILDE <〜> (U+FF5E).

6.1.4.4. Spacing rules

- Typically in Latin script there is no spacing around the en-dash. This differentiates it from the phrase stop use of the en-dash. Some style guides recommend spacing where it would avoid ambiguity in scope, e.g. *12 June – 3 July*; others do not, e.g. *12 June–3 July*.

6.1.4.5. Special handling

- In French, the range mark can be open on either side of the interval: *~3* means "up to 3", *100~* means "100 or more".
- In CJK, verbal range is conveyed through en-dash. Tilde and single em-dash can also be used for verbal range.
- Numeric range in CJK is normally wave dash. Some Japanese academic writing uses colons instead of wave dash, though this is not universal.

6.1.5. Name separator

6.1.5.1. Primary function and purpose

Particularly in CJK, components of a name are separated from each other using a punctuation mark where necessary for clarity. This applies to names in the broadest sense, including personal names (particularly foreign names), professional titles and positions next to names, and names of works (i.e. titles, separating the title from the subtitle).

EXAMPLE 1: Simplified Chinese: 李奥納多·達·文西 "Leonardo·da·Vinci"

EXAMPLE 2: Japanese: 部長補佐・鈴木 "Assistant Department Head·Suzuki"

6.1.5.2. Range of semantic functions

- There is a continuity of function between the name separator and the word divider in CJK, particularly as CJK uses ideographs for words (so words are not consistently broken down), and it uses word dividers only sparingly.
- The following are the more refined classes of name separator; because of the fluidity of classes, they are enumerated here rather than being broken down into separate semantic functions in the framework.
  - Components of foreign names (Chinese, Japanese)
  - Components of an organsiation name
  - Title and subtitle of a work
  - Professional title, professional position, personal name

- Components of a bibliographic reference entry (e.g. *Smith, J. 1989. The passage of time. New York: Wiley.*)

- Caption markers (6.2) include a subset of name separators, but these are discussed separately, as a core concern of Metanorma configuration.

6.1.5.3. Punctuation mark in scripts, languages, and locales

- Latin, Cyrillic: In Western scripts, this function is conflated, with the normal word separator (6.1.7) (space), with phrase stops (comma by default (5.2.2), colon for title/subtitle (5.2.4), both for organisation names), or with word divider (6.1.8) (hyphen, used in personal names).
  - The colon or em-dash is used for title/subtitle: *Star Trek V: The Final Frontier*, *Star Trek V—The Final Frontier*
  - The comma, colon, or em-dash is used for organsiation name: *Yamato Bank: Osaka Branch*
  - The hyphen is used within a single component of a personal name, a surname or a given name: *Jean-Jacques*, *Zeta-Jones*
- CJK: Various forms of interpunct (5.3.2.3): MIDDLE DOT < · > (U+OOB7) in China, HYPHENATION POINT < · > (U+2027) in Taiwan, KATAKANA MIDDLE DOT < • > (U+30FB) to separate Katakana words in Japanese.
  - Japanese also uses WAVE DASH <〜> (U+301C), to separate title from subtitle.
  - Japanese also uses the KATAKANA-HIRAGANA DOUBLE HYPHEN < = > (U+U+30A0) and the FULLWIDTH EQUALS SIGN <＝> <U+FF1D>. This double hyphen is used to render Latin hyphen in transliteration of foreign names, i.e. in the function of separating single components of a Western name. That function occasionally uses the interpunct instead. More rarely, the double hyphen also is used instead of the interpunct to delimit names and titles in Western names, e.g. サー = アーサー = コナン = ドイル "Sir=Arthur=Conan=Doyle"
  - Japanese occasionally uses the interpunct for Japanese names, particularly when there would otherwise be confusion as to where one name ends and another begins.
  - Japanese can use ideographic space to delimit components of an organisation name: 大和銀行　大阪支店 "Yamato Bank, Osaka Branch".

6.1.5.4. Special handling

- The interpunct is half-width in Chinese in print, but full-width in online material.

6.1.6. Attribution mark

6.1.6.1. Primary function and purpose

The attribution mark is used to delimit a quotation from its source.

EXAMPLE 1:

To be or not to be, that is the question.

—William Shakespeare

EXAMPLE 2: To be or not to be, that is the question" —William Shakespeare

6.1.6.2. Range of semantic functions

- The attribution mark is comparable to the caption separator (6.2.3) and the breaking phrase separator (5.2.5). It accordingly shares punctuation with them.

6.1.6.3. Punctuation mark in scripts, languages, and locales

- The em-dash is typically used. In less formal writing, the colon and the comma are also used.

6.1.6.4. Spacing rules

- When applied to a block quote, the attribution starts on a separate line, with the attribution mark dash. The separate line is typically indented.
- When used inline, the attribution mark dash follows the quotation directly.

6.1.7. Word separator (*space*)

6.1.7.1. Primary function and purpose

The word separator is used to separate words from each other.

6.1.7.2. Punctuation mark in scripts, languages, and locales

- In Latin and Cyrillic, this function is fulfilled by the space, which is not regarded as punctuation at all. The space is applied universally as a word separator.
  - This was not the case in antiquity. Roman inscriptions used an interpunct (5.3.2.3) as a word separator; Greek and Roman manuscripts did not use word separators. (The latter practice is known as *scriptio continua*, "continuous writing")
- CJK by contrast defaults to not using word separators, outside of special contexts (here considered as name separators). CJK therefore has *scriptio continua*.
- There are circumstances in which a word separator is necessary in careful text in CJK. Japanese specifically uses the interpunct to separate ordinary Japanese words, outside of the special cases of name separators (6.1.5), where the intended meaning would be unclear if the characters were written side-by-side. It is more commonly used to separate foreign words and names when written in kana: パーソナル・コンピューター (*pāsonaru·konpyūtā* "personal computer").

6.1.8. Word divider

6.1.8.1. Primary function and purpose

The word divider indicates the compound structure of a word, dividing it into pieces smaller than a word.

6.1.8.2. Range of semantic functions

- The word divider differs from the name separator in scope: name dividers operate on groupings of words, the word divider operates within a word.
- The grammatical word divider (6.1.9) is a special case of the word divider.
- The hyphen (6.1.10) is a special case of the word divider.
- The divisions that a word divider makes can vary in scope. What is divided up may be a compound consisting of meaningful units: independent words, or morphemes (e.g. prefixes: *de-emphasise*). Or, what is divided up may be a single meaningful unit, broken up into phonetic units (syllables, or letters: syllabification or spelling out of a word).

  EXAMPLE 1 — Meaningful units: *post-war*
  Indo-European

  EXAMPLE 2 — Phonetic units: *R-E-S-P-E-C-T* (spelling out)
  in-cre-di-ble (syllabification)

- The extent to which the word divider is used in compound words varies by language and by style. The use of hyphen to indicate compound words (e.g. *pigeon-hole*, *craftily-constructed chair*) is on the decline in English.
- There may be two levels of division represented, with different dividers used to differentiate between them. This applies because here are two actual levels of grouping, or because a component being connected is a multi-word expression with space in it

  EXAMPLE 3 — Two levels of division: *Pre–Indo-European*
  San Francisco–area

6.1.8.3. Punctuation mark in scripts, languages, and locales

- In Latin and Cyrillic, the default word divider is the hyphen, HYPHEN-MINUS <-‌> (U+002D). Unicode also has HYPHEN <-> (U+2010) to disambiguate it from the minus sign (8.1.3), but this "Unicode hyphen" is little used.
  - The EN DASH <–> (U+2013) is used as a higher level of division in careful typesetting (*Pre–Indo-European*, *San Francisco–area residents*).
  - In lexicography, interpunct is also used for syllabification.
  - In some transcription practice, e.g. for Arabic, Japanese and Chinese into English. the apostrophe is used to indicate syllabification, as disambiguation (#### transcribed as *mus'haf*, to avoid the pronunciation "moo-shaf"; しんいち transcribed as *Shin'ichi* to indicate the syllabification is "Shi-n-i-chi" rather than "Shi-ni-chi").

— This is also done in Pinyin: *Xi'an* is two syllables, *Xī-ān*, whereas *xian* would be read as a single syllable.

6.1.8.4. Spacing rules

Space is not meant to appear after word dividers normally, as any word separators would contradict the intent of the word divider as showing divisions within a single word.

6.1.8.5. Special handling

Some instances of word divider are not meant to be conflated with the hyphen, and a line break between such word divisions is inappropriate. For such cases, NON-BREAKING HYPHEN <-> (U+2011) is used.

6.1.9. Grammatical word divider

6.1.9.1. Primary function and purpose

The grammatical word divider, like the word divider, indicates the compound structure of a word, dividing it into pieces smaller than a word.

The word divider can indicate the compound structure of any compound word, and is not a normal part of conventional orthography. The grammatical word divider, by contrast, is used to designate specific grammatical functions of morphemes, and is part of the conventional orthography of a language.

6.1.9.2. Range of semantic functions

- The use of a grammatical word divider distinct from the general word divider is idiosyncratic, and inconsistent between languages.
- The use of the grammatical word divider is also less widespread than the elision mark, with which English conflates it.
  - Wikipedia lists instances in Danish, Estonian, Finnish, Polish, Turkish, and Welsh. In the first four it is restricted to foreign words, where the inflection would be hard to parse as distinct from the unfamiliar word; in Turkish it is restricted to proper nouns, for similar reasons; and in Welsh it is used to disambiguate infixed pronouns. None of them use it regularly for every instance of the grammatical function, the way English does for possessive *-'s*.
  - English attaches possessive *-s* to nouns using a grammatical word divider. No other Germanic language does.

    EXAMPLE 1: English: *Julia's*
    German: *Julias*

  - Current practice in English uses a grammatical word divider for apostrophes, but a normal word divider for unexpected verb inflections:

    EXAMPLE 2: *Julia's*
    to-ing and fro-ing

— Practice in English is in flux for particular functions: unexpected plurals, and verb inflections after vowels, used to be separated by a grammatical word divider; increasingly, that is not used at all.

EXAMPLE 3 — Older practice using grammatical word dividers: *bastinado'd*
KO'd
B's and C's

EXAMPLE 4 — Newer practice avoiding grammatical word dividers:
bastinadoed
KOed
Bs and Cs

— Grammatical word dividers are also increasingly avoided in formal labels, such as geographical place names

EXAMPLE 5: *Coffs Harbour* (town in Australia, originally Korff's Harbour)
Earl's Court, Barons Court (adjacent London Underground stations)

6.1.9.3. Punctuation mark in scripts, languages, and locales

- In Latin and Cyrillic, the grammatical word divider is usually the apostrophe.
  - — In fine typography, RIGHT SINGLE QUOTATION MARK < ' > (U+2019) is used as the apostrophe. In typewriter usage, which is inherited in online usage, APOSTROPHE < ' > (U+0027) is used.

6.1.9.4. Special handling

The grammatical word divider function of the apostrophe has the same special handling as the elision mark function (7.3.3).

6.1.10. Hyphen

6.1.10.1. Primary function and purpose

The hyphen is a special case of word divider (6.1.8). It is used to divide a word at a line-break, so that the amount of empty space in a line is reduced.

EXAMPLE: We, therefore, the represen-
tatives of the United States
of America ...

6.1.10.2. Punctuation mark in scripts, languages, and locales

- The hyphen is in routine use in Latin and Cyrillic.
- The word delimiter is alien to CJK; accordingly, the hyphen is also alien to CJK.

6.1.10.3. Spacing rules

The hyphen is only meaningful in that function at the end of a line, and no characters are meant to appear after it.

6.1.10.4. Special handling

- An optional hyphen, SOFT HYPHEN (U+00AD) is often used in word processing, which is only rendered when near a line break.

6.1.11. Identifier divider

6.1.11.1. Primary function and purpose

The identifier divider is used to break up a token that is not a linguistic expression.

EXAMPLE: Phone number: *555-8787*
Date: *2000-01-01*
Document identifier: *ISO 639-2*
Document entity identifier: *Table 3.1*, *Clause 4.1.4*

6.1.11.2. Range of semantic functions

- The function of the identifier divider and the word divider are quite similar: they both are breaking up a token and showing its internal structure. They differ in that one divides a linguistic word, and the other divides a non-linguistic token.
- The function of the identifier divider is also similar to the decimal point, as a number token divider; however decimal points are dealt with separetely, as numeric formatting.
- Identifier dividers can apply to numbers as identifiers, but are distinct from decimal points in that they don't have an arithmetic function. So when separating groups of digits in a phone number or a date, the hyphen is distinct from a decimal point.
- The components of a numeric date are also split by identifier divider.
- In the case of hierarchical identifiers of entities in a document, such as clauses, Metanorma Presentation XML notates this as `<span class="fmt-autonum-delim">`.

6.1.11.3. Punctuation mark in scripts, languages, and locales

- The usual identifier divider is the hyphen. In Unicode, the identifier divider function is properly indicated with FIGURE DASH <–> (U+2012), as that symbol is designed to have the same width as Arabic numerals.
- Period, slash, colon and parentheses are also used idiosyncratically, through analogy with their core meanings (hyphen as a word divider; sentence stop as delimiter of a meaningful unit; slash as a disjunctive mark; colon as an introductory stop; parentheses as additional material).
  - Period is the most common for hierarchical identifiers of document entities; colon is the most common for delimiting a following date.

6.1.12. Identifier delimiter

6.1.12.1. Primary function and purpose

The identifier divider is used to separate an identifier from surrounding text, whether at the start, the end, or both.

EXAMPLE: *Formula (1)*
Figure 1 a)

6.1.12.2. Range of semantic functions

- Identifier delimiters distinct from word delimiters are very rare. The most common use of them is in structured documents such as under Metanorma, for subfigures and formulas. Whether an identifier delimiter is used is a matter of style convention.
- The identifier delimiter is similar in function to title marks (7.3.6) and emphasis marks (7.4.2).
- Metanorma Presentation XML currently notates this function as `<span class="fmt-autonum-delim">`, although this is properly the notation for identifier dividers.

6.1.12.3. Punctuation mark in scripts, languages, and locales

- The usual identifier delimiter is the close parenthesis, or surrounding parentheses.

6.2. Caption markers

6.2.1. General

6.2.1.1. Primary function and purpose

In formal documents (particularly the standards in scope of Metanorma), titles of cross-referencable entities within the document (figures, tables, lists, list items, footnotes, clauses, etc.) are labelled in such a way as to isolate the identifier of the entity. This is done both in labelling the entity, and in cross-referencing that entity.

6.2.2. Caption number delimiter

6.2.2.1. Primary function and purpose

The identifier of the entity may be signalled with punctuation acting as a delimiter. In Metanorma Presentation XML, this is notated as `<span class="fmt-label-delim">`.

EXAMPLE 1: *Table 1.*

EXAMPLE 2: *3)*

6.2.2.2. Range of semantic functions

- Different kinds of entity can take different punctuation. Ordered list items are commonly delimited with a closing parenthesis instead of a period.

6.2.2.3. Punctuation mark in scripts, languages, and locales

- Latin, Cyrillic: Period is the most common caption number delimiter, if a delimiter is used at all. Closing parentheses are common for ordered list items and footnotes.

6.2.3. Caption separator

6.2.3.1. Primary function and purpose

If the number of the entity appears together with a caption or title, giving more information about the entity, punctuation is usually used to separate the two. In Metanorma Presentation XML, this is notated as `<span class="fmt-caption-delim">`

EXAMPLE 1: *Table 1. Distribution of rice yields*

EXAMPLE 2: *2.1: Soil erosion*

6.2.3.2. Range of semantic functions

- Different kinds of entity can take different punctuation. For example, a document may use period as a caption separator for clauses, but dash for tables and figures.
- This function overlaps with the caption separator. If a caption number delimiter is used, a distinct separator is usually not used after it. The following illustrates different possible configurations.

  EXAMPLE 1 — No caption number delimiter, colon as caption separator: *Table 1*
  Table 1: Distribution of rice yields

  EXAMPLE 2 — No caption number delimiter, period as caption separator: *Table 1*
  Table 1. Distribution of rice yields

  EXAMPLE 3 — Period as caption number delimiter, caption delimiter blocks distinct caption separator: *Table 1.*
  Table 1. Distribution of rice yields
  NOT: *Table 1.: Distribution of rice yields*

6.2.3.3. Punctuation mark in scripts, languages, and locales

- Latin, Cyrillic: All of the following may appear in this function: space (i.e. no special punctuation), colon, comma, em-dash, period

6.2.4. Caption stop

6.2.4.1. Primary function and purpose

The caption or title of an entity in a document may be terminated with a punctuation mark. In Metanorma Presentation XML, this is notated as `<span class="fmt-label-delim">` (conflating it with the Caption number delimiter.)

EXAMPLE: *Table 1: Distribution of rice yields.*

6.2.4.2. Range of semantic functions

- This function is an extension of the sentence stop, since captions can be seen as sentences. Whether it is used or not depends on the publisher's style conventions.

6.2.4.3. Punctuation mark in scripts, languages, and locales

- Latin, Cyrillic: if a caption stop is provided, it is a period.

6.2.5. Hierarchical caption separator

6.2.5.1. Primary function and purpose

The hierarchical caption separator separates hierarchical components of a cross-reference to entity in a document (e.g. *Clause 5, Note 1*). This is used when the entity cross-reference by itself is ambiguous (e.g. note numbering restarts each clause, so "Note 1" on its own is ambiguous between the first note of clause 4 and of clause 5.) In Metanorma Presentation XML, this is notated as `<span class="fmt-comma">`.

6.2.5.2. Range of semantic functions

- This is a special case of a name separator (6.1.5), and shares punctuation with it: comma in Latin script.

6.2.5.3. Punctuation mark in scripts, languages, and locales

- Latin, Cyrillic: if a caption stop is provided, it is a period.
- In some languages and styles, this function is conveyed by linguistic words instead of punctuation; e.g. in Japanese, the particle の "of" is used instead.

7. Functions: adding meaning

7.1. Annotation markers

7.1.1. General

7.1.1.1. Primary function and purpose

Annotation markers provide supplementary information or references related to the main text content.

7.1.1.2. Range of semantic functions

- Annotation markers are routinely conflated with grouping markers (7.2), which enclose, separate, or highlight specific text elements.
- Different kinds of annotation are indicated by different punctuation conventions, as described below, but these are usually not rigorously differentiated.

EXAMPLE: The Leiden conventions for publishing ancient inscriptions and papyri use a wide range of brackets with quite distinct meanings, in order to convey editorial approaches to a text. (They are a form of machine readable text changes.) This is the upper limit of distinct semantic functions in annotation markers, and most use of annotation markers is semantically much more ad hoc:

- ạḅ: letter is unclear in original
- [abc]: letters missing from original (because text is broken off), restored by editor
- ⌊ abc⌋ : letters missing from original, but restored from another source, e.g. a mediaeval manuscript of the same text
- ⟨abc⟩: letters left out from original text (because the scribe never wrote them), restored by editor
- a(bc): abbreviation in original, letters expanded by editor
- {abc}: letters written in the original by the scribe, deleted as errors by the editor
- ⟦abc⟧: letters deleted in the original by the scribe, restored by the editor
- \abc/: letters interpolated in the original by the scribe (typically between lines)

7.1.1.3. Special handling

CJK full-width punctuation (5.1.1.4) applies.

7.1.2. Parenthetical annotation (*parentheses*)

7.1.2.1. Primary function and purpose

Parenthetical annotation marks enclose supplementary or explanatory information. This function is represented in Metanorma i18n files as `punct.open-paren` and `punct.close-paren`.

7.1.2.2. Range of semantic functions

- Parenthetical annotations can appear either within a sentence, or as a group of one or more sentences, between sentences.
- Parenthetical annotations can be conflated with grouping markers (7.2), although they are less common in that function than for other paired punctuation marks.
- As an extension of their role, parenthetical annotations can also occur within a word, to indicate that a part of the word is optional in some sense. This is a special case of a word divider. Examples include indication of singular and plural, or masculine and plural; both give the letter distinguishing those grammatical categories in parentheses, as optional:

  EXAMPLE 1: *(s)he*
  plan(s)

- Within a sentence, paired breaking phrase (em-dashes) separators (5.2.5) can overlap in function with parenthetical annotations: the parenthetical information is presented as a break from the main sentence, and another breaking phrase separators resumes the main sentence.

  EXAMPLE 2: *The red-nosed reindeer (Rudolph was his name) had a very shiny nose.*
  The red-nosed reindeer—Rudolph was his name—had a very shiny nose.

7.1.2.3. Punctuation mark in scripts, languages, and locales

- Latin, Cyrillic: By default, parenthetical annotations are rendered as LEFT PARENTHESIS <(> (U+0028) and RIGHT PARENTHESIS <)> (U+0029)
- Paired em dashes are used within a sentence, when the parenthetic annotation can be presented as a sentence break and resumption. This is associated with less formal style.
- CJK: FULLWIDTH LEFT PARENTHESIS <（> (U+FF08) and FULLWIDTH RIGHT PARENTHESIS <）> (U+FF09)
  - Use of a paired em dash equivalent is rare in CJK.
  - Japanese: Japanese in addition uses double wave dash, WAVE DASH <〜> (U+301C)

    EXAMPLE: 〜〜答え〜〜

7.1.2.4. Special handling

- Languages differ as to whether parentheses enclosing italicised text should themselves be in italics. In German, they are expected to; in English, they are expected not to.
- Parenthetical annotations can be nested within other parenthetical annotations. In informal writing, parentheses are used for both nesting and nested annotations. In formal writing, concern for ambiguity drives many style guides to require that the nested annotation be in brackets instead of parentheses, with any third level of nesting in curly brackets.

  EXAMPLE: From Wikipedia:

  > *Parentheses may be nested (generally with one set (such as this) inside another set). This is not commonly used in formal writing (though sometimes other brackets [especially square brackets] will be used for one or more inner set of parentheses [in other words, secondary {or even tertiary} phrases can be found within the main parenthetical sentence]).*

7.1.3. Footnote annotations (*footnote marks*)

7.1.3.1. Primary function and purpose

Footnote annotations are text annotation marks that provide supplementary information or references related to the main text content, like parenthetical annotations. Unlike parenthetical annotations, the annotation does not appear inline with the text it is annotating, but in a separate place: either the bottom of a page (**footnote**), the end of a chapter or book (**endnote**), or, in older practice, the margin of a text (**marginalia**). A footnote mark is used to cross-reference the place annotated in the text (where it is a **footnote reference**), to the annotation content (where it is a **footnote label**).

7.1.3.2. Punctuation mark in scripts, languages, and locales

- Latin, Cyrillic: there are two repertoires of marks drawn on for footnote marks
  - Where a small repertoire of footnote marks is required (e.g. footnotes are cycled through once for each page), and in older practice, footnote marks are drawn from a set of typographical symbols, in the traditional order <* † ‡ § ‖ ¶>, supplemented in ad hoc function by other symbols.
  - Where a large repertoire of footnote marks is required (e.g. footnotes are cycled through once per chapter or document), and in newer practice, footnote marks are drawn from an ordered sequence of numerals or letters. Arabic numbers are the most commonly used, but Roman numbers and letters of the alphabet also appear.

7.1.3.3. Spacing rules

- There is normally no space between text being annotated and the footnote reference.
- There may be space between the footnote label and the annotation.

7.1.3.4. Special handling

- Footnote marks are normally superscripts, both as footnote references and as footnote labels.
- Footnote references normally appear after sentence and phrase stops.
- There are different conventions on whether the repertoire of footnote symbols is cycled through (i.e. restarted from 1 or *) each page, each chapter, or once in a document.
- There may be a caption number delimiter after the footnote mark. Documents may differ in whether they place a caption number delimiter after footnote references and after footnote labels.
- In some documents, a semantic distinction is made between different footnote mark sequences.

7.2. Grouping markers

7.2.1. General

Grouping markers include a range of paired punctuation markers, used to indicate that the enclosed items or word belong together in some sense. While grouping markers have some semantic functions associated with editorial interventions, the association is loose, and the markers are instead differentiated by visual form.

The listing of grouping markers here is limited to those in use in normal language. Grouping markers that only appear in formal notation systems, e.g. double brackets (MATHEMATICAL LEFT/RIGHT WHITE SQUARE BRACKET <⟦ ⟧> U+27E6/U+27E7) and floor brackets (LEFT/RIGHT FLOOR <⌊ ⌋> U+230A/U+230B) are out of scope of this document.

7.2.1.1. Range of semantic functions

- Grouping markers overlap with parenthetical annotations (7.1.2).
- Grouping markers are used to indicate editorial interventions in text; the Leiden conventions (7.1.1.2, Example) are a very formalised equivalent of the less granular conventions expressed by grouping markers. Editorial interventions are not here treated as a distinct semantic function.

7.2.2. Bracket

7.2.2.1. Range of semantic functions

- Brackets can be used to insert explanatory material, instead of parentheses. Typically the brackets represent an editorial intervention, rather than the authorial voice.

EXAMPLE 1 — Use of brackets to indicate editorial change of a text's case: *[m]y cause is just*

EXAMPLE 2 — Use of brackets to indicate an editorial explanatory interpolation: *I appreciate it [the honor], but I must refuse*

EXAMPLE 3 — Use of brackets to interpolate the original language word in a translation: *He is trained in the way of the open hand* [karate].

7.2.2.2. Punctuation mark in scripts, languages, and locales

- Latin, Cyrillic: LEFT SQUARE BRACKET <[> (U+005B), RIGHT SQUARE BRACKET <]> (U+005D)
- CJK: FULLWIDTH RIGHT SQUARE BRACKET <［> (U+FF3B), FULLWIDTH RIGHT SQUARE BRACKET <］> (U+FF3D)

7.2.3. Brace

7.2.3.1. Range of semantic functions

- The main use of braces in English normal text is to indicate editorial additions and interpolations. They are overwhelmingly used in formal notation systems instead.

7.2.3.2. Punctuation mark in scripts, languages, and locales

- Latin, Cyrillic: LEFT CURLY BRACKET <{> (U+007B), RIGHT CURLY BRACKET <}> (U+007D)
- CJK: FULLWIDTH RIGHT CURLY BRACKET <｛> (U+FF5B), FULLWIDTH RIGHT CURLY BRACKET <｝> (U+FF5D)

7.2.4. Angle bracket

7.2.4.1. Range of semantic functions

- Angle brackets have limited use to indicate editorial interpolation, or to indicate that a text was thought by a character instead of spoken. They are overwhelmingly used in formal notation systems instead.

7.2.4.2. Punctuation mark in scripts, languages, and locales

- Latin, Cyrillic: Good typography expects MATHEMATICAL LEFT ANGLE BRACKET <〈> (U+27E8), MATHEMATICAL RIGHT ANGLE BRACKET <〉> (U+27E9) for mathematical and normal text usage. In informal usage, and in many formal notation schemes such as computer programming, LESS-THAN SIGN <<> (U+003C) and GREATER-THAN SIGN <>> (U+003E) are used instead.

- CJK: LEFT ANGLE BRACKET < 〈 > (U+3008), RIGHT ANGLE BRACKET < 〉 > (U+3009).

7.3. Semantic identification

7.3.1. General

7.3.1.1. Primary function and purpose

Semantic identification punctuation identifies a span of text as conveying a specific semantics, as distinct from the structural information conveyed by punctuation marks in the foregoing usage categories.

7.3.1.2. Range of semantic functions

- Semantic identification overlaps routinely with emphasis marks (7.4.2), and with quotation markers (5.4).

7.3.1.3. Punctuation mark in scripts, languages, and locales

- Latin, Cyrillic: Punctuation marks for most semantic identification, other than abbreviation marks, are not used in contemporary Latin and Cyrillic practice—although number marks were commonplace in ancient and mediaeval versions of those writing systems.

7.3.2. Abbreviation mark

7.3.2.1. Primary function and purpose

Abbreviation marks are used to indicate that a word is an abbreviation of another word, and is to be understood as such rather than as a phonetically intact word.

7.3.2.2. Punctuation mark in scripts, languages, and locales

- Latin, Cyrillic: the usual abbreviation mark is the period.
  - The ambiguity of abbreviation mark and declarative sentence stop (5.1.2) is particularly pernicious, and is a pressing issue for Metanorma converting between Latin and CJK punctuation.
  - The slash, SOLIDUS </> (U+002F), is used for some abbreviations in English, especially involving initials of separate words or morphemes (*w/o* = "without")
  - The apostrophe, as an elision mark (7.3.3), is also used in some abbreviations: *gov't* < *government*.
- CJK: while CJK languages do form truncated expressions (e.g. *Běidà* 北大 for *Běijīng Dàxué* 北京大学 "Peking University"), abbreviation marks are not used explicitly.

7.3.2.3. Spacing rules

- There is variation as to whether an abbreviation of a multi-word phrase retains the spaces of its source phrase (*U. S.* vs *U.S.*).

7.3.2.4. Special handling

- There is variation by language, style, and instance as to when to use the abbreviation mark, and whether to use abbreviation marks within an acronym (*N.A.T.O.* vs *NATO*, *etc.* vs *etc*).

7.3.3. Elision mark

7.3.3.1. Primary function and purpose

The elision mark is used to indicate that a word is derived from a word of the same meaning in an older, more formal or more standard version of the language, through the deletion of letters or numbers in the base form.

7.3.3.2. Range of semantic functions

- Letters may be deleted from the beginning, the middle, or the end of the base form.

  EXAMPLE 1: *'tis < it is*
  shootin' < shooting
  'n' < and
  bo'sun < boatswain

- Letters may be deleted from a single word derived from a phrase; the resulting word is called a contraction.

  EXAMPLE 2: *isn't < is not*
  won't < woll not (Middle English variant of *will not*)

- Numbers may also be deleted, specifically in expressions for decades (this is an English-specific convention):

  EXAMPLE 3: '70s < 1970s

- Elision marks often differentiate colloquial forms from formal forms, and therefore words using elision marks are often avoided in formal style.
- Elision marks prioritise the formal form of a language, deriving elided forms from it; that can make its use politically contentious, if the prioritisation of that form becomes controversial.

  EXAMPLE 4: In the 19th century, Scots forms were derived from Standard English forms in spelling, using elision marks (e.g. *a' < all, gi'e < give*); this was rejected in the 20th century, as Scots was increasingly regarded as a distinct language from English (with the words now spelled *aw, gie*), rather than as a corrupt form of London English; the older use of the elision mark is dismissed as the "apologetic apostrophe".

A similar backlash has occured in the Anglosphere against other such transcription of non-standard English ("eye dialect").

7.3.3.3. Punctuation mark in scripts, languages, and locales

- Latin, Cyrillic: in fine typography, RIGHT SINGLE QUOTATION MARK < ’ > (U+2019) is used as the elision mark. In typewriter usage, which is inherited in online usage, APOSTROPHE < ' > (+0027) is used.
  - Both forms of the punctuation mark are called apostrophe.

7.3.3.4. Spacing rules

- There is variation as to whether an abbreviation of a multi-word phrase retains the spaces of its source phrase (*U. S.* vs *U.S.*).
- There is no space before an elision mark in a contraction, even if the elision mark denotes the start of a new day (*it's* < *it is*.)

7.3.3.5. Special handling

- Data-entry of the straight APOSTROPHE character is automatically corrected by word processors to the curved RIGHT SINGLE QUOTATION MARK character. The conversion uses the identical mechanism as smart quotes for quotation marks (5.4.3).
- The RIGHT SINGLE QUOTATION MARK punctuation mark is ambiguous between a quotation mark and an elision mark, as indeed its Unicode name indicates. That means that initial elision marks are incorrectly corrected from the straight quote: it is treated as an initial quotation mark, and changed to LEFT SINGLE QUOTATION MARK, whereas initial elision marks are always RIGHT SINGLE QUOTATION MARK.

  EXAMPLE: ' n ' (data entry)
  ’n’ (correct rendering)
  ‘n’ (incorrect smart quotes conversion, treating elision mark as quotation mark)

7.3.4. Cardinal number mark (out of scope)

7.3.4.1. Primary function and purpose

In writing systems where number symbols are ambiguous with letters of the script, it is routine to highlight the numeric use of letters to prevent ambiguity, through a cardinal number mark.

EXAMPLE: Latin *VI* "by power"
Roman numeral *V͞I*"6"

7.3.4.2. Punctuation mark in scripts, languages, and locales

- Hebrew and Greek have distinct cardinal number marks (*gershayim* and *geresh* in Hebrew, *keraia* in Greek), including disambiguating marks for counts of thousands (quote mark in Hebrew, *lower keraia* in Greek) and for reciprocal fractions (double *keraia* in Greek).

  EXAMPLE — Greek numerals: δ' "4" (Greek letter delta, used as a number)
  ,δ "4000"
  δ″ "1/4"

- Roman numbers used an overbar in antiquity to indicate that they were numbers; as it happens, an overbar (*vinculum*) was also used in the same period, to indicate that the number was to be multiplied by a thousand (so V̄I̅ was used for both "6" and "6,000"). The number mark is rare in contemporary usage of Roman numbers.

7.3.5. Ordinal number mark (out of scope)

7.3.5.1. Primary function and purpose

In some languages, ordinal numbers are preceded by a mark indicating that this is an ordinal, and which is often derived from an abbreviation of the spoken word for "number".

EXAMPLE: № 29 Acacia Rd
#29 Acadia Rd

7.3.5.2. Punctuation mark in scripts, languages, and locales

- In Romance languages such as French and Spanish, an abbreviation of a word for "number" is used, usually an abbreviation of Italian *numero*. In French and Spanish, this is n^{o}, with the *o* superscript. In British English, *No.* is used. In German, *Nr.* is used. Such abbreviations are not to be regarded as punctuation.
  - In informal use in the French- and Spanish-speaking world, DEGREE SIGN <°> (U+00B0) is used, as a truncation of n^{o}.
  - Some languages, such as Portuguese and Italian, use the masculine ordinal indicator <º> (U+00BA) instead of a superscript *o*.
- The number sign NUMERO SIGN <№> (U+2116) is very common in use in Russian and Bulgarian; it is derived from the Latin script abbrevation.
- In American English, NUMBER SIGN <#> (U+0023) (hash) is used.

7.3.5.3. Spacing rules

- The abbreviations of "number" and their derivatives, including the Numero sign, are followed by space.
- The number sign (hash) is not followed by a space.

7.3.6. Title mark

7.3.6.1. Primary function and purpose

The title mark is used to indicate the title of a document. This is a core function in bibliographic rendering, but it is also applied outside of formal bibliographic rendering, in running text. This function is represented in Metanorma i18n files as `punct.open-title` and `punct.close-title`.

7.3.6.2. Punctuation mark in scripts, languages, and locales

- Latin, Cyrillic: this is by default handled through italics rather than a punctuation mark. Quotation marks are often used instead, although in more formal referencing, quotation marks are used as secondary title marks instead.
- Traditional Chinese: WAVY LOW LINE (U+FE4F) is used as official title markup, particularly in texts that also use the proper name mark.
  - Square brackets 【】 and double quotation marks 『』 are also used.
- Simplified Chinese: uses《...》for titles.
  - Square brackets 【】 and double quotation marks 『』 are also used for song titles.
- Japanese: 『...』(double quotes) are used for book titles.
  - Western quotes are unofficially used in press for titles in Traditional Chinese, Japanese, and Korean.

7.3.6.3. Special handling

- The title mark is associated with the name separator (6.1.5), to separate components of a bibliographic reference entry.

7.3.7. Subsidiary title mark

7.3.7.1. Primary function and purpose

The subsidiary title mark is used to indicate the subtitle of a document.

7.3.7.2. Range of semantic functions

- The title mark is associated with the name separator (6.1.5), to delimit titles from subtitles.

7.3.7.3. Punctuation mark in scripts, languages, and locales

- Latin, Cyrillic: subtitles are not marked up as a separate span from titles, and Western script referencing relies instead on the name separator to differentiate titles from subtitles.
- Japanese: WAVE DASH <〜> (U+301C) can be used to mark subtitles: 〜概要〜

7.3.8. Secondary title mark

7.3.8.1. Primary function and purpose

The secondary title mark is used to indicate the title of a subsidiary document in bibliographic referencing. This applies to article and chapter titles, as distinct from book and journal titles. This function is represented in Metanorma i18n files as `punct.open-secondary-title` and `punct.close-secondary-title`.

7.3.8.2. Punctuation mark in scripts, languages, and locales

- Latin, Cyrillic: this is by default handled through quotation marks.
- Traditional Chinese, Simplified Chinese: Single title marks 〈...〉 () are used for article and chapter titles (LEFT ANGLE BRACKET < 〈> (U+3008) and RIGHT ANGLE BRACKET <〉 > (U+3009).)

7.3.9. Proper name mark

7.3.9.1. Primary function and purpose

The proper name mark is used to indicate a proper name within a document.

7.3.9.2. Range of semantic functions

- Writing systems can use the name separator instead (6.1.5), to differentiate names from surrounding text.

7.3.9.3. Punctuation mark in scripts, languages, and locales

- Traditional Chinese: an underline is occasionally used, in didactic and ambiguous contexts (teaching materials, movie subtitles).

7.3.10. Clause mark (out of scope)

7.3.10.1. Primary function and purpose

A clause mark is used to indicate that an identifier is a cross-reference to a clause in a document.

7.3.10.2. Range of semantic functions

- Use of the clause mark is now infrequent outside of legal documents, with the word for "clause" typically supplied instead. In ISO documents, a dot-delimited numeral on its own (i.e. featuring an <identifier-divider>>) is understood to be a clause reference, without an explicit indication of the cross-reference scope ("clause").

7.3.10.3. Punctuation mark in scripts, languages, and locales

- Latin: SECTION SIGN <§> (U+00A7)

7.3.10.4. Spacing rules

- The clause mark is linked to the following identifier with a non-breaking space.

7.3.10.5. Special handling

- When multiple clauses are referenced, the section sign may be duplicated: "§§ 13–21". This is analogous to cross-reference abbreviations of reference locality types, such as "pp." for "pages".

7.3.11. Paragraph mark (out of scope)

7.3.11.1. Primary function and purpose

A paragraph mark is used to indicate that an identifier is a cross-reference to a paragraph in a document.

EXAMPLE: *17 U.S.C. § 411 ¶ 5* "Title 17 of the United States Code, section 411, paragraph 5"

7.3.11.2. Range of semantic functions

- Use of the paragraph mark is infrequent outside of legal documents, with the word for "paragraph" typically supplied instead.
- There is limited use of the paragraph mark to represent carriage return at the end of a paragraph, or the concept of a paragraph: that use is not properly punctuation.

7.3.11.3. Punctuation mark in scripts, languages, and locales

- Latin: PILCROW SIGN <¶> (U+00B6).

7.3.11.4. Spacing rules

- The paragraph mark is linked to the following identifier with a non-breaking space.

7.4. Highlight markers

7.4.1. General

7.4.1.1. Primary function and purpose

Highlight markers highlight and draw attention to specific text elements through visual modification or annotation, for a range of functions.

7.4.1.2. Range of semantic functions

- The range of functions for which highlight markers are used is open-ended, and often overlaps with other semantic functions. This includes emphasis on the part of the author; emphasis on the part of a speaker in direct speech; use–mention distinction (so citation of words as subjects of discussion, as opposed to use of words in language); titles and names of entities; and foreign words.
- The extent of use of highlight markers varies by language and style.

7.4.1.3. Punctuation mark in scripts, languages, and locales

- Latin and Cyrillic do not use punctuation in this function, and resort instead to typographical styling, including italics, boldface, and underlining. These are not in scope of this framework, except where Metanorma needs to alternate between typographical styling in Latin/Cyrillic, and punctuation marks in CJK.
  - While some documents make meaningful distinctions between italics, boldface, and underlining, they are idiosyncratic to the document, and cannot be generalised readily. Italics is the default highlight marking device in English.
- CJK does not use italics, although it does use boldface.
  - Instead of italics, CJK uses a range of explicit punctuation, including single or double quotes, brackets, lenticular brackets, or emphasis marks.
  - For example, Japanese uses lenticular brackets in dictionaries for quoting Chinese characters and Sino-Japanese loanwords.

7.4.2. Emphasis marks

7.4.2.1. Punctuation mark in scripts, languages, and locales

- CJK emphasis marks are a diacritic applied to every character in the text span to be emphasised. Chinese is usually restricted to using an underdot. Japanese uses a much wider range of emphasis marks: filled or hollow dot, circle, double circle, triangle, and "sesame" marks.

7.4.2.2. Special handling

- While the underdot emphasis mark could be realised as a distinct Unicode character, CJK emphasis marks are treated in computer typesetting as markup. They are realised

in CSS using the `text-emphasis-style` attribute. This means that Western and CJK emphasis marks are both treated as markup rather than as punctuation.

- CJK emphasis marks are rare online and unsupported in many Word processors. (Microsoft Word also supports a smaller range of emphasis marks in Japanese than does HTML CSS or XSL:FO.)

> *Metanorma does not currently support emphasis marks. When it does, it will be as a discretionary expansion of `<em>` markup in Semantic Metanorma XML.*

8. Functions: other

8.1. Numeric punctuation

8.1.1. General

Punctuation used to convey different types of number overlaps with other punctuation discussed here; Metanorma handles it with distinct mechanisms (document attributes and `number:[]` macro parameters: see Metanorma documentation). It is briefly outlined here for completeness, and to identify potential points of ambiguity.

The punctuation used to convey different types of number also overlaps with some symbols for mathematical functions (notably plus and minus signs); mathematical functions are out of scope of this framework.

8.1.2. Decimal point

8.1.2.1. Primary function and purpose

The decimal point delimits integer and decimal portions of a number.

8.1.2.2. Range of semantic functions

- The decimal point is a special case of an identifier divider (6.1.11).

8.1.2.3. Punctuation mark in scripts, languages, and locales

- Latin, Cyrillic: there is variation among Western script languages: the Anglosphere uses a period, while Continental Europe uses a comma. In the case of ISO, even English-language texts adhere to Continental norms in using comma as a decimal point.
- Traditional Chinese: the HYPHENATION POINT <‧> (U+2027) (interpunct) is used as a decimal point in Chinese numbers: 三‧五 "3.5".

- Japanese: the same practice is followed, using the Japanese version of the interpunct, KATAKANA MIDDLE DOT < • > (U+30FB): 三 • 一四 “3.14”.

8.1.3. Minus sign

8.1.3.1. Primary function and purpose

The minus sign indicates negative numbers.

8.1.3.2. Range of semantic functions

- The same sign is used for negative numbers and for the arithmetic subtraction operator; the former is in scope of this framework, as negative numbers can appear outside the context of mathematical typesetting, inserted into normal text.

8.1.3.3. Punctuation mark in scripts, languages, and locales

- Fine typography prefers to use the distinct MINUS SIGN <−> (U+2212) in mathematical use. Informal use uses HYPHEN-MINUS <-> (U+002D).

8.1.4. Ratio sign

8.1.4.1. Primary function and purpose

Indicates a ratio or proportion of two numbers.

8.1.4.2. Range of semantic functions

- The ratio function is a generalisation of both the fraction sign and the division sign in mathematics.
- Proportions can include a proportion of completion, e.g. *Page 17/35*, indicating how many pages have been traversed out of a total of 35.

8.1.4.3. Punctuation mark in scripts, languages, and locales

- In English, both colon and slash are used: *1:7*, *1/7*.
  - Unicode differentiates the FRACTION SLASH <⁄> (U+2044), the division sign DIVISION SLASH <∕> (U+2215), and the non-mathmetical use of slash.

8.1.5. Approximation sign

8.1.5.1. Primary function and purpose

Indicates that a numeric value is approximate.

EXAMPLE: ~40 metres

8.1.5.2. Punctuation mark in scripts, languages, and locales

- TILDE <~> (U+007E) is used by default
  - Some languages such as French also use ALMOST EQUAL TO <≈> (U+2248).

8.2. Miscellaneous

8.2.1. Missing text mark

8.2.1.1. Primary function and purpose

The missing text mark represents omitted text. This function is represented in Metanorma i18n files as `punct.ellipse`.

8.2.1.2. Range of semantic functions

- The ellipse is conflated with the hesitancy phrase separator (5.2.6), but the two interact differently with other punctuation.

8.2.1.3. Punctuation mark in scripts, languages, and locales

- The same mark, the elllipse, is used as for the hesitancy phrase separator.
  - Occasionally triple asterisks, <* * *> or <***>, are used as an explicit missing text mark. This is done in some legal writing in English. In this regard, the missing text mark usually has scope at paragraph-level rather than sentence-level, and it is a variant of the section separator (5.2.8).
- When the omitted material is part of a word, e.g. out of religious or social taboo, or anonymisation, one or two em dashes are used traditionally. For taboo deletion, asterisk is also used; for more expressive or jocular deletion, a range of symbols is used.

  EXAMPLE 1 — Anonymisation: *It was alleged that D—— had been threatened with blackmail.*

  EXAMPLE 2 — Religious taboo: *In the name of G–d*

  EXAMPLE 3 — Social taboo: *F—— you!*
  *F*ck you!*
  F@# you!*

8.2.1.4. Spacing rules

- When the missing text mark represents one or more missing words, it is usually treated for spacing as a separate word, rather than as punctuation. Therefore unlike the hesitancy phrase separator, the missing text mark is preceded by space as a word separator, just as a word would. This extends to treating the missing text mark as a distinct sentence.

So British practice, as reflected in the *Oxford Style Guide*, has the hesitancy mark replace the sentence stop, but the missing text mark appear after a sentence stop. It also has space before both the missing text mark and the hesitancy mark.

EXAMPLE 1: *The … fox jumps …*
_The quick brown fox jumps over the lazy dog. … And if they have not died, they are still alive today.
It is not cold … it is freezing cold.

- But as an illustration of the flux around spacing rules: the *University of Oxford Style Guide* (which applies to the university rather than to the commercial printer) requires no space around the missing text mark, and space after and not before the hesitancy mark:

  EXAMPLE 2: *The…fox jumps…*
  _The quick brown fox jumps over the lazy dog…And if they have not died, they are still alive today.
  It is not cold… it is freezing cold.

8.2.1.5. Special handling

- The same special handling applies as for the hesitancy phrase separator.
- As a sign of editorial intervention, the missing text mark is often put inside brackets (7.2.2) in English, to make the editorial intervention explicit, and to disambiguate it from the hesitancy phrase separator. In Spanish, parentheses are preferred in that role.

  EXAMPLE: Original text: *The President said that, for as long as this situation continued, he would not be satisfied*
  Reported text: *The President said that […] he would not be satisfied*

8.2.2. Repetition mark

8.2.2.1. Primary function and purpose

The repetition mark indicates a place where text is repeated from a previous passage, and it is obvious visually what the repeated text is.

8.2.2.2. Range of semantic functions

- Repetition marks are infrequent in contemporary use outside of specific contexts, to avoid ambiguity. Space saving is not now as pressing a concern as it was in the past in texts.

8.2.2.3. Punctuation mark in scripts, languages, and locales

- English traditionally uses a ditto mark under each word to be repeated from a previous line; these can take the form of straight quotes, right smart quotes, or apostrophe.

FIGURE 1

```
Black pens, box of twenty ... $2.10 + Blue " " "
" ... $2.35
```

- In Quebec French and Greek, a right guillemet is used. In other European languages (e.g. Italian), a double prime is used; the English use of apostrophes is an approximation of this. In yet others (e.g. Swedish), the double prime ditto mark is set on the baseline, and surrounded by em-dashes.
- In CJK, DITTO MARK <〃> (U+3003) is used.
- These are clearly visual variants of each other, that have been conflated historically with other punctuation in the same local writing systems.

- In some citation conventions, if two consecutive reference entries in a bibliography have the same author(s), a repetition mark is used instead of repeating the author, consisting of two or three em-dashes.

  EXAMPLE: Smith, J. & J. Doe. 1990. *The elements of style*. New York: Wiley.
  ——— 1991. *The rudiments of style*. New York: Wiley.

8.2.3. Iteration mark

8.2.3.1. Primary function and purpose

The iteration mark indicates the repetition of a word or a component of a word.

8.2.3.2. Punctuation mark in scripts, languages, and locales

- In Chinese, IDEOGRAPHIC ITERATION MARK <々> (U+3005) is used in casual usage to repeat a character, but is no longer used in formal usage.
- In Japanese, IDEOGRAPHIC ITERATION MARK <々> (U+3005) remains in common use for repeating a single kanji, and follows different norms from the Chinese instance of the punctuation mark.
  - Japanese also has repeat marks for repeating the preceding word or phrase of two or more characters: VERTICAL KANA REPEAT MARK <〳〵> (U+3031) for exact repetition, VERTICAL KANA REPEAT WITH VOICED SOUND MARK <〴〵> (U+3032) for repetition with the first letter voiced. The repeat marks are restricted to vertical directionality, and are no longer in common use.

    EXAMPLE 1: ところ *tokoro* “place”
    ところ〳〵; *tokorodokoro* “in places”
  - Hiragana and Katakana have distinct iteration marks. They are no longer in common use, but they still appear in names.

EXAMPLE 2: *Isuzu* in Japanese is いすゞ, using a voiced iteration mark: I-su-{repeat, voiced}

EXAMPLE 3: Japanese Wikipedia refers to "heart" in Kanji: 心. Its Hiragana entry こころ includes a dozen literary works titled *kokoro* "heart". There is a distinct reference こゝろ to the 1914 novel by Soseki; being a much older work than the others, it has retained the Hiragana repetition mark, "ko-{repeat}-ro".

- Among Latin-script languages, a normal or superscript number <2> is used as an iteration mark in Filipino, Malay, and Indonesian, although in Indonesian this is no longer official practice.

9. Whitespace handling

9.1. General

Rules about whitespace handling specific to various punctuation marks have already been given inline in the foregoing discussion. The following is general discussion about whitespace handling in paragraphs, and specific to CJK writing systems—which are in *scriptio continua*, and rarely use whitespace at all typographically.

9.2. Paragraphing

Paragraphs may be indented in Western typography. If they are, this done both in word processing and HTML as document configuration, setting the indentation of the first line in a paragraph, rather than by inserting a spacing character.

In Japanese, it is common to indent the first line of a paragraph using a full-width space (U +3000).

9.3. Alternation between scripts

When alternating between CJK and Latin scripts, Chinese does not insert space at the boundary: it preserves the non-spacing nature of CJK. The word delimiter within spans of Latin text is still preserved. The same applies to Korean.

EXAMPLE: 她唱的I dreamed a dream(我曾有夢)，唱得肝腸寸斷 "She sang *I dreamed a dream*, so heartbreakingly"

In Japanese, however, spacing is required at the boundary between Japanese text and Latin characters in fine typography. This takes the form of a quarter-em space (U+2005),

and it applies to any Latin characters, including numbers, and even full-width punctuation marks that are not Japanese in origin, such as exclamation marks and question marks.

The whitespace to be applied between runs of CJK and Latin text can be configured in Metanorma i18n files, as `punct.cjk-latin-separator`. In Chinese and Korean, it is set by default to the empty string `""`. In Japanese, it is set to `\u2005`.

9.4. CJK use of whitespace

In Chinese, use of full-width space (U+3000) is very rare, and is only used in specific contexts. For example, it is used as an honorific spacing preceding certain names (such as that of Chang Kai-shek in Taiwan).

Japanese uses full-width space (U+3000) more commonly, as an optional disambiguating word divider, especially in text that mostly consists of hiragana or katakana and not kanji. There is also some usage of full-width space as a name separator (6.1.5).

10. Metanorma Implementation

10.1. Classes of punctuation localisation

Metanorma applies different types of punctuation localisation in different contexts, which provide different levels of semantic context for the punctuation to be applied; the implemented approaches are described here. The approaches are liable to change in the near future, and this document was authored in order to provide a framework for such changes, improving the quality and consistency of punctuation localisation.

The different approaches are, in order of increasingly rich semantic context:

- Smart quotes translation
- Auto-text punctuation translation
- Direct i18n configuration values
- Number localisation
- Bibliographic punctuation

10.1.1. Smart quotes translation

Like other word processing systems, Metanorma translates ASCII-based input of punctuation marks to Unicode-based punctuation marks as a post-processing step. While Metanorma input uses Asciidoctor, Metanorma does not use the native capability of Asciidoctor to do smart quotes translation; it instead does so at the end of Semantic XML

processing, using the `sterile` gem. This translation step is applied to all user-provided text in the document.

The main class of punctuation handled in this way is quotation marks; `sterile` has contextual mechanisms to differentiate between apostrophe as quotation mark and apostrophe as elision mark (left curly vs right curly single quote at the start of a word), and Metanorma adds to the configured contexts for translation (e.g. converting `'70s` to *’70s*)

Metanorma implements the English smart quotes equivalents to the straight quotes `'` and `"`; for other languages, users are expected to enter the correct Unicode punctuation marks directly in the document, rather than expect Metanorma to translate the straight quotes correctly.

10.1.2. Auto-text punctuation via `l10n()`

Metanorma implements a generic mechanism for translating text with English punctuation and spacing into other languages and scripts; as of this writing, it supports Simplified Chinese, Traditional Chinese, Korean, and French. This translation is done in the `Isodoc::I18n::l10n()` function, implemented in the `isodoc-i18n` gem.

The `l10n()` function performs the following conversions:

- Introduce French spacing for French punctuation.
- Translate Latin punctuation to full-width CJK punctuation.
- Remove spacing between CJK characters.
- Introduce space between CJK and Latin text in Japanese.

The text passed to `l10n()` is automatically generated text in Metanorma: this is for the most part document element captions and cross-references, such as “Figure 1”, “Table 2”, “Section 3.4”, etc. Bibliographic entries are also passed to `l10n()`, to normalise their punctuation to language expectations. Such text is typically generated by templates, using Latin word delimiters and punctuation between template slots. Rather than define distinct templates for English, French, Chinese and Japanese, the `l10n()` function allows the one template for auto-text to be applied to multiple languages.

The text passed to `l10n()` can be a string of text, or a fragment of XML. If it is a fragment of XML, `l10n()` is able to traverse the XML tree, and identify preceding and following context in adjacent nodes, for the purposes of punctuation and spacing localisation. For example, in a fragment like—

EXAMPLE

```
<p><strong>你想</strong> <strong>去吗</strong>
<strong>Do you want</strong> <strong>to go</strong></p>
```

—the `l10n()` method works out that the first space occurs between two spans of CJK text, despite them being inside XML elements, and deletes the first space; it will leave the second space alone, because the method also works out that it occurs in a Latin text context.

Often the text entered into the template cells for automatically generated text is in Latin script, and must not be subject to the same localisation as the surrounding template;

for example, a Japanese bibliographic citation may contain a Latin script author, whose punctuation should not be translated into Japanese.

(The particular problem with authors is the use of period as an abbreviation mark for initials, which should not be conflated with the sentence stop, and translated into fullwidth period: *Simpson, W.* should not become *Simpson ．W ．* in Japanese bibliographic entries.) In order to prevent translation in a subspan of the string, that subspan is passed wrapped in `<esc></esc>`.

While this approach works some of the time, the ambiguity of English punctuation, well-documented in this framework, limits the accuracy of such mappings; occurrences of period are particularly fraught. The punctuation of normal English text does not contain all distinctions needed in other languages: the English comma, for instance, conflates the Chinese enumeration delimiter and the Chinese minor phrase separator.

This issue will be addressed in two ways:

- Use of `l10n()` for punctuation translation will be reduced, in favour of the more well-controlled direct use of i18n configuration values: the correct punctuation for the language will where practical be looked up and inserted into the input string, so that the role of `l10n()` is reduced to dealing with spacing rules.
- `l10n()` will reduce the ambiguity of its input by assigning only one semantic function to input punctuation. (So `<,>` will never be interpreted as an enumeration delimiter.) Disambiguating punctuation markup may also be introduced.

10.1.3. Direct i18n configuration values

10.1.3.1. General

The internationalisation configuration files for Metanorma, managed in YAML format, include configuration values for punctuation marks, under the `punct:` heading. These values can be invoked directly in the templates for automatically generated text, and filled in with language-specific values, matching the requested semantic function.

The configuration files are set globally per language in the `isodoc` gem; their values are overridden in the configuration for a Metanorma flavour (and/or a Metanorma taste), and can optionally be overridden further in a configuration file supplied with the document. There is no mechanism for providing multiple configuration values for a document, and selecting which of the values to apply for a particular instance: once the overrides have been worked out, there is one configuration value applied per punctuation mark per document.

The overrides per flavour allow Metanorma to deal with SDO-specific alterations in punctuation; JIS for example uses punctuation more closely aligned with Western practice, whereas Plateau uses traditional Japanese punctuation.

10.1.3.2. Configurable punctuation marks

As of this writing, the following punctuation values can be configured in internationalisation files. While the names correspond to English punctuation marks, the intention of the

configuration is to support punctuation semantic functions; the specific semantic functions involves are named in the foregoing.

- colon (5.2.4)
- comma (5.2.2)
- enum-comma (5.3.2)
- semicolon (5.2.3)
- period (5.1.2)
- close-paren (7.1.2)
- open-paren (7.1.2)
- close-bracket (7.2.2)
- open-bracket (7.2.2)
- question-mark (5.1.3)
- exclamation-mark (5.1.4)
- emphasis-mark (7.4.2)
- em-dash (5.2.5)
- en-dash (6.1.4)
- number-en-dash (6.1.4)
- open-quote (5.4.2)
- close-quote (5.4.2)
- open-nested-quote (5.4.3)
- close-nested-quote (5.4.3)
- ellipse (8.2.1)
- open-title (7.3.6)
- close-title (7.3.6)
- open-secondary-title (7.3.8)
- close-secondary-title (7.3.8)

A subset of the foregoing is be translated from Latin punctuation in automatically generated text, via `l10n()`:

- colon: :
- comma: ,
- semicolon: ;
- period: .
- close-paren:)
- open-paren: (
- close-bracket:]
- open-bracket: [
- question-mark: ?

- exclamation-mark: !
- em-dash: —
- en-dash: –
- number-en-dash: – (invoked depending on context — both surrounding characters need to be Unicode numbers)
- open-quote: "
- close-quote: "
- ellipse: ...

In this list,

- Quotation marks are out of scope, because of the ambiguity of apostrophe, and the high variability of quotation marks between languages of `l10n()` translation.
- Enumeration delimiters are not supported directly in `l10n()`, because of the ambiguity of the comma. They are handled instead through the `multiple_and` i18n YAML value.
- Title marks require explicit semantics to drive them, which are not available in the `l10n()` context, and introducing them is deferred to bibliographic processing.

10.1.3.3. Self-references in YAML configuration

Values in the internationalisation configuration files can reference other entries in the configuration (including values overridden downstream), using the Ruby-derived syntax `#{ self.label }`, where `label` is the key of the value to be referenced. These values can include punctuation. For example, the following is how a YAML configuration value references the current rendering of the enumeration comma:

EXAMPLE

```
#{ self["punct"]["enum-comma"] }
```

10.1.4. Number localisation

Punctuation associated with numerals (8.1) is handled separately from other punctuation, as part of the localisation of numbers in Metanorma, which separates the semantic value of the number from how it is rendered. This currently allows the decimal point to be configured.

EXAMPLE 1 — Document attributes specifying decimal point and minus sign choices

```
:number-presentation-profile: notation=scientific,exponent_sign=nil,
decimal=","
```

EXAMPLE 2 — Configuration of number rendering inline

```
number:327428.7432878432992[decimal=".",notation=exponential]
```

10.1.5. Bibliographic punctuation

Bibliographic punctuation is distinct from auto-text punctuation, and is catered for in the `relaton-render` gem. The `relaton-render` gem is passed full details of the bibliographic entry to be rendered, including what the primary and secondary title are, and therefore it is the right place for title marks to be rendered, whether as punctuation or as italics.

The handling of punctuation has been overhauled in `relaton-render`. In particular, the use of punctuation to separate bibliographic fields has ceased being treated as sentence and phrase stops, since they are handled in CJK as tabs instead, and needs to avoid conflation with the abbreviation usage of period. Bibliographic rendering thus adds two new values to i18n configuration:

- `biblio-field-delimiter` is the main delimiter between (groupings of) bibliographic fields in a citation; e.g. in "Smith, J. (ed.) $ 1980 $ The smell of success $ New York: Doubleday", it is indicated by pass:c,q,a,m,p[`$`]. Note that some fields are grouped together with different publication, to indicate more tight coupling; e.g. "Smith, J" and "(ed.)", and "New York" and "Doubleday", are distinct bibliographic fields, but they are grouped together more tightly. In Western practice, the bibliographic field delimiter is usually period, though in some styles it is comma. In CJK practice, it is traditionally horizontal spacing.
- `biblio-terminator` is the final punctuation of a bibliographic entry. In some styles, it is supplied in a bibliography (e.g. NIST); in some, it is absent (e.g. ISO); in some, its use depends on the content of the entry (e.g. for BIPM, home standards vs. other standards). The bibliographic terminator is typically identified with the declarative stop (period in Western use), since a bibliographic entry is treated as a sentence. The bibliographic terminator is not used in in-document citations (citations given in running text), since they appear there as part of a sentence of running text.

10.2. Writing mode parameterization

Punctuation in CJK is handled differently in vertical and horizontal directionality. Currently there is no provision for different punctuation configuration for vertical and horizontal: the same punctuation is used in both cases, and rotating the punctuation glyphs as required is left up to the rendering engine.

If it turns out that this is not adequate, and separate configuration is needed for vertical directionality, directionality will be a separate parameter passed to `l10n()`; we propose to use the CSS `writing-mode` style values, `vertical-rl`, `vertical-lr`, `horizontal-tb`. (There is no standard encoding of directionality.)

Bibliography

[1] Chinese Standard GB/T 15834-2011, General Administration of Quality Supervision, Inspection and Quarantine; Standardization Administration of China. *General rules for punctuation*. 2011. http://openstd.samr.gov.cn/bzgk/gb/newGbInfo?hcno=22EA6D162E4110E752259661E1A0D0A8.

[2] Taiwan Ministry of Education. *Revised Handbook of Punctuation* (Second Edition). Ministry of Education. 2017.

[3] Japan Cultural Affairs Agency. *Guidelines for Creating Public Documents (Cultural Affairs Council Recommendation)*. Cultural Affairs Council. 2022.

[4] Japan Electronic Publishing Association. *Requirements for Japanese Text Layout*. W3C Working Group Note. 2012.