Transcription and Editorial Interventions

Magdalena Turska

September 2014

Transcription: a special kind of reading

What is the goal of a transcription?

- to make a primary source accessible ...
- ... and comprehensible
- which may imply adding or using much additional information

Hence,

- all transcription is selective
- all transcription is imaginative

Transcription

In transcription for the digital edition, some textual phenomena which commonly attract editorial attention:

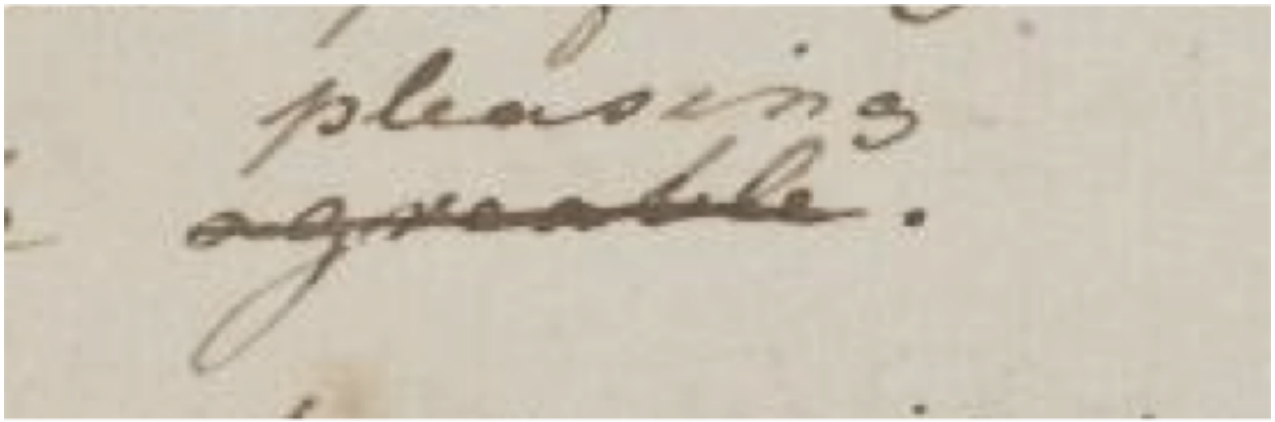
- original layout information
- abbreviations or other arcana
- ‘evident’ errors which invite correction or conjecture
- scribal additions, deletions, substitutions, restorations
- non-standard orthography (etc.) which invites normalisation
- irrelevant or non-transcribable material
- passages which are damaged or illegible

TEI Transcription

- <teiHeader>: provides metadata for the whole thing, at various levels, notably including a <msDesc>
- <text>: contains a structured reading of a document's intellectual content ... its ‘text’
- <facsimile>: organizes a set of page images representing a document
- <sourceDoc>: a non-interpretative transcription of a physical document, e.g. for a *dossier génétique*

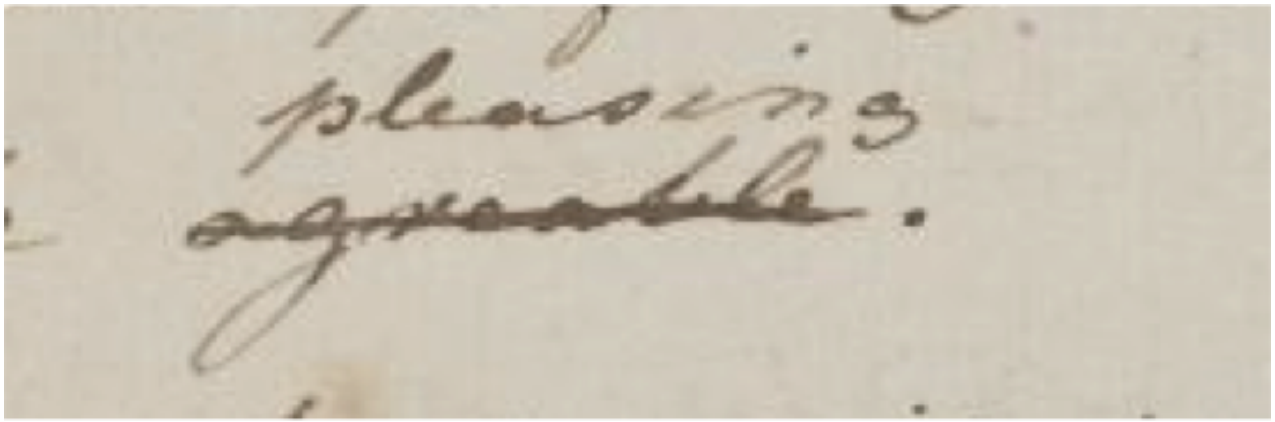
Does a transcription encode a ‘text’ or a ‘document’ ?

What's this ?



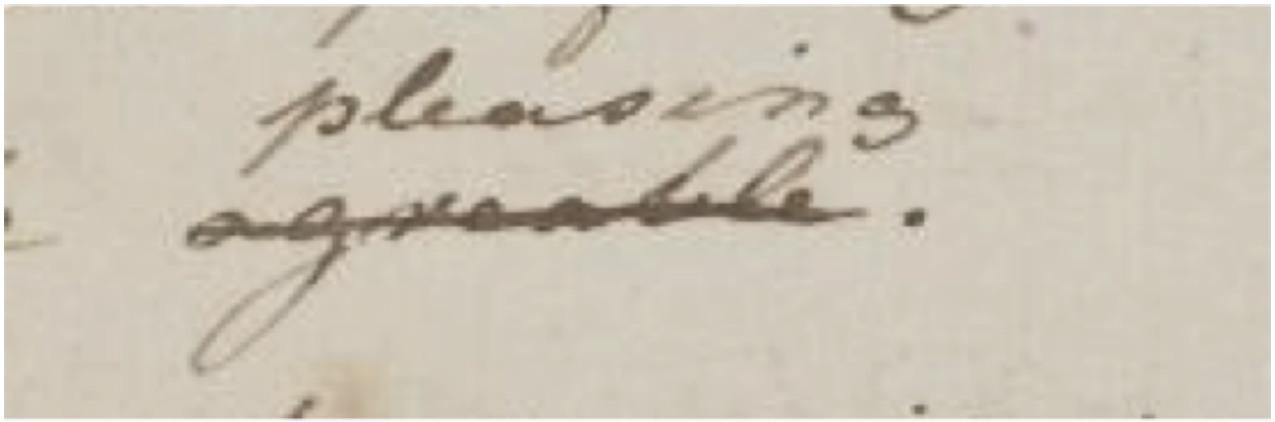
1. ‘agreable’ is struck-through, ‘pleasing’ is written above it, in the interlinear space.
2. ‘agreable’ is deleted and replaced by ‘pleasing’
3. Originally, the text read ‘agreable’, but at some subsequent stage this word was deleted; the word ‘pleasing’ was added in the same context.

What's this ?



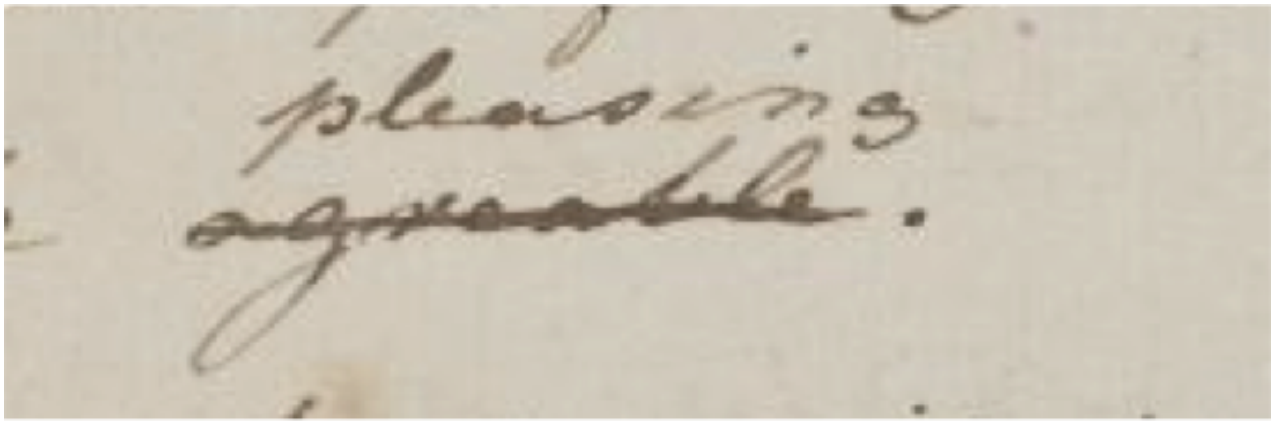
1. ‘agreable’ is struck-through, ‘pleasing’ is written above it, in the interlinear space.
2. ‘agreable’ is deleted and replaced by ‘pleasing’
3. Originally, the text read ‘agreable’, but at some subsequent stage this word was deleted; the word ‘pleasing’ was added in the same context.

What's this ?



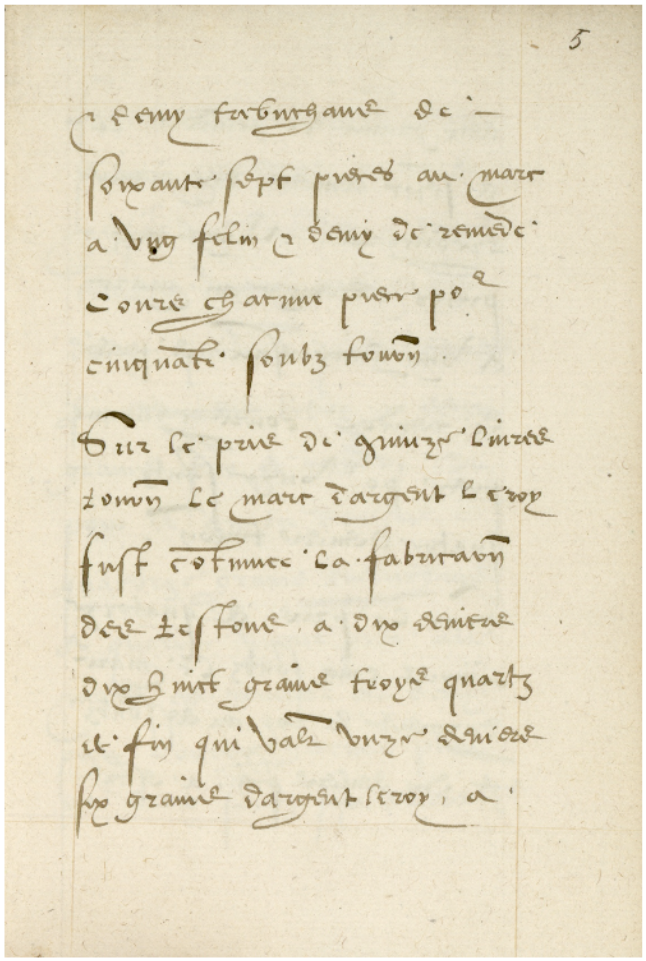
1. ‘agreeable’ is struck-through, ‘pleasing’ is written above it, in the interlinear space.
2. ‘agreeable’ is deleted and replaced by ‘pleasing’
3. Originally, the text read ‘agreeable’, but at some subsequent stage this word was deleted; the word ‘pleasing’ was added in the same context.

What's this ?



1. ‘agreable’ is struck-through, ‘pleasing’ is written above it, in the interlinear space.
2. ‘agreable’ is deleted and replaced by ‘pleasing’
3. Originally, the text read ‘agreable’, but at some subsequent stage this word was deleted; the word ‘pleasing’ was added in the same context.

A typical minimal encoding

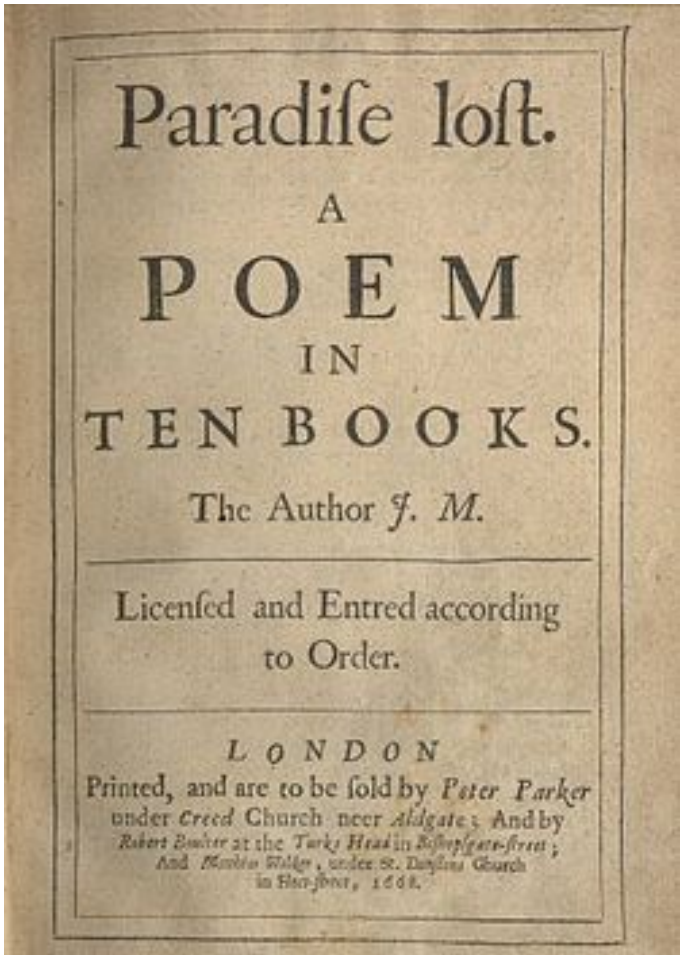


```
<p>
<!-- ... -->
<pb n="5r"/><fw>5</fw>
<lb/><expan>et</expan> demy trebuchans de
<lb/>soixante sept pieces au marc
<lb/>a ung felin <expan>et</expan> demy de
remede
<lb/>Cours chacune piece
<expan>pour</expan>
<lb/><expan>cinquante</expan> soubz
<expan>tournois</expan> <pc>.</pc>
</p> <p>
<lb/>Sur le pris de quinze livres
<lb/><expan>tournois</expan> le
marc dargent le roy
<lb/>fust <expan>continuee</expan> la
<expan>fabricacion</expan>
<lb/>des testons a dix deniers
<lb/>dix huict grains troys quartz
<lb/>de fin qui <expan>valent</expan> unze
deniers
<lb/>six grains dargent le roy<pc>,</pc> a
```

Character annotation

- distinguish allographical forms of a letter
- represent non-standard characters

<g> element (character or glyph) represents a glyph, or a non-standard character


Paradiſe loſt.

A

POEM

IN

TEN BOOKS.

The Author J. M.

Licenſed and Entred according to Order.

LONDON

Printed, and are to be ſold by Peter Parker under Creed Church neer Aldgate; And by Robert Boulter at the Turks Head in Biſhopſgate-ſtreet; And Matthias Walker, under St. Dunſtans Church in Fleet-ſtreet, 1668.

Abbreviations &c.

In Western MSS, we commonly distinguish :

Suspensions the first letter or letters of the word are written, generally followed by a point : for example ‘e.g.’ for ‘exempla gratia’

Contractions both first and last letters are written, generally with some mark of abbreviation such as superscript strokes, or points : e.g. ‘Mr.’ for ‘Mister’

Brevigraphs Special signs such as the Tironian *nota* used for ‘et’, the letter p with a barred tail used for ‘per’, the letter c with a circumflex used for ‘cum’ etc.

Superscripts Superscript letters (vowels or consonants) used to indicate various kinds of contraction: e.g. ‘w’ followed by superscript ‘ch’ for ‘which’.

Most of the symbols needed are available in Unicode, though not necessarily in all fonts.

Abbreviation and Expansion

An abbreviation may be viewed in two different ways:

- as a particular sequence of letters or marks upon the page: thus, a ‘p with a bar through the descender’, a ‘superscript hook’, a ‘macron’
- as another way of representing the letter or letters it is believed to be standing for: thus, ‘per’, ‘re’, ‘n'

Two Levels of Encoding Abbreviations

TEI proposes elements for two levels of encoding:

- the whole of an abbreviated word and the whole of its expansion may be marked using <abbr> and <expan> respectively
- abbreviatory signs or characters and the ‘invisible’ characters they imply may be marked using <am> and <ex> respectively

Compare ev(er)y (per)sone

```
 ev<choice>
  <am>
    <g  ref="#b-er"/>
  </am>
  <ex>er</ex>
</choice>y
<choice>
  <am>
    <g  ref="#b-per"/>
  </am>
  <ex>per</ex>
</choice>sone
...
```

```
<choice>
  <abbr>ev<am>
      <g  ref="#b-er"/>
    </am> y</abbr>
  <expan>ev<ex>er</ex>y</expan>
</choice>
<choice>
  <abbr>
    <am>
      <g  ref="#b-per"/>
    </am>sone</abbr>
  <expan>
    <ex>per</ex>sone</expan>
</choice>
```

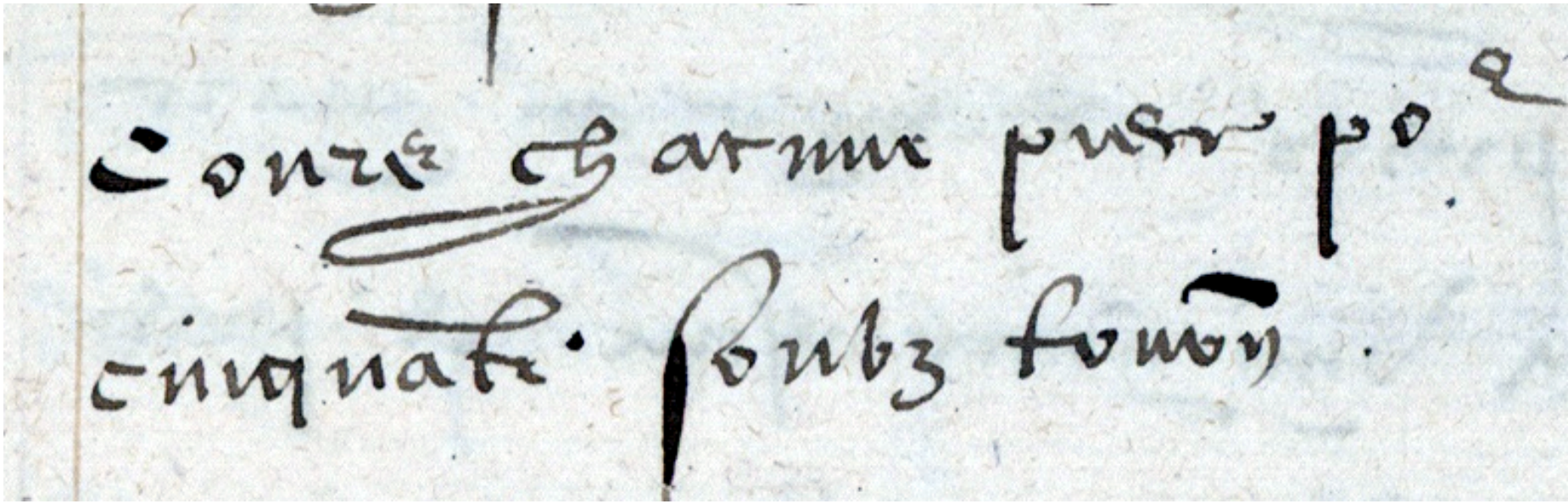
<ex> and <am>

Using these elements, from the 'transcr' module, a transcriber may indicate the status of the individual letters or signs within both the abbreviation and the expansion.

- <ex> (editorial expansion) contains a sequence of letters added by an editor or transcriber when expanding an abbreviation.
- <am> (abbreviation marker) contains a sequence of letters or signs present in an abbreviation which are omitted or replaced in the expanded form of the abbreviation.

Previously, people have re-purposed existing elements such as <hi> and <supplied> to mark individual letters/signs in abbreviations and expansions. The new P5 elements <am> and <ex> are the TEI's attempt to support this desire.

A simple example (1)



Editorial strategy may be simply to note that we have expanded the abbreviations:

```
<p>
  <lb/>Cours chacune piece <expan>pour</expan>
  <lb/>
  <expan>cinquante</expan> soubz
<expan>tournois</expan>
  <pc>.</pc>
</p>
```

A simple example (2)

As you noticed, ‘pour’ was actually written ‘po’ followed by an ‘r’ subscript; ‘cinquante’ as ‘cinquāte’ with a macron on the ‘a’ to indicate nasalisation. We could therefore encode as follows:

```
<p>
  <abbr>po&#xFFFD;</abbr> .... <abbr>cinqu&#x0101;te</abbr>
</p>
```

... or we could choose one of the following styles:

```
<p> po<am>&#xFFFD;</am> ... or po<ex>u</ex>r </p>
```

```
<abbr>po<am>&#xFFFD;</am>
</abbr>
```

```
<expan>po<ex>u</ex>r</expan>
```

Simple example (3)

And of course TEI permits both cake and the eating of it:

```
<p> po<choice>
    <am>&#xFFFD;</am>
    <ex>ur</ex>
  </choice>
</p>
```

```
<choice>
  <abbr>po<am>&#xFFFD;</am>
  </abbr>
  <expan>po<ex>u</ex>r</expan>
</choice>
```

Choice

Children of a choice element all represent alternative ways of encoding the same sequence and in most cases they are mutually exclusive.

```
<choice>
  <abbr>Dat</abbr>
  <expan>Dat<ex>ae</ex>
  </expan>
</choice>
```

Where the purpose of an encoding is to record textual variants, rather than to identify multiple possible encoding decisions, the app element & company should be preferred.

```
<app>
  <rdg>
    <expan>Dat<ex>ae</ex>
    </expan>
  </rdg>
  <rdg>
    <expan>Dat<ex>um</ex>
    </expan>
  </rdg>
</app>
```

A <choice> reminder

- <choice> (groups alternative editorial encodings)
- Abbreviation:
  - <abbr> (abbreviated form)
  - <expan> (expanded form)
- Errors:
  - <sic> (apparent error)
  - <corr> (corrected error)
- Regularization:
  - <orig> (original form)
  - <reg> (regularized form)

Classifying abbreviations

The *@type* attribute on <abbr> is a useful way of categorising abbreviations, whether for statistical purposes, or to allow for different types to be rendered differently:

```
<choice>
  <abbr  type="brevigraphe">po<am>&#xFFFD;</am>
  </abbr>
  <expan>po<ex>u</ex>r</expan> en <choice>
    <abbr  type="suspension">fin<am>.</am>
    </abbr>
    <expan>fin<ex>ir</ex>
    </expan>
  </choice>
</choice>
```

This encoding might be displayed as : ‘po(u)r en fin*ir*’

As elsewhere, the *@resp* and *@cert* attributes can also be used to indicate who is responsible for an expansion, and the degree of certainty attached to it.

Corrections and emendations

The <sic> element can be used to indicate that the reading of the manuscript is erroneous or nonsensical, while <corr> (correction) can be used to provide what in the editor's opinion is the correct reading:

```
<sic>relea</sic>
```

```
<corr>relicta</corr>
```

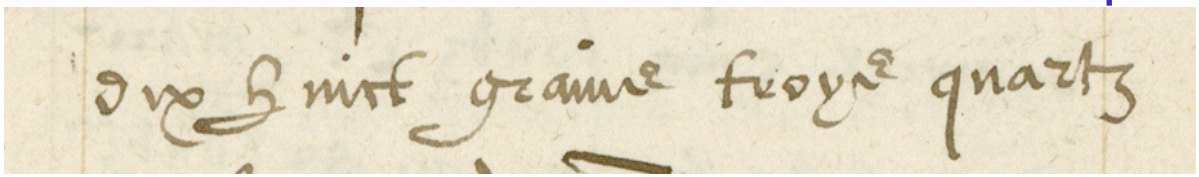
The two may, of course, be combined within a <choice> element:

```
<choice>
  <sic>relea</sic>
  <corr  cert="high">relicta</corr>
  <corr  cert="low">relatio</corr>
</choice>
```

Normalization

Source texts rarely use modern orthography. For retrieval and other processing reasons, however, the modernized form may be very. The <reg> (regularized) element is available used to mark a normalized form; the <orig> (original) element to indicate a non-standard spelling. These elements can optionally be grouped as alternatives using the <choice> element:

Normalisation example



```
<lb/>dix <choice>
  <orig>huict</orig>
  <reg>huit</reg>
</choice> grains

<choice>
  <orig>troys
     quartz</orig>
  <reg>trois-quart</reg>
</choice>
```

In this case, a further semantic regularisation is possible :

```
<lb/>
<measure  quantity="18.75"  unit="gr">dix
<choice>
    <orig>huict</orig>
    <reg>huit</reg>
  </choice> grains
<choice>
    <orig>troys
         quartz</orig>
    <reg>trois-quart</reg>
  </choice>
</measure>
```

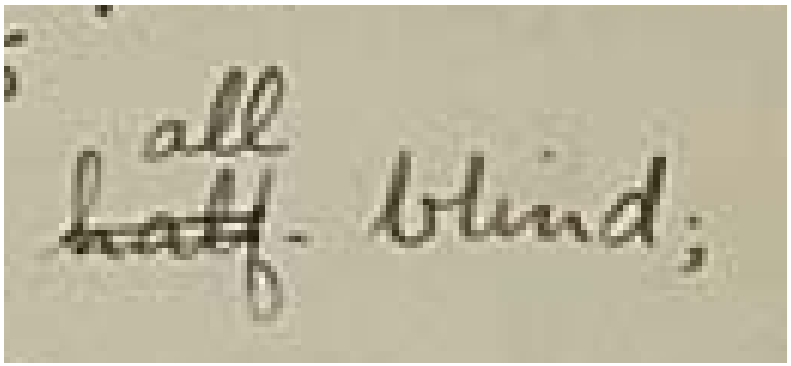
Additions, deletions, substitutions, and modifications

Alterations made to the text, whether by the scribe or in some later hand, can be encoded using <add> (addition) or <del> (deletion).

Where the addition and deletion are regarded as a single *substitution*, they can be grouped together using the <subst> (substitution) element :

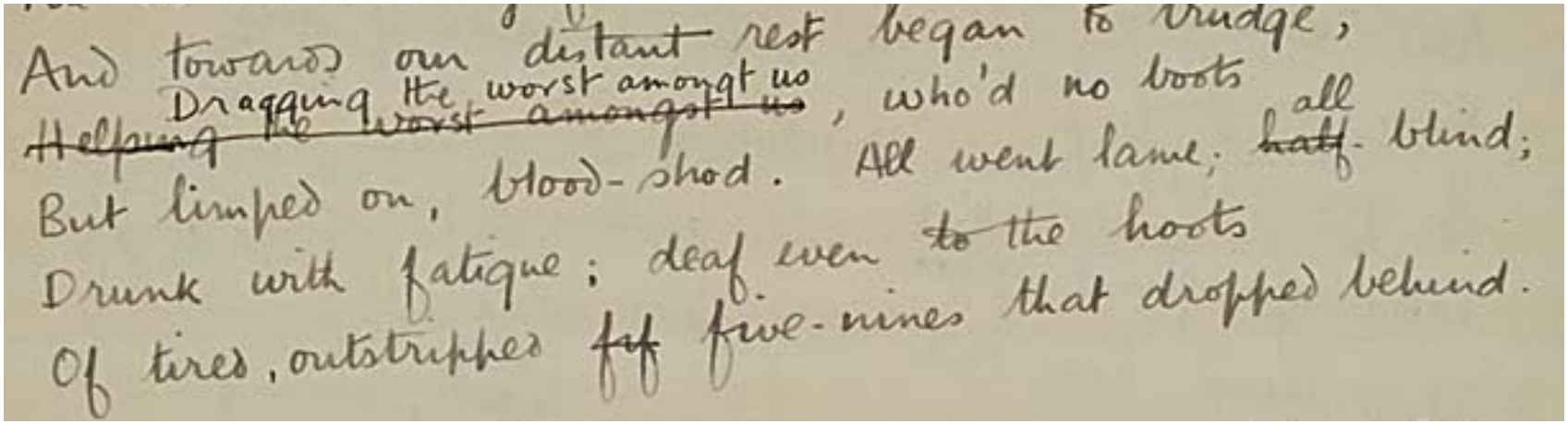
- <add> (addition) or <del> (deletion) are used for evident alterations in the source
- a combined addition and deletion may be marked using <subst> (substitution)
- <mod> (modification) represents any kind of general modification without interpretation

A substitution?




```
<subst>
  <del>half-</del>
  <add>all</add>
</subst> blind
```

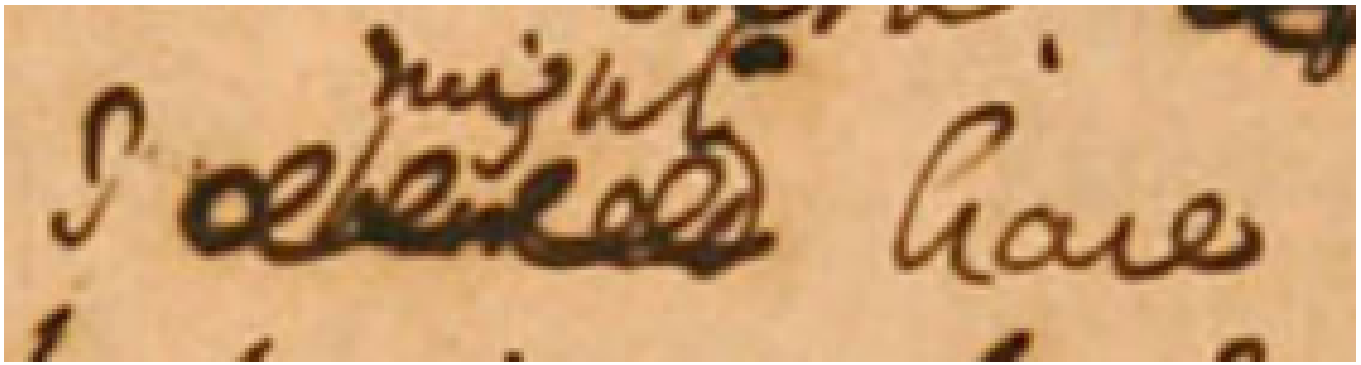
More context for Wilfred Owen




```
<l>And towards our distant rest began to trudge,</l>
<l>
  <subst>
    <del>Helping the worst amongst us</del>
    <add>Dragging the worst
         amongt us</add>
  </subst>, who'd no boots
</l>
<l>But limped on, blood-shod. All went lame; <subst>
    <del  status="shortEnd">half-</del>
    <add>all</add>
  </subst>
blind;</l>
<l>Drunk with fatigue ; deaf even to the hoots</l>
<l>Of tired, outstripped <del>fif</del> five-nines that dropped
behind.</l>
```

Semi-legible text

Use <unclear> if the text is partly illegible i.e. it can be read but without perfect confidence. The *@reason* attribute here states the cause of the uncertainty in transcription.



```
I <subst>
  <add  place="above">might</add>
  <del>
    <unclear  reason="overinking"
      cert="medium"  resp="#LDB"> should</unclear>
  </del>
</subst>have
```

Supplied and damaged text

Use the <supplied> element if the transcriber has provided a reading not actually visible in the text, whether because of damage or scribal error : *@reason* here indicates why the text has been supplied.

```
…Dragging the worst
among<supplied reason="authorialError">s</supplied>t us…
```

Use the <damage> element to record the existence of physical damage to the document, whether or not the damaged text is readable :

```
<l>The Moving Finger wri<damage agent="water" group="1">es; and</damage> having
writ,</l>
<l>Moves <damage agent="water" group="1">
   <supplied>on: nor
       all your</supplied>
 </damage> Piety nor Wit</l>
```

Lacunae

When missing text cannot be confidently supplied or is intentionally omitted the <gap> element should be used with a *@reason* to explain why. It can use its *@extent* and *@unit* attributes indicate its size.

```
<gap reason="wormhole" extent="7"
 unit="mm"/>
```

```
I am dr Sr yr <gap reason="illegible" quantity="3"
 unit="word"/>Sydney Smith
```

Some difficulties

These methods are perfectly adequate where variation is comparatively simple. They rapidly encounter problems when:

- overlap happens (as it always does)
- the sequence of interventions is important or indeterminate
- the layout and the meaning of the writing are not easily separable

Additions and deletions crossing element boundaries

When additions and deletions are not conveniently well-nested within other parts of the structure, we can use spanning techniques.

The elements <addSpan> and <delSpan> delimit a span of text by pointing mechanisms rather than by enclosing it.

@spanTo indicates the end of a span initiated by the element bearing this attribute.

```
<l> .... </l>
<addSpan spanTo="#id4"/>
<l>
<!-- an interpolated line -->
</l>
<anchor xml:id="id4"/>
<l> .... </l>
```

Using attributes to clarify who did what when

- The author (WJ) wrote ‘One must have lived... ’
- The author added the word ‘But’ before ‘One’
- An editor (FB) corrected ‘One’ to ‘one’

```
<add place="supra" hand="#WJ"
  cert="medium"> But</add>
<choice>
  <sic>One</sic>
  <corr resp="#FB" cert="high">one</corr>
</choice> must have lived ...
<!-- elsewhere -->
<respStmt xml:id="FB">
  <resp>editorial changes</resp>
  <name>Fredson
     Bowers</name>
</respStmt>
<respStmt xml:id="WJ">
  <resp>authorial changes</resp>
  <name>William
     James</name>
</respStmt>
```

Using <restore> to indicate authorial change of mind

- The author writes ‘For I hate this my body’
- The word ‘my’ is deleted
- The author writes ‘stet’ in the margin

The <restore> element can be used to indicate that a deletion has been reversed:

```
<l>[...] For I
hate this <restore hand="#dhl"
   type="marginalStetNote">
   <del>my</del>
  </restore> body [...]</l>
```

... note that we have not encoded the <metamark> ‘stet’, but rather its effect.

Text Omitted from or Supplied in the Transcription

- <gap> indicates a point where material has been omitted in a transcription, whether for editorial reasons described in the TEI header, as part of sampling practice, or because the material is illegible or inaudible.
- <supplied> signifies text supplied by the transcriber or editor for any reason, typically because the original cannot be read because of physical damage or loss to the original.

<gap> and <supplied> examples

```
expansion <gap reason="illegible" agent="water"/> river denominated
```

```
expansion <supplied reason="illegible"
  source="#SH1862">of the</supplied>river
denominated
```

<gap> Example

```
<div>
  <head>Lectio x.</head>
  <p> Hic itaque paterfamilias ad excolendam <gap  extent="20"  unit="words"
     reason="not transcribed"  resp="#DC"/>
     congregare non desistit. </p>
</div>
```

More <supplied>

Where the transcriber considers that one or more words have been erroneously omitted in the original source and corrects this omission, the <supplied> element should be used in preference to <corr>.

```
 by the ancient Dutch
navigators <supplied>of</supplied> the Tappan Zee
```

<supplied> Example

```
<p>Oblatus est <supplied  reason="omitted"  resp="#DC"> quia ipse
    voluit</supplied>. </p>
```

<damage>, <space>, and <unclear> Example

```
 Revelabunt caeli
iniquitatem Judae et <damage agent="rubbing"/> consurget et <space/>
manifestum erit peccatum ipsius in die furoris
do<unclear agent="rubbing" resp="#JC">mini</unclear> cum eis qui dixerunt
domino deo recede a nobis scientiam viarum tuarum nolumus
```

Damage and Illegibility

Use <damage> if the text can be read with perfect confidence

```
<p>
  <pb  n="5r"/>
  <damageSpan  agent="rubbing"
    extent="whole leaf"  spanTo="#damageEnd"/>
</p>
<p> .... </p>
<p> .... <pb  n="5v"  xml:id="damageEnd"/>
</p>
```

Disjoint Damage

```
 IN the bosom <damage  group="1">o</damage>f one of those spa<lb  n="2"/>cious
coves
wh<damage  group="1">ich inde</damage>nt the eastern <lb  n="3"/>shore
of the <damage  group="1">Hudson, at </damage>that broad <lb  n="4"/>expansion
<damage  group="1">of the r</damage>iver denominated <lb  n="5"/>by the
ancie<damage>nt</damage> Dutch navigators
```

Original layout information

The TEI privileges the logical view, but does permit the physical view to ‘show through’ as empty milestone elements :

- <gb> the start of a new gathering or quire
- <pb> the start of a new page
- <cb> the start of a new column
- <lb> the start of a new written line

These are primarily useful to establish a reference system.

The <fw> element can be used to mark ‘paratextual’ features such as running heads, foliotation etc.

The <handShift> element can be used to mark changes of hand or writing in a document.

<fw>

<fw> (forme work) contains a running head (e.g. a header, footer), catchword, or similar material appearing on the current page.

```
<fw place="top-centre" type="head">Poëms.</fw>
<fw place="top-right" type="pageno">29</fw>
```

Changes of hand: <handShift>

A special kind of milestone the <handShift> can be used to mark the beginning of a sequence of text written in a new hand, or the beginning of a scribal stint.

```
<l>When wolde the cat dwelle in his ynne</l>
<pb  n="f.23v"/>
<handShift  medium="greenish-ink"
  new="#h1"/>
<l>And if the cattes skynne be slyk <handShift  medium="black-ink"
    new="#h2"/> and gaye</l>
```

```
<handNotes>
  <handNote  xml:id="h1"
    script="copperplate">Carefully written with
     regular descenders</handNote>
  <handNote  xml:id="h2"  medium="pencil">Unschooled scrawl</handNote>
</handNotes>
```

<handNote> and <handShift>

The <handNote> element is used to provide information about each hand distinguished within the encoded document.

- When the 'transcr' module is used, the element <handNotes> is available, within the <profileDesc> element of the Header, to hold one or more <handNote> elements. (brief)
- When the 'msdescription' module is included, the <handDesc> element also becomes available as part of a structured manuscript description. (more robust)

It is possible to use the two elements together if, for example, the <handDesc> element contains a single summary describing all the hands discursively, while the <handNotes> element gives specific details of each.

<handShift> Example

```
<handShift new="#h1" resp="#das"/>... and that good Order Decency and
regular worship may be once more introduced and Established in this Parish
according to the Rules and Ceremonies of the Church of England and as under
a good Consciencious and sober Curate there would and ought to be
<handShift new="#h2" resp="#das"/> and for that purpose the parishioners pray
```

Editorial phrase-level elements

A summary list of some of the more important phrase-level transcription elements might include:

- Core module: <abbr>, <add>, <choice>, <corr>, <del>, <expan>, <gap>, <orig>, <reg>, <sic>, <unclear>
- 'transcr' module: <am>, <damage>, <ex>, <metamark>, <mod>, <redo>, <restore>, <retrace>, <space>, <subst>, <supplied>, <surplus>, <transpose>, <undo>

<sourceDoc>

- <sourceDoc> contains a transcription or other representation of a single source document potentially forming part of a dossier génétique or collection of sources.
- An embedded transcription is one in which words and other written traces are encoded as subcomponents of elements representing the physical surfaces carrying them rather than independently of them.
- A <sourceDoc> usually contains one or more <surface> elements with <zone> or <line> elements.
- <line> is a specialisation of <zone> that contains the transcription of a topographic line in the source document
- Some editorial markup (such as <add> <del>) are available in <line>, but it is not meant to be used for interpretative judgements like <persName>

<zone> in <sourceDoc>

Provides a embedded non-interpretative transcription

```
<zone  facs="#postcard-back_zone4"
  rend="printed">
  <line>
    <choice>
      <abbr>
        <g  ref="#RN_logo">RN</g>
      </abbr>
      <expan>Reinthal &
             Newman</expan>
    </choice>
  </line>
</zone>
<zone  facs="#postcard-back_zone5"
  rend="printed">
  <line>THIS SPACE MAY BE
     USED FOR</line>
  <line>CORRESPONDENCE</line>
</zone>
<zone  facs="#postcard-back_zone6"
  rend="printed">
  <line>PRINTED IN
     AMERICA</line>
</zone>
```

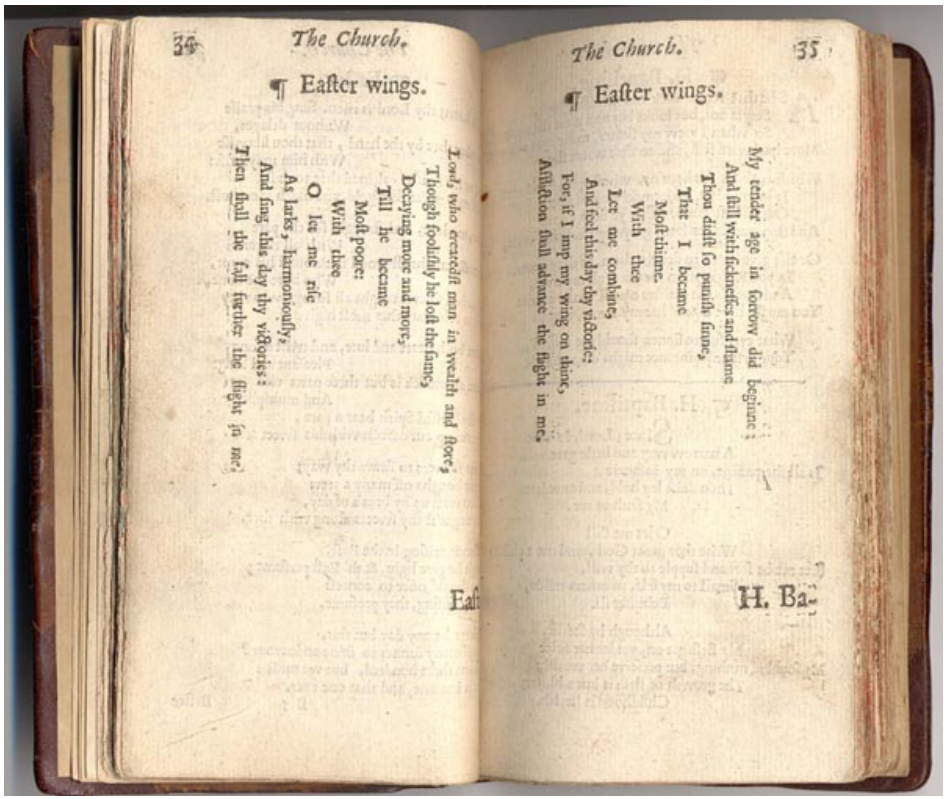
How far will the TEI take us ?

In particular, is the TEI scheme adequate for the needs of those transcribing ‘modern’ manuscripts ?

- surviving medieval or early modern manuscripts generally have a public function, and a more or less conventionalised (if complex) format
- modern manuscripts or authorial drafts however often contain entirely private or idiosyncratic signs, with no clear communicative function

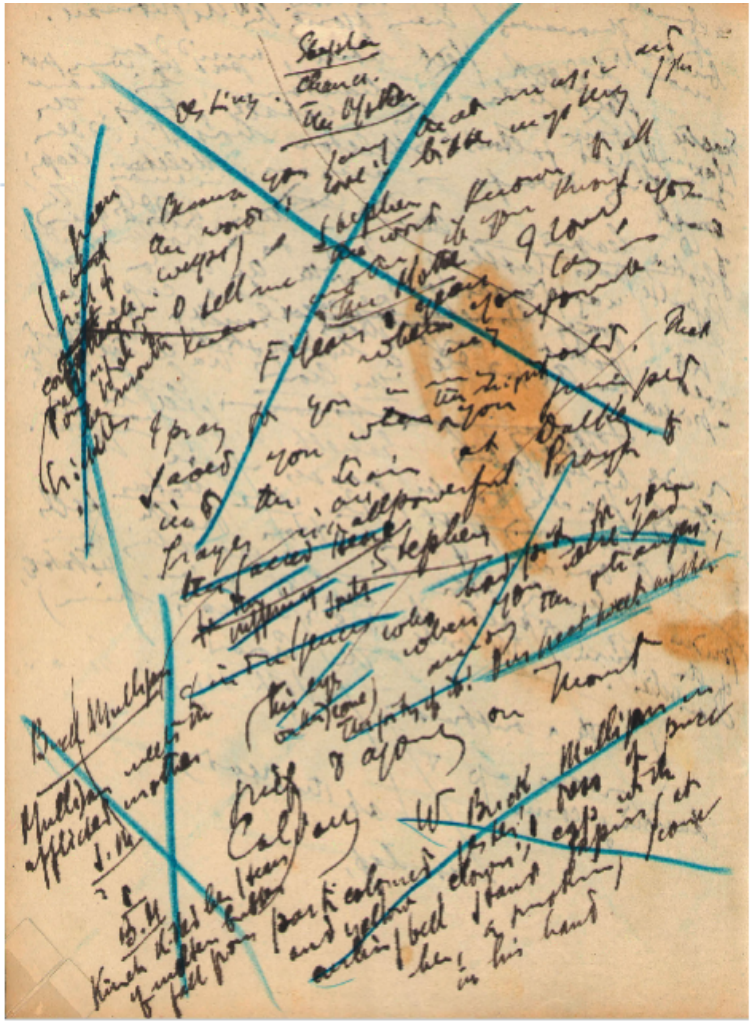
Text/Image

At all periods we find ‘playful’ texts whose meaning is conveyed by their documentary appearance as much as by their linguistic properties, or by the interplay between the two.

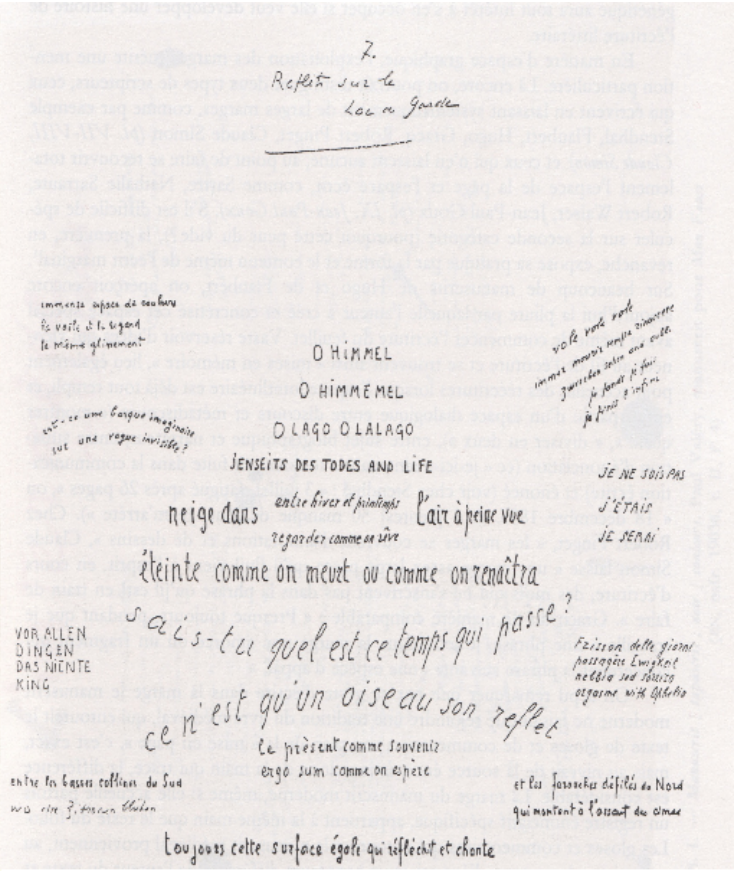


The TEI initially ruled such texts out of scope.

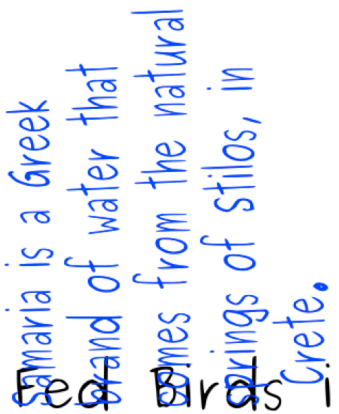
Difficult documents



More examples ...



Diary example



1 April 2009

Red Birds in the park today.
Might write an article about
the Thick-billed Warbler.

Next

Any Questions? Next we're going to do an exercise!