TEI Encoding and Syntactic Tagging of an Old French text

Dominique Estival & Nick Nicholas
Department of Linguistics & Applied Linguistics
University of Melbourne
Parkville, Victoria 3052
AUSTRALIA
{d.estival,n.nicholas}@linguistics.unimelb.edu.au

13 October 1998

1 Introduction

In this paper, we report on two aspects of SGML tagging which have emerged from a research project undertaken at the University of Melbourne on diachronic syntax. The first concerns the practical issues involved with tagging an Old French text for typographical formatting in the TEI Lite framework. The procedure we describe here represents a case study in such typographical formatting, as carried out under the particular constraints of the facilities and resources available to us (micro-computers, HTML-based processing). We hope that the experiences and decisions reported here may be of use to researchers undertaking similar tasks, for instance scholars who want to make texts available electronically.

The second issue concerns the adaptation to the TEI framework and to Old French of a syntactic tagset which had been developed for a different tagging framework and a different language, namely the Penn-Helsinki coding scheme which was created for Middle English (Taylor & Kroch 1994). Syntactic tagging, and particularly diachronic syntactic tagging, is a process still in its infancy; but the Penn-Helsinki coding scheme has already emerged as a standard which corpus syntacticians cannot ignore. Adapting the coding scheme to Old French has entailed certain expansions of the tagset which should prove of interest to corpus syntacticians working outside English, and may be seen as a contribution towards the ideal of a language-independent syntactic tagset. The adaptation of this tagset to TEI also provides a practical illustration of the issues involved in applying SGML to syntactic tagging tasks, a process which has not yet been fully explored. This project should thus be relevant to researchers in corpus-based studies as well as to researchers in diachronic syntax.

2 Brief description of the project

A fundamental goal of the research done in the dominant syntactic framework of *Principles & Parameters* (P&P) (which remains in place in Minimalist Theory: Chomsky 1995) is to account for and explain the "logical problem of language acquisition" (see Hornstein & Lightfoot 1981), in terms of parameters of Universal Grammar which a child sets while learning a language. Such parameters might be for example the direction of case assignment, whether the language is head-first or head-final, or whether it allows empty subjects.

A number of attempts have been made at implementing Principle-Based parsers, (i.e. parsers implementing a P&P approach, for instance Abney 1989, Stabler 1992, Fong 1991, Lin 1994) which researchers can use to investigate whether these implementations shed any light on on the process of First Language Acquisition.

Our goal parallels this, but takes as a starting point another fundamental observation about language: all languages change. There has been increasing research in the past two decades in diachronic syntax and in the study of syntactic change (e.g. Li 1977, Lightfoot 1979). While several researchers have been studying the consequences of particular aspects of syntactic theories for a theory of syntactic change (e.g. Harris & Campbell 1995, Lightfoot 1991), no attempt has yet been made to map out the implications of those theoretical choices for the processing aspects of syntactic change.

More specifically: it may be clear that, in a P&P approach, syntactic change is predicted to be a possible (but not necessary) consequence of a switch in the setting of a parameter, but it is not yet evident:

- how such a switch takes place in the first place (cf. Kroch 1989, Keller 1994);
- what kind of switches lead to impossible grammars (cf. Optimality Syntax, e.g. Grimshaw 1997);
- how parameters and their switches are to be implemented;
- whether the implementation tells us anything about possible (or impossible) changes;
- what the relations (if any) are between parameters and parsing mechanisms.

The goal of our research project, entitled *Computational Modelling of Syntactic Change*, is to elucidate these questions; the main hypothesis underlying this research is that the human parsing mechanism (the *HSPM* of Crocker 1996) must be robust enough to allow processing of diverging grammars while still being constraining enough to reject unacceptable input. Ultimately, we intend to model a flexible human processor capable of dealing with language change. This goal is still far away; in the short term we are collecting historical texts, and are tagging them syntactically.

As part of this project, we have obtained from Tony Kroch at the University of Pennsylvania access to the syntactically tagged Penn-Helsinki Parsed Corpus of Middle English (`http://www.ling.upenn.edu/mideng/`). This corpus uses the syntactic tagging system developed at the University of Pennsylvania, applied to the Middle English portion of the Helsinki Corpus of Historical English (Kytö & Rissanen 1988). We have also begun a collaboration with Barbara Vance at the University of Indiana, who works on Old French syntax, and are currently encoding and tagging a text obtained from her, having adapted the Penn-Helsinki coding schema (Taylor & Kroch 1994) for that purpose.[1]

[1] Syntactically-tagged corpora are still a rarity in computational linguistics, and they are of inestimable value to syntacticians trying to describe the statistical and distributional characteristics of constructions in found text. So the development of a syntactically-tagged Old French corpus is of interest to a wide range of researchers.

3 SGML encoding of Joinville

3.1 Description of the text

The text we obtained from Vance is Jehan de Joinville's *La Vie de Saint Louis,* a biography of King Louis IX of France, written around 1305. This text is written in a mixture of standard (Ile-de-France) and Champenois French, and is extremely important for historical and literary, as well as for linguistic reasons.[2]

Historically, Joinville's text is the major source of information on Louis IX, one of the preeminent rulers of medieval Western Europe. From a literary point of view, not only is this one of the first major prose texts in French (prose had started being used only a century earlier, in such works as Geoffroy de Villehardouin's *La Conqueste de Constantinople*), but it also marks the first incidence in French of literature written from a personal vantage point. It is an at times chatty and anecdotal biography, written by a close personal friend of Louis IX.

Linguistically, Joinville's text is situated at the cross-over point between Old and Middle French; it is conventionally regarded in French linguistic histories (e.g. Pope (1952 [1934])) as the last of the Old French texts. As a result, it is transitional in many respects. For instance, the Old French declension system, which separates Old French from Middle French, is already severely curtailed in Joinville. Because this is a text written in a period of morphological and syntactic transition, we hope that Joinville's text will afford us insights about linguistic change and the flexibility of a human parser in the face of such change.

The text of *La Vie de Saint Louis* survives in three manuscripts: A (Brussels manuscript), B (Reims manuscript), and L (Lucques manuscript). Manuscript A has long been regarded as the most authoritative source of the work and is the version used by Corbett (1977), the latest critical edition of the work.[3] Corbett's edition is the text which was given to us by Vance (in Macintosh Microsoft Word 5 format) after it had been scanned and proofread under her supervision. One of our major concerns in our tagging of the text was to distinguish Corbett's editorial interventions from the manuscript text: the typographical conventions of Corbett's text, including his use of italics and paragraph and line breaks, are clearly noted as such, and distinguished from the information given about the original text, such as the page breaks in manuscript A (see section 3.2).

As Corbett wished to minimize the amount of guesswork involved in his critical edition, his edition stays fairly close to A. When he emends the text in A with reference to the BL text tradition, the emendations are given in square brackets in his text. This distinction between A and BL readings has been retained in our formatted text, with use of the `<add>` and `<corr>` tags.[4] The

[2] *La Vie de Saint Louis* is available in English translation as Shaw (1963) and online at the University of Virginia E-Text Project as:
`http://etext.lib.virginia.edu/etcbin/toccernew?id=WedLord&tag=public&images=images/modeng&data=/texts/english/modeng/parsed&part=0.`

[3] The best-known editions of the work remain those by Wailly in 1867, 1868, and 1874.

[4] The `resp` attribute was given as `'CORB'`, indexing the editorial decision as made by Corbett.

reading in manuscript A which Corbett rejects in emendation is given through the `sic` attribute.

Corbett's critical apparatus appears as endnotes, rather than footnotes. The text of the endnotes was not included in the document we received from Vance, but space has been allocated for these endnotes in our document, through the `<note>` element. Since we also have access to Corbett's edition in printed form, and the syntactic tagging is proceeding on the basis of the critical edition rather than the variant manuscript readings, we have not attempted to incorporate the critical apparatus into the text at this stage; however, this can readily be done at a later date.[5]

3.2 TEI encoding

The TEI tagging of the text began in October 1996 by converting the Microsoft Word text into HTML, taking advantage of the fact that HTML is a subset of SGML. Work proceeded in fifteen distinct stages, requiring occasional small computer programs, written in C, to automate formatting changes. The major stages involved included the following:

- Converting Corbett's typographical indication of sections (using a section character (§) and a marginal section number) into numbered TEI `<div1>` elements. The section numbering follows Wailly's 1874 edition of the text, which is taken as the standard reference.
- Resolving the treatment of emendations in Corbett's text, as already discussed, and allocating space for Corbett's footnotes (critical apparatus).
- Interpreting typographical italics in Corbett's text in terms of the various TEI highlight tags (`<foreign>`, `<mentioned>`, `<q>`, etc.). Where italics are used for extensive quotation (as with the ordinance cited in sections 694–714), the use of italics is indicated through the `rend` attribute. This task was performed manually (the NLP resources to automatically resolve the semantics of italicization would be formidable indeed!); however italicization was infrequent enough in the text for this not to be too onerous.
- Consistently treating the various manifestations of quotation in Corbett's text. In accordance with Corbett's typographical conventions, three renderings of quotations are noted: *italics*, *em-dash*, and *continue* (where a quotation spans over a section boundary, and the `<q>` tag needs to be resumed in the new section). Where no rendering type is indicated in the tagging, it is understood that the original text uses quotation marks.

[5] The linguistic reliability of critical editions of medieval works remains a controversial issue, as the critical text is frequently prepared on the basis of philological, literary and historical rather than strictly linguistic criteria. As a result, the linguistic realism of the critical text is subject to distortion. It has been suggested that computational tools may in fact break the stemmatological deadlock and make it easier to deal with manuscripts witnesses directly (Robinson 1996), undermining the notion of a single standard text.

- Adding indication of Corbett's line and page breaks (using the `<pb>` and `<lb>` tags), and the page and paragraph breaks from manuscript A. Paragraph breaks are rendered as `<milestone>` elements, with `unit` attribute `'para'`.

  In Corbett's text, the page breaks in manuscript A are indicated with a pipe character (|) and a marginal page number. In our tagging, both Corbett's and manuscript A's page breaks are treated uniformly as `<pb>` elements; they are differentiated through the `ed` attribute — set to CORB and JOIN, respectively. Treating a manuscript of a work as an edition is not an unreasonable extension of the `ed` attribute.

  Note that the `<p>` tag could not be used to indicate the paragraph breaks in manuscript A, which are also the typographical paragraph breaks employed in Corbett's edition. This is because Corbett's paragraphs overlap with Wailly's sections, and these are rendered as `<div>` elements, whereas the `<p>` element is a child of a `<div>` element in TEI. Furthermore, Corbett's paragraphs are not numbered, while Wailly's sections are, and serve as text references for anyone working with this text. As Wailly's sections have to take priority, they need to be tagged as `<div>` elements; thus the coding of paragraphs, rather than that of sections, needed to be adjusted for the encoding.[6]

- Deleting line-break hyphenations, and distinguishing line-break from orthographic (or morphological) hyphenation — the latter, being part of the orthographic representation of the Old French word, was retained. This was a painstaking task, since there is no typographical differentiation between the two, and knowledge of Old French morphology was required to make a proper distinction. Initially we removed all hyphens preceding line-breaks, intending to reinsert those hyphens which were actually morphological. However, we subsequently decided to reinstate these hyphens, and had them checked. All instances of actual line-break hyphenation were then removed manually, and only occurrences of actual Old French orthographic hyphens were retained. Indications of line and page breaks frequently occur mid-word in both Corbett's text and in the pagination of manuscript A. These indications have been moved to the end of the orthographic words they were found in.

- Tagging and numbering each orthographic sentence in the text as a `<seg>` element of type `sentence`. This was done for consistency with the Penn-Helsinki (PH) syntactic tagging scheme: all syntactic tagging of the text takes place within the bounds of these sentential `<seg>` elements. Also

[6] This is an instance of the overlapping hierarchies problem, which is well-known amongst SGML users. TEI offers another solution to this problem, using the `part` attribute, whereby a tagged section can be marked as incomplete, constituting either the beginning, middle, or end of a greater whole. However the paragraph division of the original manuscript did not seem important enough to warrant imposing two overlapping segmentations onto the text (one of them discontinuous), thus impeding the ease of processing and the clarity of the text structure; the `<milestone>` solution has proven to be quite workable.

following PH conventions, sentence number rather than line number is used to refer to a specific location in the text.

- Compilation of the `<teiHeader>` information. This information includes a description of the editorial and rendering conventions used in tagging the text, as well as one or two interventions necessary in the text (e.g., assuming in one case that for one emendation in Corbett's text the ensuing ghost-footnote mark is in fact a typographical error for the missing right square bracket).

3.3 Differences with Helsinki encoding

One of our goals is to build a corpus of historical texts, encoded and tagged in similar ways to facilitate comparison between them, and to make these texts available to historians and literary analysts as well as to linguists and philologists. We thus decided not to adopt the text mark-up system employed for the Helsinki corpus, even though it is one of the most prominent diachronic text corpora, because it does not adhere to a currently commonly accepted scheme, and because it is quite minimalist in the type of formatting it allows. The goal of the creators of the Helsinki corpus (which predates TEI and SGML, having started in the early part of the '80s) was exclusively linguistic, so that much of the formatting information which other scholars might find of interest was simply not relevant to their aims.

We illustrate this point by presenting a comparison of the way the Joinville text would have been formatted in the Helsinki corpus scheme in (1), the way it would appear in the Penn-Helsinki scheme in (2), and the way we have tagged it in TEI in (3).

(1) Helsinki

```
 Lors il me dist:  " Donc devez-vous croire fermement touz les articles
de la foy, lesquiex les apostres [{ nous {] tesmoingnent, aussi comme
vous oez chanter au dymanche en la (\ Credo \) [^ sample comment ^] .
"

<R 46> Il me dist que l'evesque Guillaume de Paris li avoit cont&eacute;
que un grant mestre de divinit&eacute; estoit venu a li et li avoit dit
que il vouloit parler a li.  Et il li dist:  " Mestre, dites vostre
 volent&eacute;.  " Et quant le mestre cuidoit parler a l'evesque, et
commen&ccedil;a a plorer trop fort.  Et l'evesque li dit:  " Maistre,
dites, ne vous desconfort&eacute;s pas, car nulz ne peut [^ Joinville
p.  24 ^] tant pechier que Dieu ne peut plus pardonner.  "
```

Unlike in the Penn-Helsinki (PH) tagging system, there is no sentence numbering in the Helsinki encoding, only page numbering. On the other hand, the Helsinki tagset retains formatting information which is discarded in PH, for example, (\... \) for italics. By contrast, the PH way of encoding texts, as shown in (2), does not preserve formatting information (e.g., the information about the string *Credo* in PH is it is a Latin word, not that it is in italics, as in the Helsinki encoding), but includes sentence as well as page numbering.

(2) Penn-Helsinki

(Lors il me dist: " Donc devez-vous croire fermement touz les articles de la foy, lesquiex les apostres < nous > tesmoingnent, aussi comme vous oez chanter au dymanche en la <latin> Credo </latin> <hc> sample comment </hc> . ") (JOINV 45.6)

(Il me dist que l'evesque Guillaume de Paris li avoit conté que un grant mestre de divinité estoit venu a li et li avoit dit que il vouloit parler a li.) (JOINV 46.1)

(Et il li dist: " Mestre, dites vostre volent'e. ") (JOINV 46.2)

(Et quant le mestre cuidoit parler a l'evesque, et commença a plorer trop fort.) (JOINV 46.3)

(Et l'evesque li dit: " Maistre, dites, ne vous desconfortés pas, car nulz ne peut <hc> Joinville p. 24 </hc> tant pechier que Dieu ne peut plus pardonner. ") (JOINV 46.4)

Our rendition of this passage with TEI formatting keeps the typographical information specified in the PH format — e.g. page breaks, line breaks, and so on.[7] In addition, it contains not only sentence and page numbering, but also section numbering, as well as editorial information.

(3) TEI

```
 <seg type='sentence' n='JOIN45.6'>Lors il me <lb ed=CORB> dist:
  <q>Donc devez-vous croire fermement touz les articles <lb ed=CORB>
 de la foy, lesquiex les apostres <add resp=CORB>nous</add>
 tesmoingnent, aussi <lb ed=CORB> comme vous oez chanter au dymanche
 en la <term>Credo</term> <note resp=editor> sample comment</note>.
  </q> <lb ed=CORB></seg> </div1>

<div1 id=JOIN46 n='46' type='section'><P><milestone unit=para ed=CORB>
 <seg type='sentence' n='JOIN46.1'>Il me dist que l'evesque Guillaume
 de Paris li avoit cont&eacute; <lb ed=CORB> que un grant
 mestre de divinit&eacute; estoit venu a li et li avoit
 <lb ed=CORB> dit que il vouloit parler a li.</seg>
 <seg type='sentence' n='JOIN46.2'>Et il li dist:  <q>Mestre, dites
 <lb ed=CORB> vostre volent&eacute;.</q></seg>
 <seg type='sentence' n='JOIN46.3'>Et quant le mestre cuidoit parler
 a l'evesque, <lb ed=CORB> et commen&ccedil;a a plorer trop
 fort.</seg><seg type='sentence' n='JOIN46.4'>Et l'evesque li dit:
```

[7] We dispense with the information given by the rendering of specially formatted words, replacing it with explicit semantic information given in the tag label; since all specially formatted words are actually in italics in the source, it would be redundant to supply this information. The type of rendering could also be noted as e.g. <foreign rend=italic>, should the need arise (e.g. if we embark on a more thorough-going tagging of terms, including items not italicized in the source.)

```
<lb ed=CORB> <q>Maistre, dites, ne vous desconfort&eacute;s
pas, car nulz ne peut <lb ed=CORB><pb ed=JOIN n='24'> tant pechier
que Dieu ne peut plus pardonner</q></div1>
```

Note in particular that the TEI tagged text includes information on line breaks, the nature of quotations (e.g. that *Credo* is a technical term[8]), the editorial responsibility for emendation (as well as, potentially, the original text) and explicit identification of which version of the text the page break is associated with.

Because of the flexibility of TEI Lite, its level of standardization, and its economical yet powerful inventory of tags, we decided to employ it as our tagging system, thereby guaranteeing for other researchers accessibility and easy processing of the text to be produced. The Joinville text has now been fully tagged for typographical formatting in TEI Lite.

This work, including proofreading of the entire text for such matters as hyphenation and italics, continued from mid-October 1996 until the start of February 1997.[9]

Work now under way involves tagging the text for syntactic information. Although this syntactic tagging is taking place within the framework of TEI (rather than PH) text-markup, we are using a version of the PH syntactic tags, which we have modified for Old French. Since in the rest of the project we will use the software developed at the University of Pennsylvania for use with the PH-tagged corpus, we are also in the process of writing a translator between our TEI-based and the PH syntactic tagging schemes.

4 Syntactic coding

As mentioned earlier, our research program involves the computational analysis of texts as facilitated by the addition of syntactic tags. Syntactic tagging is still a young discipline, particularly in diachronic linguistics, and the major venture of this kind to date is the Penn-Helsinki Corpus. While this corpus has established a standard in tagged diachronic corpora in linguistics, there are several problems with the scheme as it stands. In particular, the PH tags were devised specifically for Old and Middle English. Although the syntactic ontology they delimit is not in itself anglo-centric, several of the distinctions nonetheless need to be modified before the coding scheme can be applied to Late Old French.

We give a brief description of the PH tagging scheme and show how the tags have been translated into SGML in section 4.1, and then present in section 4.2 the modifications to the tagset introduced for Old French.

[8] If this had been considered an instance of code-switching, we could have used the `<foreign lang=Latin>` tag. However, the use of Latin terms in Old French is better viewed as technical terminology than as code-switching, and we have chosen to tag them accordingly.

[9] The text is currently 750 K in size (from an original HTML file of 490 K). Its SGML was grammar-checked using the Macintosh implementation of the YASP parser developed by David Durand.

4.1 Translation of Penn-Helsinki (PH) tags into SGML

The PH tagset conveys six kinds of information:

1. The syntactic role of the constituent is represented by a single character after the opening square bracket surrounding the constituent. For example, direct object is denoted as `[d ]`.

2. The clausal status of the constituent is conveyed by the case of the character. Thus, nominal direct objects are denoted as `[d ]`, while clausal direct objects (taken to be instances of *THAT-clauses*) are denoted by `[T ]`.

3. Whether a clause is main or subordinate is conveyed by the choice of square or round brackets. Thus, embedded indirect questions, being subordinate clauses, are marked as `[Q ]`, while direct questions, which count as main clauses, are marked as `(Q )`.

4. The depth of embedding of subordinate clauses is indicated by numerals around the beginning and end brackets. The numbering is to be interpreted breadth-first; e.g., a subordinate relative clause embedded within the second subordinate clause in a sentence is labelled as `2.1[L ]2.1`.

5. Empty categories inserted into the text for syntactic analysis are indicated by percent signs instead of square brackets around the constituent. For example, in (4), the elided copula in the second of the conjoined clauses is represented as `%vt%` (i.e., an empty transitive verb):

   ```
   (4) ( [+ but ] [t si+t ] [vt was ] [s +tis tare ] [m schewyd ] )
       (WYCSER,I,374.416)
       ( [+ and ] [s Godis lawe ] %vt% [m hid ] , ) (WYCSER,I,374.417)
   ```

6. Coindexing can be indicated by adding `-n` (where `n` is a numeral) to the tag code. For instance, presentational *there* is coindexed to the logical subject of the sentence, as shown in (5):

   ```
   (5) [s-1 there ] [vt were ] [n-1 many kyngis ]
   ```

   In the PH corpus, the main use of this coindexing device is to coindex phrases with empty categories.

The excerpt from the Middle English text *The Ancrene Riwle* given in (6) provides a brief example of a text tagged in the PH tagset:[10]

```
(6) ( [+ For ] [s uh <P I.44> an ] [at schal ] [v halde ] [d +tuttere]
    [p efter 1[B [c +tet ] [s heo ] [b best ] [at mei ] [p wi+d hire ]
    [vseruin ] [d +teo inre ] .  B]1]
    {For each one shall hold the outer according that she best may serve
    with it the inner} )(ANCRIW,I.44.25)
```

[10] The digraphs `+t` and `+d` denote "thorn" and "edh".

We decided to encode the entire Joinville document, including its syntactic tagging, in TEI, and then to translate the TEI text to the PH tagset in order to manipulate the text using the software designed at the University of Pennsylvania. In this manner, the formatting information not included in the PH scheme is not lost, but is preserved for use outside the PH framework by other researchers.

The issue then arises of how to render syntactic tags in TEI.[11] TEI has three tags which could be used to that end: `<s>`, `<seg>`, and `<interp>`. Of these, `<s>` is used to delimit orthographic sentences only, and does not allow for embedding of other `<s>` within it; the only attribute it allows is the number of the current sentence. So `<s>` is obviously not appropriate as the general solution to the syntactic tagging problem. It is appropriate for sentence numbering, which is how PH requires the text to be organised; however, since the orthographic sentence is itself considered a syntactic unit in PH, it is preferable to handle all constituents uniformly by means of the same tag.[12]

Both of the two remaining tags (`<seg>` and `<interp>`) allow embedding and can take on arbitrary type attribute values. Of the two, `<seg>` additionally requires that its constituents be properly nested (no branch-crossing), and that its type attribute have a single coded value. The `<interp>` tag does not have these restrictions: it allows an arbitrarily complex hierarchy of interpretation types to be set up in one part of the document, and referred to anaphorically from another. This is why `<interp>` is explicitly suggested in Burnard & Sperberg-McQueen (1995) as a means of encoding syntactic information; they give the following as an illustration:

```
(7) <interp id=NP1 type=pos value='noun phrase, singular'>
    <interp id=VV1 type=pos value='inflected verb, present-tense singular'>
    ...
     <seg type=sentence ana=SVO>
    <seg type=lex ana=NP1>John</seg>
    <seg type=lex ana=VVI>loves</seg>
    <seg type=lex ana=NP1>Nancy</seg>
    </seg>
```

However, the actual encoding of constituents becomes unnecessarily unwieldy: since `<interp>` itself is an empty element, constituents would still have to be delimited using something akin to `<seg>`, with the **ana** attribute connect-

[11] We do not use feature structures to render syntactic tags, as is done in the Corpus Encoding Standard (CES: Ide, Priest-Dorman & Véronis 1996), for two reasons. First, the CES guidelines only became available after this project was already started, and more importantly, we want our coding scheme to be compatible with the PH framework. In addition, the fine-grained distinctions made in the CES would be much more difficult to establish consistently than the rather simple PH scheme. Of course, it should be possible to establish a correspondance between our tagset and a subset of the CES tagset.

[12] We can always easily reinstate the `<s>`, since orthographic sentences are clearly distinguished from other constituents by attribute.

ing the <interp> element to this delimiter.[13] Furthermore, the strict nesting required by <seg> is not only consistent with standard syntactic constituency, but also useful in error-checking the tagging.

We have thus elected to use the <seg> element to encode syntactic tagging. The information on the meaning of the various tags is not encoded anaphorically within the document, although this can easily be done if it proves necessary. The ontology of information encoded in the PH tagset (see next section) is simple enough that it can be rendered as attributes of the <seg> element, as follows:

- To convey the distinction between main clauses (round brackets), subordinate clauses (square brackets, uppercase tag), and clause-internal constituents (square brackets, lowercase tag), the `function` attribute is used, and is set to `'m'`, `'s'`, and `'c'`, respectively.

- The PH tags for constituent type are given in the `type` attribute. For instance, the equivalent of the PH direct object tag, `[d ]`, is as in (8):

  ```
  (8)     <seg function='c' type='d'>
  ```

- We could notate coindexing in TEI through the <ana> element. However, since the processing of such syntactic coindexing is to take place within the PH framework, it is sufficient to manually insert the coindexing number into the `type` attribute, thus retaining consistency with the PH scheme. For example, a coindexed presentational *there* is tagged in our system as in (9):

  ```
  (9)     <seg function='c' type='s-1'>
  ```

- Empty types can be either clausal or intra-clausal, so the information that a constituent is empty cannot be conveyed through the `type` attribute which gives information about constituent type. Instead, we use the `subtype` attribute, set to `%`. Thus, an elided copula is tagged as in (10):

  ```
  (10)    <seg function='c' type='vt' subtype='%'>
  ```

- Since the indexing of subordinate clause embedding results from an entirely mechanical traversal of well-nested clausal constituents, this information is not included in our syntactic tagging: it is to be calculated and inserted by the TEI-PH translator.

Thus we see that the ontology of syntactic tags required is just simple enough to be encompassed within the attributes of the <seg> element. Any additional complexity would require the more complex hierarchies encodable by the <interp> element, although <seg> elements would still be required, and placed around the constituents to index the <interp> definitions.

[13] The <interp> element does not have to be connected to <seg>-delimited text segments, so <interp> is not susceptible to the overlapping hierarchies problem like <seg> is. However, in most syntactic models well-defined nestedness is desirable, and we do not see intrasentential typographical tagging conflicting with syntactic tagging significantly.

4.2 Adaptation of the PH coding scheme to Old French

Our guiding principle throughout the modification of the PH tagset has been not to modify the interpretations of the existing PH tags, thereby ensuring that future comparisons between different tagged texts will remain meaningful. Thus, we have limited ourselves to making additions to the existing tagset where necessary. Some of these additions are motivated by differences between Middle English and Old French syntax. Other additions concern distinctions which are possible in Middle English syntax but were not considered important enough to record in the PH corpus, and which are necessary in Old French (e.g. the use of different prepositions with infinitive complements, or the coindexing of discontinuous constituents).

4.2.1 New Tags

We list here a few examples of the tags which had to be created for Old French.

Impersonal subjects Impersonal subjects are not specifically coded for in the PH corpus, but as they are of interest in French, we use the new tag `[si]` (for "sujet impersonel"). In presentational constructions, we use coindexing between the impersonal subject and the post-verbal NP, as in (11):

```
(11)   il y a X =    [si-1 il] y a [n-1 X]
       there is X
```

Infinitive constructions Diachronically, infinitive constructions are of special interest in French, so we have devised a set of new tags for infinitive constructions to complement the simple infinite construction marker `[U ]` in PH:

- `[U ]` : verbal complement, as in (12):

```
(12)   commencer [U [p a] [v travailler]]
       to begin to work
```

- `[UN ]` : infinitive + negation, as in (13):

```
(13)   [s il] [vt est] [parti] [UN [p sans] [v payer]]
        he left without paying
```

- `[UP ]` : purpose, as in (14), (15), and (16):

```
(14)   [UP [p pour] [v travailler]]
       in order to work

(15)   [UP [p pour] [d le peuple de la terre] [v garantir]]
       in order for the people of the earth to guarantee
       (JOIN, 12.4)

(16)   [UP [Et pour moy [v aquitier] [p de ma promesse]]
       and for me to make good on my promise
       (JOIN, 18.2)
```

Note that *peuple* in (15) and *moy* in (16) are the deep subjects marked syntactically as direct objects in the infinitive construction.[14]

Reflexive pronouns Reflexive pronouns are not specifically coded for in the PH corpus, but their development is particularly interesting in Romance languages. They are tagged as `[r ]`. This tag is used irrespective of whether the pronoun is semantically a real reflexive as in (17.a), an inherent reflexive as in (17.b) or (17.c), or a reciprocal as in (17.d).

```
(17) a   [s il] [r se] [vt leva]
         he got up
         (JOIN, 41.4)
     b   [r s'] [p en] [vt feust] [b bien] [m soufers]
         would have suffered
         (JOIN, 20.2)
     c   a cui [s il] [r s'] [vt attendoit]
         which he was expecting
         (JOIN, 16.1)
     d   et [vi devisiens] [r l'un a l'autre]
         and we were talking to each other
         (JOIN, 31.2)
```

Vocatives The Joinville text contains many vocatives. As vocatives are left uncoded in the PH corpus, we have created a new tag `[!]` for this function, as in (18.a), (18.b) and (18.c):

```
(18) a   [! toi]
         you!
     b   [! toi, Joinville]
         you, Joinville!
     c   [! Joinville]
         Joinville!
```

Prepositional phrases

- Usually PH leaves prepositional phrases undifferentiated; however in some of the texts in that corpus, a distinction is made between different types of PP, e.g. `[pt ]` (temporal), `[pl ]` (locative) etc. We have decided to systematically mark the function of the PP when it is unambiguously clear, as in (19) for manner and (20) for reason.

```
  (19)   [vt ama] [d Dieu] [pm de tout son cuer]
         loved God with all his heart
         (JOIN, 20.1)
```

[14] Cf. the Latin *Accusativum cum infinitivo* construction: *dixit canem carnem edere* (he said the dog eats meat) with both *canem* (dog) and *carnem* (meat) in the accusative case.

(20) [pr pour le soutil senz] dont vous estes
for the subtle mind which you have
(JOIN, 26.1)

Otherwise, as in PH, a PP is simply tagged `[p]`.[15]

Example (21) is an instance of a sequence of PPs, some of which can be disambiguated:

(21) [p ou pelerinage de la croiz], [p par l'espace de sis anz]
on the pilgrimage of the cross, in the space of six years
[w que] [s je] [vt fu] [p en sa compaignie].
that I was in his company.
(JOIN, 5.1)

- A PP which is a subcategorized argument of a verb, an adjective or a noun (we leave open here the issue of how to determine that a given PP is an argument or an adjunct), is marked `[pa ]`, as in (22) and (23):

(22) [vt parle] [pa de ses granz chevaleries]
relates his great actions
(JOIN, 2.3)

(23) [pa de bon cendal] [m enforcier]
with good silk reinforced
(JOIN, 25.4)

- A predicative PP is marked `[pd ]`, as in (24):

(24) [s elle] [vt sera] [pd de refus]
it will be out of commission
(JOIN, 15.1)

- In order to help record the transition from the inflectional case system of Old French, we tag constituents which function as PPs but do not contain a preposition, with the `[p0 ]` tag. This is only done when there is evidence that the phrase is functionally equivalent to a full PP, and in our text this only occurs in frozen archaic expressions. See (25), which contains a preposition before *moy*, but no preposition before *Dieu*:

(25) Si vous pri-je [p pour l'amour] [p0 Dieu premier]
so I beg you for the love of God
et [p pour l'amour] [p de moy]
and for the love of me
(JOIN, 29,4)

[15] As the clitic adverb *y* stands for a locative PP, is is always tagged `[pl y]`, while the clitic *en* is ambiguous between a locative and a partitive interpretation and remains unresolved, as `[p en]`, as in (i):

(i) [s les freres] [r s'] [p en] [vt furent] [m partis]
the brothers had left
(JOIN, 27.3)

If other researchers do not accept the analysis we propose here, it would be very easy to treat the <[p0]> tag as an instance of <[n]>, i.e., a noun phrase.

We extend this analysis and notation to instances of possessive constructions without overt possessive marking, as in (26), overt possessive marking having become obligatory in Modern French. Note that this analysis deviates from PH, where the whole NP would be tagged as <[n]>.

(26) un livre des saintes paroles et des bons faiz
[p0 nostre saint roy Looys]
a book of the saint sayings and of the good deeds
[ ] our saint king Louis
(JOIN, 2.1)

4.2.2 Extension of existing tags or notation

Negation [-] is the negation tag in PH. As verbal negation in French is often discontinuous, we have decided to tag separately each part of the negation, as in (27.a) and (27.b). Note that the second negative element in (27.b) is tagged as an adverb.

(27) a [- ne] [- ne] scé comment
I don't know how
(JOIN, 98.3)

b Et [- ne] voult [t onques] demourer avec le roy,
and he never wanted to remain with the king
(JOIN, 56.2)

Coindexing

- We extend the coindexing mechanism to cover cases of discontinuous constituents, as in (28):[16]

  (28) et [c-1 pource] [b meismement] [c-1 que] il...
  and because in the same way he...
  (JOIN, 5.2)

- When a relative clause unambiguously modifies the last conjunct of a coordinated structure, we break apart the coordinated structure and mark as coreferential the relative pronoun and its antecedent. Compare (29.a) and (29.b): in such cases, PH does not break apart the coordinated structure and would leave both structures as tagged in (29.a), although in (29.b) the relative clause unambiguously modifies the second conjunct only.

  (29) a et je vi [d la femme et l'enfant], 1[L [w qui] [j seulz]
  and I saw the woman and the child, who alone(PL)
  [pl de ceste nef] [vt eschaperent] L].

[16] The discontinous negation particles of examples (27) could also be coindexed, if desired.

```
            of this ship escaped(PL)
            (JOIN, 14.3)

    (29) b  [d la femme] [+ et] [d-1 l'enfant], 1[L [w-1 qui] [j seul]
            the woman and the child, who alone(SG)
            [pl de ceste nef] [vt eschapa] L]
            of this ship escaped(SG)
```

Causative constructions Causative constructions are analyzed as a succession of verbs, with the causee NP marked as a direct object, as in (30):

```
(30)    [s il] [d leur] [vt fait] [v donner] [p pour Dieu]
        he made them give for God (to the Church)
        (JOIN, 32.2)
```

We give in (31) an example of a complete sentence tagged with the TEI syntactic tagset:

```
(31)  <seg function='m' type='sentence' n='JOIN2.2'>(<seg function='c'
      type='s'> La premi&egrave;re partie </seg> <seg function='c'
      type='r'> si</seg> <seg function='c' type='vt'>devise</seg>
      <seg function='w' type='b'> <seg function='c' type='w'>comment</seg>
      <seg function='c' type='s'>il </seg> <seg function='c' type='r'>se</seg>
      <seg function='c' type='vt'> gouverna</seg> <seg function='c' type='t'>
      tout<lb ed=CORB> son tens</seg> <seg function='c' type='p'> selonc Dieu
      et selonc l'Eglise</seg>, <seg function='c' type='+'>et</seg>
      <seg function='c' type='g'> au profit de son <lb ed=CORB> regne</seg>
      </seg>.)  <lb ed=CORB></seg>
```

La première partie si devise comment il se gouverna tout son tens selonc Dieu et selonc l'Eglise, et au profit de son regne.
The first part explains how he governed himself during all his time according to God and to the Church, and to the benefit of his kingdom.

5 Tools for tagging

The entire document of the Joinville text, including its syntactic tagging, has been encoded in TEI, and then translated to the PH tagset, in order to manipulate the text using the software designed at the University of Pennsylvania.

Both the typographical and syntactic tagging of the text have been undertaken entirely on microcomputers. For formatting, TEI tags were automatically entered either through translation of the Microsoft Word original document to HTML, or through small programs automating further tag translation. This task was completed relatively quickly.

Syntactic tagging, on the other hand, is a long and tedious process, which cannot be substantially automated. So it is desirable to make the task easier through a graphical visual interface. By analogy with other graphical interfaces, such an interface involves highlighting text sections to be tagged, and specifying the appropriate tag either by pressing a button on screen or by keying in a short command. Owing to the limited availability to date of SGML editors on microcomputers, the syntactic coding in this project has been carried out on HTML editors we have customized for this task, namely *PageSpinner* on the Macintosh, and *HTMLedPro* on PC.

Similar functionalities had to be provided for both platforms. For example, in *PageSpinner*, a phrase is tagged as a direct object by highlighting it with the mouse and pressing `control-D`, or by selecting **direct object** from the *Custom Tags* menu. In *HTMLedPro*, the equivalent task is performed either by pressing `alt-D`, by selecting **direct object** from the *Custom Tags* menu, or by pressing the button on screen marked `d`.

A challenge in this project has been finding a microcomputer HTML editor allowing a sufficiently large number of custom tags, in order to deal with the Penn-Helsinki tagset. The two editors we chose were, to our knowledge, the only non-commercial instances for their respective platforms allowing such customisation to our satisfaction. Other editors either did not allow any customisation of the tagset, severely restricted the number of custom tags (for example, to six or ten, whereas the number required for our task is closer to forty), or did not allow the custom tags to bear complex structure (for instance, attributes).

6 Conclusion

6.1 What can be learned?

Converting the Microsoft Word document we received into TEI was relatively straightforward conceptually; the prevalence of HTML formatting tools, and RTF-to-HTML converters in particular, was of great help. Similarly, preparing the software needed to automate significant portions of formatting, such as the numbering of sections, sentences, and pages, did not prove a major challenge. However, the process was structurally complex, consisting of many discrete stages.

In order to modularize the task, and to make sure that particular editorial choices could be undone, it was necessary to keep these stages separate, and to produce and save a new document version for each major formatting task. This approach proved crucial when, for example, the decision to remove all line-final hyphens and manually reinsert morphological hyphens was rescinded, in favour of manually removing all line-final hyphens. By that stage formatting had proceeded along several other stages; but because an earlier version of the document with hyphenation in place was intact, it was possible to write code to automatically restore line-final hyphens as appropriate.

An aspect of our project which can be seen as characteristic of computing in the humanities has been its reliance on microcomputers (although a Unix workstation has more recently been obtained for use in the project). The kind of formatting and encoding tasks involved in such a project need to be carried out using locally available computational resources and, particularly in humanities departments (and especially outside the United States), workstation-level computers, including high-end microcomputers, are not always readily available. Furthermore, even if they are available, they are not necessarily accessible by the research assistants who are most likely to be doing the tagging. The wide availability and increasing computational power of microcomputers makes them well-suited to the task, but since SGML has been primarily associated with UNIX-driven machines, several issues arise.

First, although SGML parsers are available for both Macintosh and PC microcomputers, the choice of software available is not as wide as for workstations. This paucity is even more acutely felt for SGML browsers.[17]

Second, the computational power of microcomputers is still limited compared to workstations — particularly the microcomputers typically available in a humanities department. In parsing the formatting-tagged Joinville text, for example, the resources of the computer we were using were severely taxed, even after the capacity variables in the SGML prologue were increased to the maximum allowed by the parser. [18] This was in fact one of the deciding factors in some of our formatting choices.

For instance, explicit coindexing within SGML using `<interp>` and `<ana>` alongside `<seg>` in syntactic tagging, instead of the attributes of `<seg>` alone, would have caused a massive increase in the number of variables the parser would have had to deal with, with no perceptible gain towards the aims of our project. As it stands, parsing the 750K SGML document (with no syntactic tagging) on a Macintosh LC475 computer running TCL YASP takes about four minutes. A microcomputer is thus clearly inadequate for the type of syntactic work envisaged, and we intend to move up to a workstation for that phase of research.

The relative lack of SGML resources for microcomputers is even more acutely felt in syntactic tagging. The formatting-tagging part of the project is based on regular expression substitution, and can make use of small programs modifying the HTML source. By contrast, syntactic tagging cannot be automated, and necessitates a proper SGML editor. Budget constraints meant it was impractical to purchase commercial microcomputer SGML software; so we have customised HTML editors on both Macintosh and PC.

The fact that only one available non-commercial editor on each platform could cope with our customisation demands indicates the problems researchers can run into when they cannot work within the commonly assumed world of workstation-based SGML.

[17] Fortunately, this discrepancy is being addressed, although less quickly for the Macintosh than for the PC.

[18] For example, TOTALCAP is the number of entities the SGML parser may recognise; the maximum TOTALCAP value allowed by our microcomputer parser was a million, although the value the parser actually reported for the text was closer to one and a half million.

6.2 Foreseen Results and Future Tasks

In the long run, we hope that this project will lead to a better understanding of two universal characteristics of language, language variation and language change. Besides the obvious interest from the point of view of historical linguistics, an understanding of the limits and types of variation in language is also necessary to deal adequately with actual language. Ultimately the integration of language variability in the modelling of Natural Language Processing should lead to models that are closer to the behaviour of humans in real language situations.

However, this is still a distant goal, and for the immediate future, this research can be seen as a contribution to the study of historical linguistics in several ways. A first result is what we have already produced: an electronic version of an important text in a standardized encoding, which can be used by other researchers studying language change. Secondly, we have devised a set of syntactic tags which builds upon the PH coding scheme and applies to Late Old French, while still compatible with the original tagset. Researchers in language change need more electronic texts and, especially for the study of diachronic syntax, syntactically tagged texts. We hope that by the completion of this project, we can provide a syntactically tagged text of Joinville's *La Vie de Saint Louis*, which will thus be our third contribution.

We intend to extend our study further to texts from the Penn-Helsinki Middle English collection, and continue the collaboration with Vance. We also hope that other researchers who have been informally pooling and exchanging texts will, like Vance and Kroch, give us access to their texts with the prospect of getting in return a TEI-encoded and syntactically tagged version of those texts. This would lead to a more general extension of the PH tagset to other types of texts and to other languages.

We conclude by noting that, in addition to encoded and tagged texts, researchers in diachronic syntax need computational tools to be available so they can actually use those texts effectively.

7 Acknowledgements

We wish to thank the University of Melbourne for the two grants (a Special Interest Grant and a Small Australian Research Council Grant) awarded to the first author, which have allowed us to undertake this project; Tony Kroch and Barbara Vance for giving us access to their data; and Katia Margolis for her help with the syntactic coding of Old French. Thanks also to the three anonymous reviewers who have helped us clarify some of the issues raised in this paper.

8 References

[1] Abney, Steven (1989). 'A Computational Model of Human Parsing.' *Journal of Psycholinguistic Research*, 18.1.

[2] Burnard, Lou & Sperberg-McQueen, C. M. (1995) *TEI Lite: An Introduction to Text Encoding for Interchange.* Document No: TEI U 5. `http://www.uic.edu/orgs/tei/intros/teiu5.tei`.

[3] Chomsky, Noam (1981). 'Principles and Parameters in Syntactic Theory.' in *Explanations in Linguistics. The logical problem of language acquisition*, Norbert Hornstein & David Lightfoot, eds. London: Longman. ch.2, pp.32-75.

[4] Chomsky, Noam (1995). *The Minimalist Program.* Cambridge: MIT Press.

[5] Corbett, Noel L. (1977). *La Vie de Saint Louis: Le témoignage de Jehan, seigneur de Joinville.* Sherbrooke, Québec: Éditions Naaman.

[6] Crocker, Matthew W. (1996). *Computational Psycholinguistics. An Interdisciplinary Approach to the Study of Language.* Dordrecht: Kluwer Academic Publishers.

[7] Fong, Sandiway (1991). *Computational properties of Principle-Based Grammatical Theories.* PhD Thesis, MIT, Cambridge.

[8] Grimshaw, Jane (1997). 'Projections, Heads and Optimality.' *Linguistic Inquiry. vol.28. n.7.*

[9] Harris, Alice & Lyle Campbell (1995). *Historical syntax in cross-linguistic perspective. Cambridge Studies in Linguistics 74.* Cambridge: Cambridge University Press.

[10] Hornstein, Norbert & David Lightfoot, eds. (1981). *Explanations in Linguistics. The logical problem of language acquisition.* London: Longman.

[11] Ide, Nancy, Greg Priest-Dorman & Jean Véronis (1996). *Corpus Encoding Standard.* Document No: CES 1. Version 1.4. `http://www.cs.vassar.edu/CES`.

[12] Keller, Rudi (1994). *On Language Change: The Invisible Hand in Language.* London: Routledge.

[13] Kroch, Anthony S. (1989). 'Function and Grammar in the history of English'. In *Language Change and Variation.* Ralph Fasold & Deborah Schiffrin, eds. *Current Issues in Linguistic Theory, vol. 52.* Amsterdam: John Benjamins.

[14] Kytö, Merja & Matti Rissanen (1988). 'The Helsinki Corpus of English Texts: Classifying and Coding the Diachronic Part'. In *Corpus Linguistics, Hard and Soft: Proceedings of the Eighth International Conference on English Language Research on Computerized Corpora.* Merja Kytö, Ossi Ihalainen & Matti Rissanen, eds. *Language and Computers: Studies in Practical Linguistics No. 2.* Amsterdam: Rodopi.

[15] Li, Charles ed. (1977). *Mechanisms of Syntactic Change.* Austin: University of Texas Press.

[16] Lightfoot, David (1979). *Principles of Diachronic Syntax.* Cambridge: Cambridge University Press.

[17] Lightfoot, David (1981). 'Explaining syntactic change.' In *Explanations in Linguistics. The logical problem of language acquisition.* Norbert Hornstein & David Lightfoot, eds. London: Longman.

[18] Lightfoot, David (1991). *How to set parameters: arguments from language change.* Cambridge (Mass.): MIT Press.

[19] Lin, Dekang (1994). 'PRINCIPAR — An Efficient, Broad-coverage, Principle-based Parser.' *Proceedings of COLING94*. pp.482–488.

[20] Pope, M. K. (1952 [1934]). *From Latin to Modern French.* Manchester: Manchester University Press.

[21] Robinson, Peter M.W. (1996). ' Computer-Assisted Stemmatic Analysis and 'Best-Text' Historical Editing '. In *Studies in Stemmatology.* Pieter van Reenen & Margot van Mulken, eds. Amsterdam: John Benjamins. pp.71–103.

[22] Shaw, Margaret R. B. (1963). *Joinville and Villehardouin: Chronicles of the Crusades.* London: Penguin Books.

[23] Stabler, Edward (1992). *The Logical Approach to Syntax: Foundations, Specifications, and Implementations of Theories of Government and Binding.* Cambridge (Mass.): ACL-MIT Press Series in Natural Language Processing.

[24] Taylor, Ann & Anthony Kroch (1994). *The Penn-Helsinki Parsed Corpus of Middle English.* MS. University of Pennsylvania.