

Transcription Questions, Problems, and Current Rules

tl;dr: Our transcription rules are as follows:




1. Our transcription collapses multiple letter forms with identical meaning.
2. Our transcription renders miniscule “should-be” capitals as minuscule letters.
3. Our transcription does not expand abbreviations, but rather marks them with Unicode diacritics.

Overview

As we compile our training transcriptions, it is essential that we have consistent rules for rendering letters and abbreviations. However, part of what makes medieval handwriting so challenging for OCR is its various letter forms and obscure abbreviations. We have identified three primary transcriptions “issues” for our manuscript: 1) multiple letter forms 2) lowercase letters used in uppercase instances 3) extensive abbreviations. We discuss these issues in detail below, and also explain how we will handle them in our transcription.

Multiple Letter Forms

Anglicana formata, like other medieval bookhands, has multiple forms for certain letters. For example, the “s” can take either a kidney, sigma, or long shape:

Manuscript Image	Character Description	Furnivall’s Transcription	Unicode Option	Our Transcription
	kidney s	s	none	s
	sigma s	s		s
	long s	s	f	s

These different letter forms do not affect meaning in any way, but are either aesthetic or functional; that is, they often depend on the place of “s” in a word or line (e.g. typically long-s can be either word-initial or word-final, whereas the kidney-shape s tends to be word-final, and the sigma only word-initial).

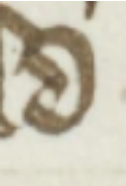
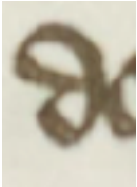
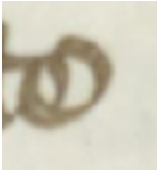
We can either transcribe these forms individually, or collapse them together.




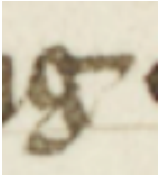
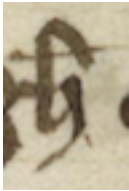

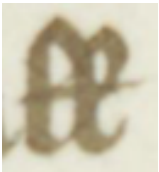
Individually transcribing these forms is preferable if it helps the OCR system identify them correctly, and if our goal is produce a very conservative digital replication of the manuscript. However, if we were to transcribe these forms uniquely, we would use different Unicode characters for each form, and certain forms do not have a matching Unicode characters (see the kidney “s” above); while matching manuscript-character and Unicode-character is not necessary (we could edit machine-made transcriptions manually later), it is still desirable.


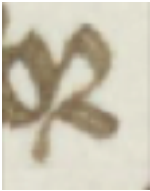
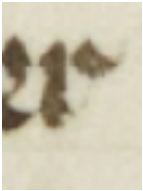
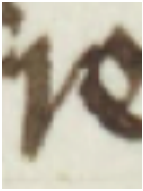


Collectively transcribing these forms is preferable if our goal is to create a digital edition of the manuscript that is accessible to a modern reader. As well, collapsing these forms is easier for creating training data (since it is tedious to insert special characters).

Since Kraken is able to collapse forms, and these varieties do not affect semantics, **we will be collapsing them.**

For more examples of multiple letter forms, see the below chart, as well as our full character guide for Scribe D:

Manuscript Image	Character Description	Furnivall's Transcription	Unicode Option	Our Transcription
	d open loop	d	none	d
	d closed loop	d	none	d
	chickpea e	e	none	e

Manuscript Image	Character Description	Furnivall's Transcription	Unicode Option	Our Transcription
	regular e	e	e	e
	f with tick	adds a tick	none	f
	biting f's	ff (connected stroke)	ff (connected stroke)	ff (disconnected strokes)
	g with extended stroke	adds tick	none	g
	h with stroke	h	h	h
	normal h	h	h	h
	Welsh l			ll

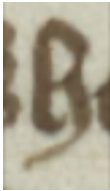
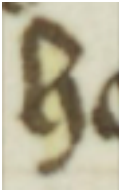
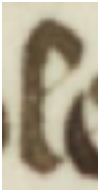

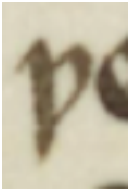
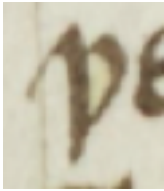
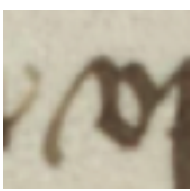
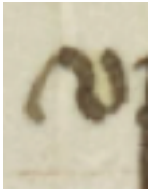
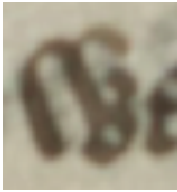
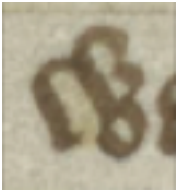
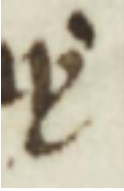
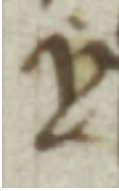
Manuscript Image	Character Description	Furnivall's Transcription	Unicode Option	Our Transcription
	biting p's	pp	none	pp
	2-shaped r	r	2	r
	miniscule r	r	r	r
	split r	r	none	r
	t with tick	adds a tick	none	t
	y dot	y	ȳ	y

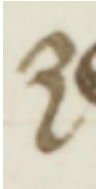
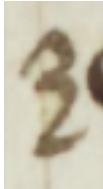
Lowercase and uppercase

It is often hard to distinguish between lowercase (minuscule) and uppercase (capital) letter forms. At times the scribe uses completely different graphs (for instance **B** and **b**, **R** and **r**, **S** and **s**); but often the distinction seems too subtle to be perceivable (for example, **W** and **w**, **P** and **p**). We may assume from location or semantic context that a letter is “meant” to be a capital, for example in a proper name or at the beginning of a line. And accordingly, our reference, Furnivall, often marks a capital where the manuscript presents a miniscule form.

But it will be difficult (if not impossible) for an OCR system to identify such a distinction, at least without a certain level of language processing. Additionally, the scribe often shades uppercase letters slightly to mark their capital form, but this shading will not be visible to Kraken, which receives a binarized (black and white) photo.

Here are some examples:


Letter	Minuscule	Shaded Capital
h		
l		
p		
v		
w		
y		


Letter	Minuscule	Shaded Capital
		

Because these differences appear negligible, and will probably be difficult for the machine to discern, we will be transcribing such characters in their lowercase form.

Note: there are formal capitals for **V** for and **H**, but often the scribe uses the miniscule forms instead. **L**, **P**, **W**, **Y** (above) do not seem to have distinct capital forms, nor do **J**, **K**, **U**, **X**, or **Z**. For capital **F**, the scribe uses a double f, which we will transcribe ff. An expanded chart of Scribe D's character forms can be found [here](#).

As well, these two forms are somewhat ambiguous:

Letter	Potential	
	Capital	Note
h		significantly larger and slightly different in shape compared to other hs

Letter	Potential Capital	Note
d		extended loop com- pared to other ds , and Mooney identi- fies this as as capital D

Given that these two forms appear exaggerated enough, and especially that this **D** has been identified by paleographers as a capital (see Late Medieval English Scribal Profile of Scribe D), we will be transcribing these two forms as capitals.



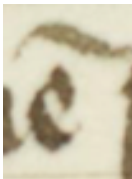
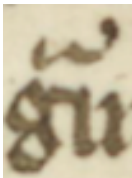
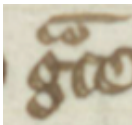
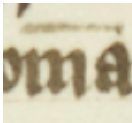
In conclusion, any character that “should” be a capital (semantically or geographically), but *looks* nearly or totally indistinguishable from the the miniscule, will be transcribed as a miniscule. To see examples of “proper” capitals and miniscules, consult the expanded character chart.


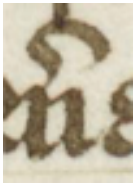

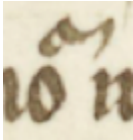

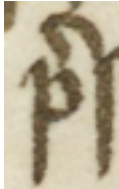
Abbreviations

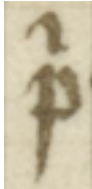
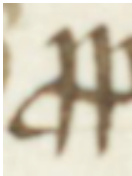
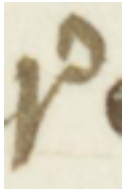
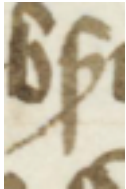

Medieval manuscripts have many abbreviations. Usually transcriptions will extend these abbreviations, marking the abbreviated letters with italics. However, since Kraken requires diplomatic transcription (one-to-one characters), we can not extend the abbreviations in our training data. And since abbreviations *do* carry semantic meaning (unlike the various letter forms above), we will be using special characters and diacritics to represent abbreviations.


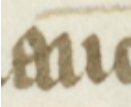
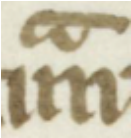
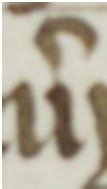
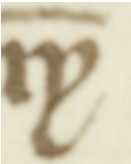
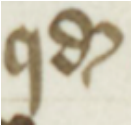
Certain diacritic choices here are quite obvious: for example, a macron over an **o** can be easily represented with **ō**. Other abbreviations have no corresponding Unicode character (see, for example, **ps** with right hook and left hook suprascripts). For these forms, we use Unicode diacritics that most closely resemble the abbreviations. We have also tried to be as consistent as possible

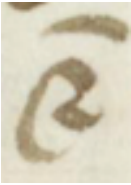

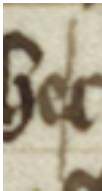
with these abbreviations: for example, we use the same diacritic for the right hook (“er” or “re” abbreviation) over the **p**, **n**, **t**, **p**, and **u**.

Manuscript Image	Abbreviation Description	Abbreviation Extension	Furnivall's Tran- scription	Alternative Option	Temporary Rule
	character resembling 3	us (i.e. bus)	<i>bus</i>	(Latin small et)	
	e with suprascript	m (i.e. em)	<i>em</i>	ě (e with hook diacritic)	ě
	e with macron	n (i.e. en)	<i>en</i>	ē	ē
	g with tilde or with macron and tilde	ra (i.e. gra)	<i>gra</i>	ḡ	ḡ
	m with macron	m (i.e. mm)	<i>mm</i>	m̄	m̄
	macron over on or n	u (i.e. oun)	n̄	ōn̄ (over o as well)	n̄

Manuscript Image	Abbreviation Description	Abbreviation Extension	Furnivall's Tran- scription	Alternative Option	Temporary Rule
	n with tilde	ra (i.e. ran)	<i>ran</i>	\tilde{n}	\tilde{n}
	n with right hook suprascript	er (i.e. ner)	<i>ner</i>	n (inverted comma diacritic)	n
	macron over o	m (i.e. om)	<i>om</i>	\bar{o}	\bar{o}
	o with tilde	ur (i.e. our)	<i>our</i>	\tilde{o}	\tilde{o}
	p with stroke through descender	er or ar (i.e. per or par)	<i>per</i> or <i>par</i>		and
	p with right hook suprascript	er (i.e. per) or re (i.e. pre)	<i>per</i> or <i>pre</i>	p (inverted comma diacritic)	p

Manuscript Image	Abbreviation Description	Abbreviation Extension	Furnivall's Tran- scription	Alternative Option	Temporary Rule
	p with left hook suprascript	pri	<i>pri</i>	p (comma diacritic) \dot{p} (hook diacritic)	\dot{p}
	p with loop on descender	ro (i.e. pro)	<i>pro</i>	(p with flourish)	
	r with left hook suprascript	e (i.e. re)	<i>re</i>	\dot{r}	\dot{r}
	long s with diagonal stroke	ser	<i>ser</i>	(long s with diagonal stroke)	
	t with right hook suprascript	er (i.e. ter)	<i>ter</i>	t	t

Manuscript Image	Abbreviation Description	Abbreviation Extension	Furnivall's Tran- scription	Alternative Option	Temporary Rule
	thorn with right hook suprascript	er (i.e. þer)	þer	þ (inverted comma diacritic)	þ
	u with macron	n (i.e. un)	un	ū	ū
	u with macron and tilde	ra (i.e. ura)	ura	ũ (macron and diaeresis)	ũ
	u with right hook suprascript	er (i.e. uer)	uer	ŭ (hook diacritic) u (inverted comma diacritic)	u
	y with macron	n (i.e. yn)	yn	ȳ	ȳ
	d with endstroke	quod	quod	qd (add nothing)	qd

Manuscript Image	Abbreviation Description	Abbreviation Extension	Furnivall's Tran- scription	Alternative Option	Temporary Rule
	tironian et	and	<i>and</i>		
	paragraph marker	paragraph	¶	¶	¶
	vertical bar	divides cramped words	nothing (space)		

“Furnivall’s transcription” refers to *The Corpus MS of Chaucer’s Canterbury Tales* ed. by Frederick J. Furnivall, a print transcription of MS 198 from 1868-79

Unicode Resources

MUFI
Unicode and Macron

Unicode HTML Codes

(add semi-colon to end of codes)

comma diacritic = ̓

hook diacritic = ̉ reverse comma diacritic = ̔

double macron = ͞ or ͞