

CATMuS

Consistent Approaches to Transcribing Manuscripts

Ariane Pinche

Alix Chagué

Malamatenia Vlachou-Efstathiou

Simon Gabay

Matthias Gille-Levenson

Olivier Brisville-Fertin

Thibault Clérice

Built 2026-07-30

Commit cd2c7ada24

Online <https://catmus-guidelines.github.io>

Contents

1	General principles	3
1.1	CATMuS Guidelines Presentation	3
1.2	Graphemic transcription principles	3
1.3	Favoring Unicode's public domain	4
1.4	Exemple	4
1.5	Notes	5
2	Letters and numbers	5
2.1	Transcription of Alphabetic and Numerical Signs	5
2.2	Ligatures	6
2.3	Ramist letters ("u"/"v", "i"/"j")	6
2.4	Capital letters	7
2.5	Digits and numerals	7
3	Linguistic segmentation	8
3.1	Word Segmentation	8
3.2	Hyphenation	9
3.3	Diastoles	10
3.4	Notes	10
4	Abbreviations	10
4.1	Abbreviations	10
4.2	General Rule	10
4.3	Medieval Documents	11
4.4	Modern and Contemporary Documents	11
5	Accents and cedillas	12
5.1	Accents	12
5.2	Cedillas	13
6	Punctuation	13
6.1	Punctuation	13
7	Corrections and functional signs	15
7.1	Functional Marks	15

8	Table of characters	17
8.1	Character Table	17
9	Datasets	26
9.1	Medieval Dataset	26

1 General principles

1.1 CATMuS Guidelines Presentation

The CATMuS guidelines aim to propose a systematic approach to transcription in the context of creating training data for automatic text recognition (ATR)¹. Our task is to find a way to translate how a text is rendered on its original medium into a machine-readable system that facilitates learning.

Our solutions will inevitably be reductive, as it is impossible to fully capture the diversity of handwriting using the limited character set available to a computer and the necessity of simplification due to the production of a generic model and the need for those rules to be accessible to non paleograph experts. The proposed rules must strike a balance between principles suited for machine learning and a palaeographic approach. They must also adhere to a principle of simplicity (as much as possible), since producing this kind of data is fundamentally a collective effort. Therefore, we must establish the simplest common ground in our practice. For any specific corpus, CATMuS guidelines can be adapted, specify, e.g if your corpus need a distinction between <macron> and <tilde> (that have been rendered with the same sign (tilde) in our corpus), as it is significant in your corpus, then if consistent it can be later converted to be compatible wiht catmus guidelines, if you want to integrated your corpus in it.

Finally, it is essential to emphasize that adhering to the guidelines is the best way to ensure compatibility between datasets. In order to facilitate access to special characters and ensure consistent character use, we provide JSON keyboard layouts that can be uploaded to eScriptorium through the Tools and Keyboard section. The keyboard is divided into three : special caracter, combining sign (sign above a letter), suscript sign (like exposant, eg. 9)

1.2 Graphemic transcription principles

To define our approach, we follow the definitions of graphemic transcription provided by Peter Robinson and Elisabeth Solopova² — a transcription method that preserves the sequence of letters and abbreviations while reducing each form to its representation within an alphabetic system.

To ensure an accessible transcription system, we rejected the idea of producing fully imitative transcriptions that capture every variation in the shape of alphabetic letters (e.g., different forms of a, d, e, r, s, etc.), as it would be impossible to establish general transcription guidelines for all medieval documents from the 10th to the 15th centuries. Pushing imitation too far would risk making transcription both impractical and unusable³. For this reason, we chose to adopt the principles of graphemic transcription.

This approach is particularly challenging for abbreviations, as the boundary between a different shape of a sign and a distinct sign is often unclear and sometimes subjective. For example, in 13th-century French manuscripts, tildes and macrons are frequently interchangeable, though this is not always the case with other types of documents. That's why we use the same sign <~> for both. At present, the list of signs is not fixed; it may be expanded as new case studies emerge. Even if imperfect, we aim to adhere to two criteria for abbreviations to regroup potential variation of the same sign: - Shape of the sign - Potential function(s) of the sign

Our goal is to strike a balance and avoid unnecessarily multiplying the number of distinct signs, ensuring they are both recognizable and sufficiently represented in the dataset. A second argument is that multiplying the

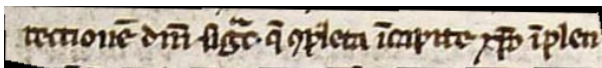
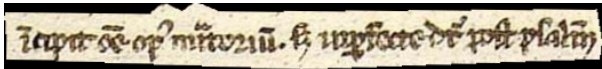
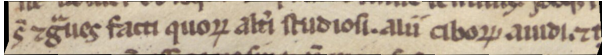
¹Paul Saenger, « La naissance de la coupure et de la séparation des mots », dans Henri-Jean Martin, Jean Vezin (éd.), *Mise en page et mise en texte du livre manuscrit*, Paris, France, 1990, p. 447-455.

²Robinson, Peter, et Elisabeth Solopova. « Guidelines for Transcription of the Manuscripts of the Wife of Bath's Prologue ». In *The Canterbury Tales Project Occasional Papers*, 19-52, 1993, doi: 10.5281/zenodo.4050360.

³As noted by Robinson and Solopova, when attempting to create imitative transcriptions, transcribers who focused too much on variations in letter shapes began making significant transcription errors that altered the meaning of the text.

number of signs can be confusing for transcribers from different backgrounds. For example, in the CATMuS guidelines, the transcription of a flattened superscript open a can be ambiguous, as it visually resembles a tilde. Some transcribe it as a tilde (since it visually resembles one), while others transcribe it as a superscript open a <“>. The choice of this specific sign is recent: it was previously transcribed with a superscript < a > as an allograph, but since some extensions of the abbreviation have lost the sense that it originally stood for an , we chose to stick to the form and use a specific sign instead.

Some examples of macrons, tildes and flattened open a:

Graz, Universitätsbibliothek, Ms. 1265, 13th c.	Transcription
	rectionē dñi sigr. q̃ 9pleta ī capite xp̃o īplen
Graz, Universitätsbibliothek, Ms. 1265, 13th c.	Transcription
	īcipit oē op̃ mītoriū. s; impfecte dē post psalm̃
Munich, BayerischeStaatsbibliothek, CLM 13027, 13th c.	Transcription
	s̃ ṡgues facti quoz alī studiosi. alīi ciboṡ auidi ṡt

However, when examining the CATMuS data, it is important to consider that: some norms have evolved over time, such as the representation of letters with strokes, which were initially treated as letters combined with a stroke. Specific signs for these cases have since been introduced in the guidelines (see Table of Special Characters). Tools like Choco-Mufin exist to harmonize, verify and document the characters used in transcriptions⁴.

Finally, one of the project's objectives is to establish a consultative board made of expert from different background with annual meetings to address specific cases or the addition of new signs to the datasets. Tricky cases will also be discussed in a FAQ to keep the guidelines as simple and systematic as possible, while still addressing the non-standardized reality of manuscripts.

NB: The transcription should only take into account characters and signs that are part of the body text. Consequently, decorative elements such as line-ending ornaments should not be transcribed.

1.3 Favoring Unicode's public domain

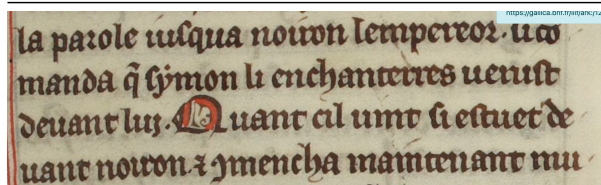
Regarding character selection, Unicode's public domain should be preferred over various private domains (known as Private Use Areas). For specialized characters, such as abbreviations or textual structuring signs, we have followed the MUFI guidelines, prioritizing characters available in the public domain. If a private-use character is chosen, this decision must be clearly documented in the project's guidelines.

1.4 Exemple

⁴Clérice, T., & Pinche, A. (2021). Choco-Mufin, a tool for controlling characters used in OCR and HTR projects (Version 0.0.4) [Computer software]. <https://doi.org/10.5281/zenodo.5356154>

BnF, Département des manuscrits, fr. 412

Transcription



la parole iusqua noiron lempereor. li comanda q̄ symon li enchanterres uenist deuant lui. Quant cil uint si estuet deuant noiron ⁊ ymencha maintenant mu-

1.5 Notes

2 Letters and numbers

2.1 Transcription of Alphabetic and Numerical Signs

Since alphabet compositions and letterform variations are not consistent over time, some letters having emerged (such as *w*) while others tended to disappear (such as *ſ* and *η*), we aimed to find a balanced approach between faithful rendering of the source and feasibility and reproducibility of the method across a broad scientific community.

2.1.1 General Principles

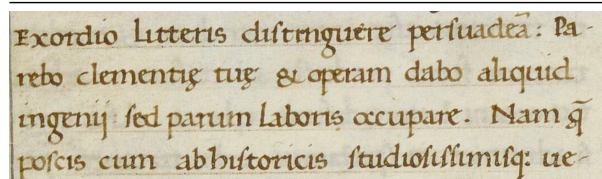
The transcription of alphabetic and numerical signs MUST follow the modern Latin alphabet. *This ensures consistency and readability, as historical variations in letter shapes are too numerous and complex to classify systematically.*

- Alphabetic signs **MUST** be transcribed using the modern Latin alphabet.
- Capital letters **MUST** be preserved.
- Small caps **MUST** be rendered as uppercase letters. *Small caps are stylistic variations, not distinct characters, so they are normalized to uppercase for simplicity.*
- Stylistic elements (e.g., italics, bold) **MUST NOT** be transcribed. *These do not involve distinct characters.*

Core Rule: The transcription **MUST** follow the letters as they appear in the source, with **NO** editorial corrections. *This preserves the source text, which is critical for historical and linguistic analysis.*

BnF, Département des manuscrits, Smith-Lesouëf 22

Transcription



Exordio litteris distinguere persuadeā: Pa-
rebo clementię tuę & operam dabo aliquid
ingenii: sed parum laboris occupare. Nam q̃
poscis cum ab historicis studiosissimisq: ue-

2.2 Ligatures

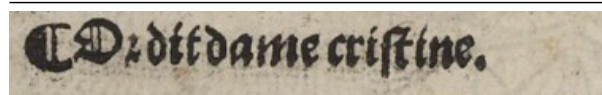
As ligatures represent graphical variations combining two letters into a single form, ligatures **MUST** be transcribed as separate letters.

In **medieval documents**, the ampersand (& [U+0026]) **MUST** be transcribed as is. This is based on the fact that this sign can bear additional abbreviation marks, such as macrons to indicate “etiam” in Latin.

In **modern and contemporary documents**, both “&” [U+0026] and “ſ” [U+00DF] **MUST** be transcribed as is.

BnF, Département Réserve des livres rares, RES-Y2-746

Transcription



¶ Or dit dame cristine.

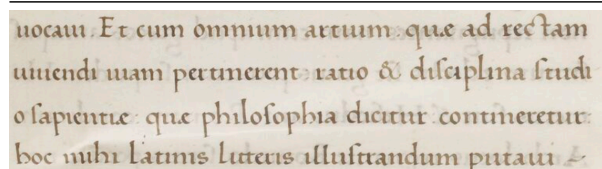
2.3 Ramist letters (“u”/“v”, “i”/“j”)

2.3.1 In medieval documents

In **medieval documents**, since the use of these signs does not distinguish between vocalic and consonantal uses, ramist letters are purely allographic variations. Therefore, Ramist letters (u/v, i/j) **MUST** be transcribed as i and u, respectively.

BnF, Département des manuscrits, Latin, 6

Transcription



uocau. Et cum omnium artium quę ad rectam
uiuendi uiam pertinerent ratio & disciplina studi
o sapientie quę philosophia dicitur contineretur
hoc mihi latinis litteris illustrandum putau

2.3.2 In modern and contemporary documents

In **modern and contemporary documents**, the transcription **MUST** follow modern usage (e.g., v for consonantal, u for vocalic).

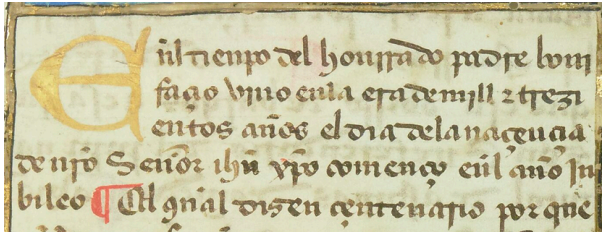
In **printed sources**, the transcription **MUST** follow the printer's usage. *Printers' conventions may differ, and respecting them ensures fidelity to the source.*

BnF, Département Littérature et art, 8-Z-18097	Transcription
8 mai. — Quelle journée admirable! J'ai passé toute la matinée étendu sur l'herbe, devant ma maison, sous l'énorme platane qui la couvre, l'abrite et l'ombrage tout entière.	8 mai. – Quelle journée admirable ! J'ai passé toute la matinée étendu sur l'herbe, devant la maison, sous l'énorme platane qui la couvre, l'abrite et l'ombrage tout entière.

2.4 Capital letters

Lowercase letters **MUST NOT** be normalized to uppercase.

If the distinction between uppercase and lowercase is unclear, the transcription **MUST** remain consistent.

BnF, Département des manuscrits, Espagnol 36	Transcription
	ñl tienpo del honrrado padre boni faço uino enla era de mill 7 trezi entos años el día dela naçençia de nro Señor ihū xpo començo eñl año iu bileo ¶ Etl gñal dizen çentenario por que

Series of capital letters or emphasized letters, such as the one found in running title, **MAY** be transcribed as series of capital letter, depending on their inner variation.

Drop capitals and decorated letters **MUST** be transcribed as uppercase. and **MUST** be considered as not a part of the same text line as the line they semantically belong to: they should be identified as drop capitals during the image segmentation phase. *Drop capitals often misalign with the baseline, so treating them separately avoids segmentation errors.*

Oversized capitals expanding above the baseline, particularly at the beginning of paragraphs, **MUST** be transcribed, even if part of the letter is missing in the mask's polygon.

2.5 Digits and numerals

2.5.1 General Rules for Numerals

Numbers **MUST** be transcribed as they appear in the document, whether as Roman or Arabic numerals. *This preserves the original format, which may be significant for historical or stylistic analysis.*

- For Roman numerals, ramist standardization (u, v, i, j) **MAY** be applied following the rules specified above.
- Numbers written with superscript letters **MUST** follow the rules specified in the section dedicated to superscript abbreviations.
- The groups of numbers **MUST** be clearly separated, and the punctuation around the number **MUST** be retained where it exists.

In **modern and contemporary documents**, when the distinction between a capital I and a i is difficult to make (as can be the case in typewritten documents or in some manuscripts), we **MAY** follow the transcription that makes the most sense in the source.

Old-style roman numerals M and D (sometimes transcribed as “CIO” and “IO”) **MUST** be transcribed as, respectively, M and D.

In modern documents, decimals **MUST** follow the convention used in the document (, or .). In case of a doubt, we recommend sticking to the system used more commonly in the rest of the document. If the source uses a symbol or a space to separate thousands, we should follow the report it in the transcription.

BnF, Département des manuscrits, fr. 654

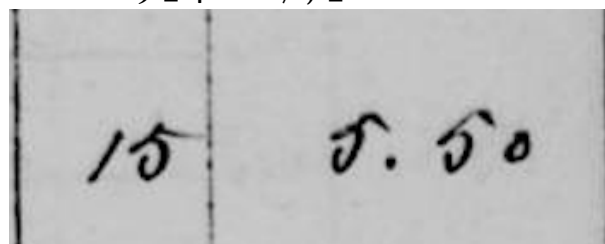
Transcription



Arnautz daniels. .iiii.

Lecturaep, fonds Dufour,
DAFANCH96_048MIC07650_L-0

Transcription



15. 5.50

3 Linguistic segmentation

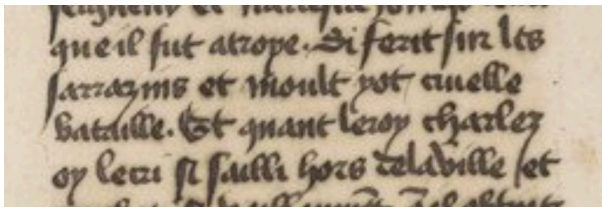
3.1 Word Segmentation

Word segmentation **MUST** be transcribed as closely as possible to the source, without normalization.

However, the medieval period marks a transition from *scripta continua* (continuous script without spaces) to the use of typographic spaces between words, a practice fully established only with the advent of print technology. Word separation in medieval manuscripts—and sometimes in manuscripts more generally—is therefore **not standardized**, making it inherently complex.⁵

In a given document, word division can be highly subjective, and two transcribers may interpret it differently. When it is impossible to determine whether a space is present or not, the irregularity of spacing in manuscripts has led to the following rule: **When unclear, word segmentation SHOULD follow modernized conventions to ensure consistency, using the typographic space (” ” [U+0020]).** However, some exceptions apply due to the medieval nature of some documents and their linguistic state, which are described below in the section *Exceptions and Special Cases*.

⁵Paul Saenger, « La naissance de la coupure et de la séparation des mots », dans Henri-Jean Martin, Jean Vezin (éd.), *Mise en page et mise en texte du livre manuscrit*, Paris, France, 1990, p. 447-455.

source	transcription
	que il fut atroye. Si ferit sur les sarrazins et moult yot cruelle bataille. Et quant le roy charlez oy le cri si sailli hors delaville et

3.1.1 Special Cases

- As the use of the apostrophe doesn't exist in medieval manuscript, original agglutination **MUST** be kept, eg. *q'il* for the normalized form *q' il*.
- The use of doubled consonants to note elision **MUST** be kept and not normalized, eg. *arriva* is kept instead of *à riva*.
- **Locutions in the process of being lexicalized** (such as for verbs like *enchargier/en chargier*, *en fuir/enfuir* or certain locutions such as *aujourd'hui*) **MUST** be transcribed as they appeared in the source.
- The separation of the initial at the beginning of a verse in poetry **MUST** be imitated.
- In the case of a drop capital at the beginning of a line, it **CAN** either be ignored (as such letters are often more decorative than textual) or transcribed on a separate line. This is because the baseline of a drop capital typically does not align with the rest of the text, and line segmentation often fails to capture it as part of the continuous line.

When in doubt, word segmentation should follow the usage of the contemporary or normalized language of the manuscript.

3.2 Hyphenation

Hyphenation refers to the indication that a word has been split at the end of a line. While it appears inconsistently in medieval manuscripts, it is frequent in modern and contemporary sources.

Hyphenation **MUST** be transcribed whenever it appears in the source.

A hyphenation symbol **MUST NOT** be added if it is not present in the source.

The character “-” [U+002D] **MUST** be used to transcribe the hyphenation symbol, regardless of its form in the source. If the hyphenation symbol is repeated at the beginning of the next line, it **MUST** also be transcribed as “-” [U+002D].

BnF, département Littérature et art, Y2-27423

Transcription

For the next eight or ten months, Oliver was the victim of a systematic course of treachery and deception—he was brought up by

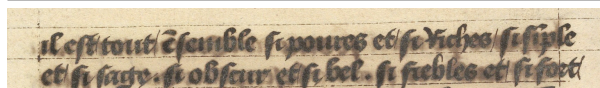
For the next eight or ten months, Oliver was the victim of a systematic course of treachery and deception – he was brought up by

3.3 Diastoles

Sometimes, diastoles (vertical or oblique pen strokes) are drawn between two contiguous letters to indicate that they belong to different words. Diastoles (vertical or oblique pen strokes) **MUST** be transcribed with the sign “/” [U+002F].

University of Pennsylvania, ms. 660

Transcription



il est/tout/ẽsemble si poures et/si riches/si siple/
et/si sage. si obscur et/si bel. si fiebles et/si fort/

3.4 Notes

4 Abbreviations

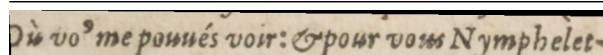
4.1 Abbreviations

4.2 General Rule

Abbreviations **MUST** be reproduced and **MUST NOT** be expanded in the transcription. *Expanding abbreviations at the transcription stage—where automated tools lack contextual and linguistic information **MUST** be avoided, as it risks introducing inconsistencies and degrading model performance. We therefore consider that abbreviation expansion **SHOULD** be treated as a separate normalization task, performed after the transcription stage.*

Paris, BnF, YE-7580, 17th c.

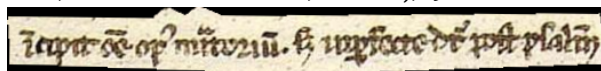
Transcription



Où vo⁹ me pouués voir: & pour vous Nymphet-

Graz, Universitätsbibliothek, Ms. 1265, 13th c.

Transcription



ĩcipit oẽ op⁹ mĩtorĩũ. s; impfecte dť post psalmĩ

4.3 Medieval Documents

4.3.1 Character Selection

For medieval documents, MUFI characters **MUST** be preferred, along with Unicode’s public domain characters. *The use of standardized characters ensures interoperability and consistency across projects.* - The **semantic value** of the unicode sign **MUST** be taken into account. *For example, to represent “p with stroke,” the Armenian letter թ (Ke) [U+0554] **MUST NOT** be used, even if it visually resembles p [U+A751], which is the correct character carrying the appropriate semantic value for “p with stroke through the descender.”*

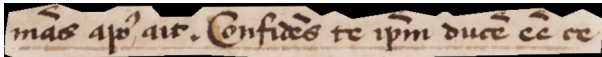
When selecting the appropriate MUFI character, the choice **MUST** be based on the **shape** of the abbreviation, keeping in mind that CATMuS does not differentiate allographs. If the shape leads to an uncertain choice, the choice **MAY** be informed by the **function** of the abbreviation sign.

Projects following CATMuS guidelines **MUST** prioritize the characters listed in the CATMuS character table.

Categories of Abbreviations Abbreviations in medieval documents fall into the following categories: - **Tildes** - **Abbreviations using special signs** - **Abbreviations using superscript letters** *Superscript letters are divided into two subcategories:* - **Combining characters:** Letters written directly above another letter, see keyboard in section Tools. - **Superscript characters:** Letters written in superscript position next to a regular letter (e.g., exponent-like forms), see keyboard in section Tools.

We provide a priority list of signs in the table below. Given the diversity of abbreviation representations across time and languages, we make no claim to exhaustivity. See also the subsection on *Graphemic Transcription Principles* in the Generalities section.

Concerning the placement of the tilde, some cases may seem complex, and certain choices are more influenced by habit than by strict rules. In instances where the tilde covers more than one element, there are two possible approaches, depending on the context: - Double the sign, as in *ēē* (esse). - Place the sign in the most appropriate position—either in the middle of the word, or over the consonant or vowel that is modified.

Paris, BnF, lat. 13388, 9th c.	Transcription
	Notā fac mihi uia in qua ambule. quia
Toronto, Thomas Fisher Rare Book Library, MSS 01237, 15th c.	Transcription
	mās apo ⁹ ait. Confidēs te ipm ducē ēē ce

4.4 Modern and Contemporary Documents

4.4.1 General Approach

For older documents where the abbreviation system resembles medieval practices, **the rules for medieval documents MAY apply**. *This allows for consistency when dealing with transitional or hybrid systems.*

For commonly abbreviated letters (e.g., “p” [U+A751], “j” [U+204A], “&” [U+0026]), “d” [U+0256]), they **MAY** be retained for the transcription of more recent sources.

Bâle, Staatsarchiv Basel-Stadt, B 168/15-9.2, 17th c

Transcription



lad Agatte lui a donne le mal p la

4.4.2 Superscript Letters

In modern and contemporary documents, abbreviations often take the form of superscripted letters (e.g., “1st” for *first*, “M^elle” for *Mademoiselle*). The following rules **MUST** be applied :

1. Superscript sequential additions must be transcribed with superscript characters, see keyboard in section Tools. *P.S : Superscript sequential additions have been transcribed in the first datasets using a pseudo markup system based on “^” [U+005E], which marks the beginning of the sequence.*
2. Interlinear corrections **MUST NOT** be confused with superscript sequential additions. If a superscript sequence forms a complete word or a sequence of words, it **MUST** be transcribed on its own line.
3. In modern documents, a unique letter used above a word, as it can be found in medieval documents, **SHOULD** be transcribed with the corresponding combining letter, combined to the character it relates to.

5 Accents and cedillas

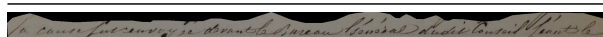
5.1 Accents

5.1.1 Modern and Contemporary Documents

Accents **MUST** be transcribed whenever they appear in the source..

Paris, AD75, DiU10 386, 19th c.

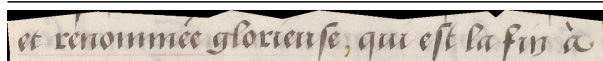
Transcription

la cause fut renvoyé devant le Bureau Général du dit
Conseil séant le

In medieval documents, accents are relatively marginal and **SHOULD** be transcribed when they occur. **Be careful: pointed *i* MUST NOT be transcribed with an acute accent, but simply as a regular *i*.**

Paris, BnF, Arsenal, 5103, 15th c.

Transcription



et renommée glorieuse, qui est la fin à

5.1.2 Medieval Documents

Accents are relatively marginal in medieval documents and **SHOULD** be transcribed when they occur.

Special Cases

- When an accent is misaligned with the letter it belongs to—often the case in cursive writing—the accent **MUST** be transcribed in its expected position, rather than at the place where it was physically drawn.
- When the orientation of an accent is uncertain (e.g., a flat horizontal stroke above a letter), modern accentuation **should** be followed in modern document.
 - In 16th-century documents, the acute accent **MAY** be preferred.

Automatic baseline or polygon detection may exclude an accent traced too far above the line. In such cases, the accent **MAY** nonetheless be transcribed as if it were included within the polygon, anticipating technological improvements that will ensure better inclusion of such marks in future line masks.

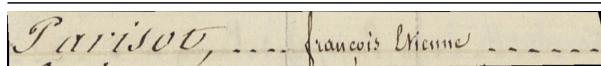
5.2 Cedillas

5.2.1 General Rule

Cedillas **MUST** be transcribed using the combining cedilla [U+0327].

Paris, AN, MC/CM - Projet Lectaurep, 19th c.

Transcription

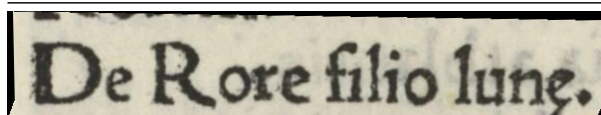


Parisot, —- françois Etienne —-

NB: In Latin manuscripts from the 10th to the 13th centuries, cedilla-marked e's frequently represent the Latin diphthong ae. Cedillas **MUST** therefore be transcribed in medieval French manuscripts as well, in order to ensure consistent encoding.

Paris, Mazarine, Inc. 59, 15th c.

Transcription



De Rore filio lunę.

6 Punctuation

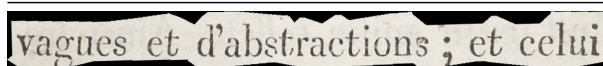
6.1 Punctuation

6.1.1 General Rule

Punctuation in the source **MUST** be transcribed exactly as it appears in the document and **MUST NOT** be normalized. Spacing **MUST** be standardized as follows: - No double spaces. - No space before punctuation marks. - A single space after punctuation marks, *except for quotation marks and apostrophes*.

Paris, BnF, RESP-Y2-2462, 18th c.

Transcription



vagues et d'abstractions; et celui

6.1.2 Medieval Sources

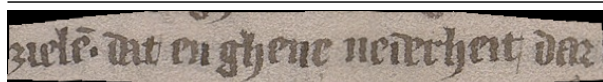
Punctuation transcription in medieval sources **MUST** be simplified to ensure consistency.

Because the dot-based punctuation system displays a wide variety of shapes (simple dot, middle dot, punctus elevatus, double dot, colon, semicolon, etc.), and since their use frequently varies across scribes and manuscripts, the following simplification rules **MUST** be applied:

- *Single dot* **MUST** be transcribed as <.> [U+002E], regardless of its height or position.
- *Double dots* **MUST** be transcribed as <:> [U+003A].

Ghent, UL, 94I, 14th c.

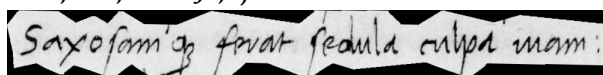
Transcription



zielē. dat/ en ghene nederheit/ dar

Paris, BnF, lat. 8236, 15th c.

Transcription



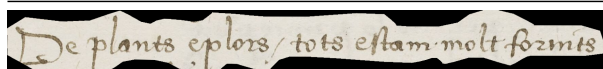
Saxosam q; ferat sedula culpa uiam:

NB: This system *MAY* be adapted to the historical period and to the geographic or linguistic context. For example, the inverted question mark <¿> occurs in Spanish manuscripts, whereas the standard question mark <?> appears in Italian ones. Any such adaptation *MUST* be explicitly documented and justified.

- The sign [U+002F] **MUST** be used for diastoles and virgula suspensiva. *These signs often serve both token segmentation (e.g., words, verses) and indicating a short pause.*

Paris, BnF, esp. 225, 15-16th c.

Transcription

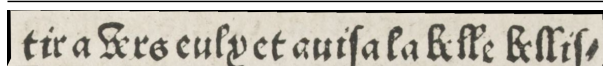


De plants eplors/ tots estam molt fornits

- At the end of a line, all hyphenation marks **MUST** be represented by the sign <-> [U+00AC].

Paris, BnF, Rés. Y2-82, 15th c.

Transcription



tira uers eulx et auisa la belle bellis-

6.1.3 Modern and Contemporary Sources

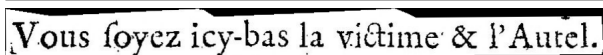
Punctuation marks in modern and contemporary sources **MUST** be transcribed exactly as they appear in the source. This includes periods, commas, exclamation marks, question marks, and similar signs, which **MUST** be represented faithfully.

Standardization of Allographic Variations Some graphical variations are considered allographic. In such cases, the representation **MUST** be reduced to a single standardized sign: - **Dashes**: All dashes, whether short or long, **MUST** be transcribed as `<->` [U+2010]. *This avoids ambiguity and ensures consistency.*

- **Hyphenation at Line Breaks**: All hyphenation marks **MUST** be transcribed as `<->` [U+00AC] [U+2010]. *This avoids confusion with hyphens used for morphological or lexical associations (e.g., “semi-automatique”).*
- **Quotation marks**: Any character functioning as a quotation marker **MUST** be transcribed as the standard Unicode quotation mark `<“>` [U+0022].
- **Apostrophes**: All apostrophe-like characters **MUST** be transcribed as the single quotation mark `<’>` [U+0027].

Paris, BnF, YE-2650, 17th c.

Transcription



Vous soyez icy-bas la victime & l'Autel.

7 Corrections and functional signs

7.1 Functional Marks

Historical documents often contain additional signs indicating textual organization or the status of certain passages. These signs **MUST** be transcribed as faithfully as possible. *However, due to the extensive variety of such signs, certain standardization principles **MUST** be applied to ensure consistency and reproducibility.*

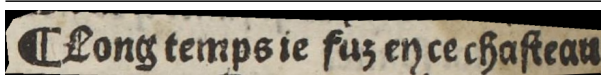
7.1.1 Section marks

Medieval and Modern Documents Section division symbols **MUST** be transcribed using the appropriate character.

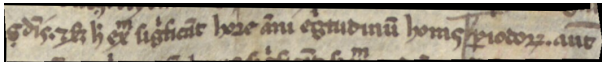
- Pilcrows and similar symbols (reverse pilcrow, ornamental variants, etc.) **MUST** be transcribed as `<¶>` [U+00B6], regardless of orientation. In **modern documents**, other signs **MUST** be transcribed as as faithfully as possible. In most of the cases, the standardized sign `<§>` [U+00A7] **SHOULD** be used to indicate sections or paragraphs.

Paris, BnF, Rés. Y2-930, 15th c.

Transcription



¶ Long temps ie fuz en ce chasteau

Paris, BnF, Rés. Y2-930, 15th c.	Transcription
Munich, Bayerische Staatsbibliothek, CLM 13027, 13th c.	Transcription
	s dīs. 7 f; h ex sigficāt hore āni eg̃tudinū horas ¶ piodoꝛ. aūt

7.1.2 Reference Marks

Medieval Documents

- *Insertion signs* **MUST** be transcribed with the caret sign <^> [U+2038].
- *Reference marks* (such as <> [U+002A] or <✕> [U+203B]) **MUST** be normalized to <> [U+002A].
- *Manicules* **MUST** be transcribed as <☞> [U+261E].

If there is an interlinear addition, it **MUST** be considered a separate text line and transcribed accordingly using the SegmOnto vocabulary.

Modern and Contemporary Documents Reference marks **MUST** be reproduced as they appear in the source. When in doubt, medieval norms **SHOULD** be applied.

Paris, BnF, lat. 16204, 13th c.	Transcription
	et p̃ponas mediū celū ascendentī. nec .ū. i. x̃.
Bamberg, Staatsbibliothek, MS Class. 30, 9th c.	Transcription
	* hęc duo uerba dysyllaba ñ excipunt' ċiugatione s̃ p̃dicta *
Paris, BnF, 4-S-1534, 18th c.	Transcription
	(*) Mémoires de l'Académie Royale des Sciences, 1736.

7.1.3 Correction marks

For in-line corrections, the combination <⌈> [U+27E6] and <⌋> [U+27E7] **MUST** be used to encapsulate a word marked as crossed out. If the word is not decipherable, as many dots as the number of missing letters **MAY** be added.

If the word is not decipherable (due to a stain, strokes, or any other reason), the combination <[> [U+005B] and <]> [U+005D] **MUST** be used empty, optionally adding as many dots as the presumed number of missing letters.

Paris, BnF, fr. 12554, 15th c.	Transcription
	Regarda le ⌈chapelain⌋ et cōmanda quil feust serui
Bâle, Staatsarchiv Basel-Stadt, B 168/15-2.2, 17th c.	Transcription
	channes de vin. dont le ⌈lieu⌋ mesme Jour.

7.1.4 Unreadable text

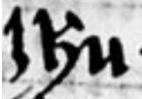
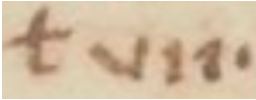
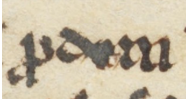
If the text is unreadable due to damage to the support (and not because of a correction), the combination <[> [U+005B] and <]> [U+005D] **MUST** be used to encapsulate the unreadable portion, optionally adding as many dots as the number of missing letters.

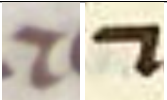
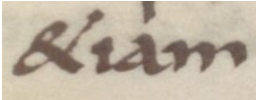
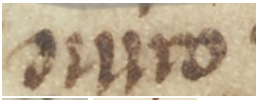
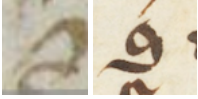
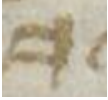
Paris, BnF, Velins 1103, 16th c.	Transcription
	Sur lesquelz estoit escr[]pt. Dēs isti sūt

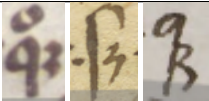
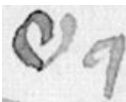
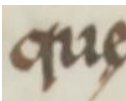
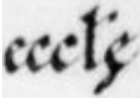
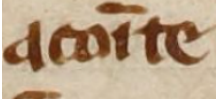
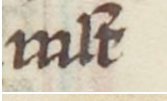
8 Table of characters

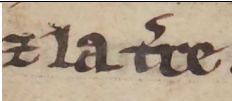
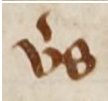
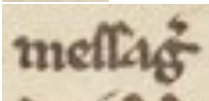
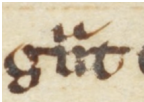
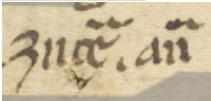
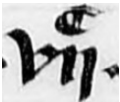
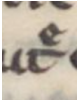
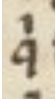
8.1 Character Table

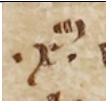
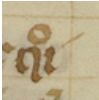
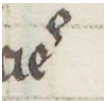
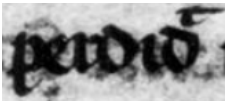
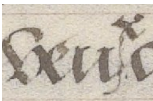
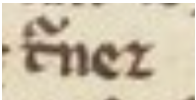
This table is currently limited to Middle Ages.

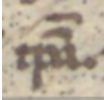
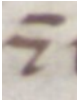
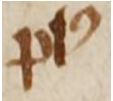
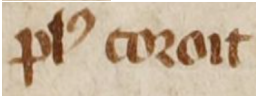
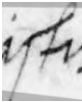
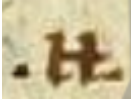
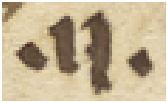
Class	Character(s)	Unicode	Name	Label	Examples	Allographs
Abbreviation (Strikes)	đ	U+0111	LATIN SMALL LETTER D WITH STROKE	d + any desinence truncation		
Abbreviation (Strikes)	ḥ	U+0127	LATIN SMALL LETTER H WITH STROKE			
Abbreviation (Strikes)	ḷ	U+0142	LATIN SMALL LETTER L WITH STROKE			
Abbreviation (Strikes)	p̣	U+A751	LATIN SMALL LETTER P WITH STROKE			
Abbreviation (Strikes)	p̣	U+A753	LATIN SMALL LETTER P WITH FLOURISH			
Abbreviation (Strikes)	q̣	U+A759	LATIN SMALL LETTER Q WITH DIAGONAL STROKE			
Abbreviation (Strikes)	ř	U+A775	LATIN SMALL LETTER RUM	-rum	     	ř
Abbreviation (Strikes)	f̣	U+1E9C	LATIN SMALL LETTER LONG S WITH DIAGONAL STROKE			

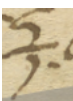
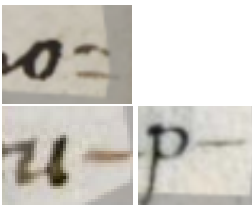
Class	Character(s)	Unicode	Name	Label	Examples	Allographs
Abbreviation	]	U+204A	TIRONIAN SIGN ET	Et		
						
						
						
Abbreviation	&	U+0026	Ampersand	Et		
Abbreviation	ſ	U+A76D	LATIN SMALL LETTER IS	-is		
Abbreviation	9	U+A76F	LATIN SMALL LETTER CON	con		
						
Abbreviation	≈	U+2248	ALMOST EQUAL TO	esse		
Abbreviation	÷	U+00F7	DIVISION SIGN	est/id est		
						

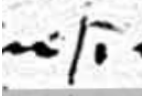
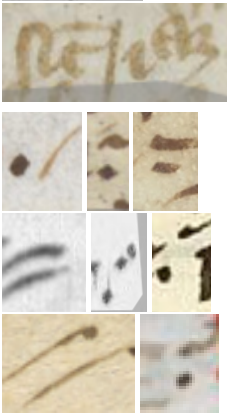
Class	Character(s)	Unicode	Name	Label	Examples	Allographs
Abbreviation	;	U+F1AC	LATIN ABBREVIATION SIGN SEMICOLON	-que/-bus/-m/-et	   	z, 3
Abbreviations	∅	U+2205	EMPTY SET	instans		
Abbreviation	☛	U+2661	WHITE HEART SUIT	coeur		
Combining Abbreviation	,	U+0327	COMBINING CEDILLA	ξ = ae	 	
Combining Abbreviation	~	U+0303	COMBINING TILDE	Nasalisation (-_m, -_n), suspension, contraction.	 	

Class	Character(s)	Unicode	Name	Label	Examples	Allographs
Combining Abbreviation	ˆ	U+033E	COMBINING VERTICAL TILDE(	-r, -er	  	4
Combining Abbreviation	ˆ	U+0363	COMBINING LATIN SMALL LETTER A			
Combining Abbreviation	ˆ	U+1DD3	FLATTENED OPEN A ABOVE		 	
Combining Abbreviation	ˆ	U+0368	COMBINING LATIN SMALL LETTER C			
Combining Abbreviation	ˆ	U+0364	COMBINING LATIN SMALL LETTER E			
Combining Abbreviation	ˆ	U+0365	COMBINING LATIN SMALL LETTER I	qui		

Class	Character(s)	Unicode	Name	Label	Examples	Allographs
Combining Abbreviation	m	U+036B	COMBINING LATIN SMALL LETTER M			
Combining Abbreviation	o	U+0366	COMBINING LATIN SMALL LETTER O			
Combining Abbreviation	q	U+F033	COMBINING LATIN SMALL LETTER Q			
Combining Abbreviation	r	U+036C	COMBINING LATIN SMALL LETTER R			
Combining Abbreviation	s	U+iDE4	COMBINING LATIN SMALL LETTER S			
Combining Abbreviation	t	U+036D	COMBINING LATIN SMALL LETTER T			
Combining Abbreviation	x	U+036F	COMBINING LATIN SMALL LETTER X			
Combining Abbreviation	ʷ	U+iDD1	COMBINING UR ABOVE			

Class	Character(s)	Unicode	Name	Label	Examples	Allographs
Combining Abbreviation	ʹ	U+1DD2	MODIFIER US ABOVE	-us, -os (does not replace MODIFIER LETTER US, below)		
Combining Abbreviation	̃̃	LETTER + U+0303 & LETTER + U+0303	COMBINING TILDE on two consecutive characters	ẽẽ = esse	 	
Combining Abbreviation	̃̃	LETTER + U+0303 + U+0303	Two stacked combining tildes			
Combined Abbreviation	̃j	U+204A + U+0303	TIRONIAN SIGN ET + COMBINING TILDE	Etiam	  	
Superscript Abbreviation	ʹ	U+A770	MODIFIER LETTER US	-us, -os	 	
Ligature	st	-	st			
Ligature	.n.	-	-	enim	 	

Class	Character(s)	Unicode	Name	Label	Examples	Allographs
Ligature	qr	-		quia		
						
Ligature	qd	-		quod		
Ligature	Et	-		Et		
Symbols	℥	U+2125	OUNCE SIGN	Ounce		
Symbols	℥	U+10192	ROMAN SEMUNCIA SIGN	Semi-Ounce		
Punctuation	¶	U+00B6	PILCROW SIGN	Content change	  	¶ ¶ ¶
Punctuation	-	U+002D	HYPHEN-MINUS	Hyphenation		

Class	Character(s)	Unicode	Name	Label	Examples	Allographs
Punctuation	/	U+2215	DIVISION SLASH	Diastole		
Punctuation	:	U+003A	COLON	Punctus elevatus		
Punctuation	:	U+003A	COLON	Punctus interrogativus		
Reference mark	^	U+2038	CARET	Omission sign 'caret'(reintroduction of content)		

9 Datasets

9.1 Medieval Dataset

- HuggingFace (Samples)

The following datasets are in the current version of the HuggingFace dataset, as seen in the paper from CAT-MuS Medieval.

All data were passed through Choco Mufin. Any attempt to replicate the HuggingFace dataset should use the conversion table provided with each dataset. The conversion table for the modified dataset will be made available.

language	url	version
Latin	DEEDS-Project/htr-dataset	v0.0.8
Latin	HTRomance-Project/medieval-latin	v0.0.8
Spanish Languages	HTRomance-Project/middle-ages-in-spain	v0.0.6
Old/Middle French	HTRomance-Project/medieval-french	v0.0.9
Italian Languages	HTRomance-Project/medieval-italian	v1.0.2
Old/Middle French	HTR-United/cremma-medieval	v2.0.1
Latin	HTR-United/cremma-medieval-lat	v0.1.2
Old/Middle French	Gallicorpora/HTR-imprime-gothique-16e-siecle	v0.0.19
Old/Middle French	Gallicorpora/HTR-MSS-15e-Siecle	v0.0.37
Old/Middle French	Gallicorpora/HTR-incunable-15e-siecle	v0.0.29
Old/Middle French	ciham-htr/fabliaux	v0.0.22
Old/Middle French	ciham-htr/liber	v0.0.5
Latin/Italian/French	Reorganized from HN2021 Boccace	last
Old/Middle French	Reorganized [Decameron-Fr]	last
Latin	Reorganized malamatenia/Eutyches	last
Spanish Languages	Reorganized & Augmented Gille-Levenson's PhD Data	last
Latin	Adapted rescribe/carolineminuscul-groundtruth	2023-10-03

9.1.1 Vocabulary:

- **Adapted:** Image were replaced / rescaled
- **Reorganized:** Data were reorganized in a way that made them more usable
- **Augmented:** Data whose digitization were not publicly available where used to extract lines.