French Language Studies **18** (2008), 103–119, © Cambridge University Press
doi:10.1017/S0959269507003183

Annotating an oral corpus using the Text Encoding Initiative. Methodology, problems, solutions[1]

JANICE CARRUTHERS

Queen's University Belfast

(Received August 2007; revised September 2007)

ABSTRACT

The objective of this paper is to describe and evaluate the application of the Text Encoding Initiative (TEI) Guidelines to a corpus of oral French, this being the first corpus of oral French where the TEI has been used. The paper explains the purpose of the corpus, both in creating a specialist corpus of *néo-contage* that will broaden the range of oral corpora available, and, more importantly, in creating a dataset to explore a variety of oral French that has a particularly interesting status in terms of factors such as *conception orale/écrite, réalisation médiale* and *comportement communicatif* (Koch and Oesterreicher 2001). The linguistic phenomena to be encoded are both stylistic (speech and thought presentation) and syntactic (negation, detachment, inversion), and all represent areas where previous research has highlighted the significance of factors such as medium, register and discourse type, as well as a host of linguistic factors (syntactic, phonetic, lexical). After a discussion of how a tagset can be designed and applied within the TEI to encode speech and thought presentation, negation, detachment and inversion, the final section of the paper evaluates the benefits and possible drawbacks of the methodology offered by the TEI when applied to a syntactic and stylistic markup of an oral corpus.

0. INTRODUCTION

This project involves the collection, transcription and encoding of a corpus of new storytelling (or *néo-contage*), a type of oral narration that has enjoyed extensive revitalisation in France since the 1980s. As such, it is a specialist corpus which, it is hoped, will not only add to the range of oral corpora available to researchers, but will also allow an analysis of a variety of oral French that poses particularly interesting questions with regard to a number of syntactic and stylistic phenomena.[2] It is the

[1] I am very grateful to the AHRC (Research Leave Scheme) and the British Academy (Small Grant award) for sponsoring the corpus project on which this paper is based; to the Oxford Text Archive for their technical expertise, in particular, James Cummings; to the two anonymous JFLS readers; and to Jacques Durand, Julien Eychenne and Nabil Hathout.

[2] For excellent lists of French and francophone corpora, see Beeching (2002, Appendix) and Cappeau and Seijido on www.culture.gouv.fr/culture/dglf/recherche/

https://doi.org/10.1017/S0959269507003183 Published online by Cambridge University Press

first corpus of oral French to use the Text Encoding Initiative (TEI) methodology and the purpose of this paper is to use the project to show how the TEI can be deployed to encode syntactic and stylistic phenomena in the oral medium, and to offer a critical evaluation.

1. RATIONALE FOR THE CORPUS

Unlike their traditional counterparts, new storytellers tend not to have acquired their stories as part of a living oral tradition passed from one generation to the next, but rather have normally worked, at least initially, with written texts from a variety of registers, historical periods and cultures. They are performed in a variety of performance contexts to eclectic audiences who may be more or less familiar with the storyline. The relationship between the storyteller's text and his/her oral performance is extremely complex and involves concepts which go well beyond a binary oral/written divide. The precise process through which what Lyons calls 'medium transferability' (1981, 11) occurs is not monolithic but, with the notable exception of stories in verse form and those where the language is closely allied to a musical accompaniment, it does not involve simple memorisation of a text.[3]

Koch and Oesterreicher (2001), building on the work of Söll (1985), offer an analysis that attempts to disentangle concepts that are frequently conflated, i.e., questions of medium, register, discourse type etc. They posit three distinct but related notions: **conception orale/écrite** (with 'conception orale' at one end of a spectrum and 'conception écrite' at the other, giving the examples of a spontaneous conversation for the former and a legal text for the latter); **réalisation médiale** (a binary divide between oral and written codes); **comportement communicatif** (a multi-factored continuum relating to communicative distance vs immediacy). New storytelling, or *néo-contage*, presents a complex and unstable pattern with respect to these three notions. The simplest part of the analysis relates to the *réalisation médiale* which clearly is transferred from the written code (the source) to the oral code (the performance). However, if we look at the *conception orale/écrite* continuum, both the written source and the final oral performance would be placed somewhere in the middle, with literary textual versions of *contes* more towards the *conception écrite* end than, for example, the most informal of children's versions, and different performance factors and individual stylistic preferences determining the precise position on the scale of a given oral performance.[4] The positioning of

corpus_parole/Presentation_Inventaire.pdf. The majority of existing corpora of oral French involve informal conversational data, but there are also a number of media corpora. There has been much discussion recently of the need to broaden the range of 'les oraux' under investigation (Bilger and Cappeau 2004).

[3] This is confirmed by a questionnaire submitted to participating storytellers who speak of memorising the storyline events and certain formulae but never the text itself.

[4] Training offered to storytellers by centres such as the CLIO would strongly suggest that the desired position for a successful oral performance is towards the *conception écrite* end of the scale (Carruthers 2005: 58–60). Cf. Montelle (1994) and De la Salle (2004).

https://doi.org/10.1017/S0959269507003183 Published online by Cambridge University Press



Figure 1.

néo-contage with regard to *comportement communicatif* is particularly interesting, with different factors suggesting a mixed and variable analysis. Figure 1 gives the relevant parameters (Koch and Oesterreicher 2001: 586):

In the case of *néo-contage*, three factors could be considered invariable, i.e. 1, where we are always dealing with a *communication publique* in the context of performance; 6, where there is always *coprésence spatio-temporelle*; and 10, where there is a high degree of *fixation thématique*. The rest are variable to greater or lesser extents: e.g. the audience may be more or less well known to the *conteur*, emotionality is normally but not always high, a story is normally but not always a monologue (sometimes there is audience participation), and the extent of the spontaneity varies. In theory factors 4 and 5 suggest distance, but with oral narrative, there is often attempt on the part of the storyteller to present the events as happening before the eyes of the audience, thus at least attempting to create immediacy. The source texts can also present a mixed pattern, but there are clearly ways in which they differ dramatically from the oral performance, notably with regard to factors such as 3, 6, 7, 8 and 9.

The purpose of the current corpus is twofold. First, given that there are no French corpora of new storytelling in the oral medium, one primary aim is to establish a corpus with full metadata for use by a variety of scholars not just in linguistics but also in fields such as literary studies or anthropology. The corpus contains 1000 minutes of live storytelling from a variety of storytellers, all of whom are native French speakers. The aim was not to create a stratified corpus as such; the primary parameters to be explored are not speaker variables such as age, gender, socio-economic status or regional origin.[5] Rather, the corpus is intended to be a specialised one (see Leech 2005: 12–13) that is broadly representative of the type of discourse that is contemporary oral storytelling. The most important goal was to obtain a spread of category of story across the range (*contes facétieux, contes d'animaux, contes d'enseignement, contes fantastiques, légendes, contes bibliques, contes de la création, mythologies* and *contes merveilleux*), partly in order to give a representative

[5] All such details were nonetheless recorded as important metadata in the Header.

https://doi.org/10.1017/S0959269507003183 Published online by Cambridge University Press

sense of the types of stories told by contemporary *néo-conteurs*, and partly to create the opportunity to test whether there are links between type of story and certain linguistic features. In order to avoid major inbuilt imbalances that might skew the analysis, the corpus aims to reflect broadly the gender balance (there are 19 participating storytellers in all, including 10 women and 9 men),[6] and a spread of ages and regional backgrounds within France. No material was included where there was a risk of simple memorisation of a written text: verse forms and recordings where the story was closely allied to music were avoided. All stories were recorded live by the *Conservatoire de littérature orale* and thus form part of their collection.[7] The performance context was also noted: the majority of the performances take place in a relatively intimate format in front of a fairly small audience (known as the *Perce Oreille* workshop), but some involve a much larger and more public format, as they are part of the CLIO's *Rencontres d'eté* or the *Rencontres d'hiver* which are frequently held outside in open-air contexts, or in large venues such as churches.

Secondly, as far as the linguistic analysis is concerned, oral storytelling is fertile ground for research on a number of linguistic phenomena where, alongside syntactic, stylistic, phonetic and semantic factors, questions of *conception orale/écrite, réalisation médiale* and *comportement communicatif* have been shown strongly to influence usage. A recent project on a smaller corpus revealed tense usage, patterns of subordination/ coordination/ parataxis at points of temporal juncture, and speech and thought presentation to be three areas where *néo-contage* raises interesting questions (Carruthers 2005). This project focuses on three different syntactic phenomena, i.e. negation (or non-use of *ne*), detachment and inversion, as well as expanding previous work on speech and thought presentation. Speech and thought presentation is the most complex of these areas to encode, and one where recent and current research has begun to explore not only literary and journalistic texts but also various types of oral material (Marnette 2005). Patterns evident in *néo-contage* appear on an initial exploration to be predominantly those traditionally associated with the *conception orale* end of the spectrum and with communicative immediacy, (e.g. large quantities of direct discourse), but there is certainly evidence of features associated with *conception écrite* and communicative distance (e.g. small quantities of free indirect discourse: Carruthers and Marnette 2007). In the case of inversion, it is widely argued that it is rare in the oral medium and increasingly restricted not only to largely to the written medium, but also to the *conception écrite* end of the spectrum, notably to literary texts and some types of journalism, where its usage is further restricted in terms of syntactic context (see Ayres-Bennett and Carruthers 2001: chapter 8; Clifford 1973; Lahousse 2003, 2006; Postal 1989). In the case of negation, variationist research suggests a strong link between the non-use of *ne* on the one hand, and both the oral medium and *conception orale* on the other, in addition to a number of significant linguistic factors such as the subject of the verb (lexical subject versus clitic, use of particular pronouns), the *forclusif*, the

[6] In terms of minutes, male speakers form 44% of the corpus, female speakers 56%.

[7] I am extremely grateful to the CLIO's artistic director, Bruno de La Salle, and the staff, for access to the recordings, and to all the storytellers who gave permission for their work to be used.

https://doi.org/10.1017/S0959269507003183 Published online by Cambridge University Press

phonetic context (e.g. pre-vocalic versus pre-consonantal contexts), use in main versus subordinate clauses, use in 'frequently used phrases' (e.g. Armstrong 2001; Armstrong and Smith 2002; Ashby 2001; Coveney 1996, 1998). Detachment is thought to be almost exclusively oral, with left-detachment in particular associated primarily with communicative immediacy (notably with spontaneous discourse); there is also extensive published research on the characteristics of left- and right-detached structures in syntactic, semantic and pragmatic terms (e.g. Ashby 1988, 1982; Barnes 1985; Blasco-Dulbecco 1999).

2. THE TEXT ENCODING INITIATIVE: KEY PRINCIPLES

No one corpus methodology has dominated the scene in terms of encoding and analysing data, in part because each project has its own specific demands and in part because the technology is evolving so rapidly. However, there is substantial consensus on good practice in annotation technique (see Leech 2005), as well as a high degree of common terminology for part-of-speech tagging.[8] The Text Encoding Initiative (TEI) was designed as a form of extensible markup, suitable for encoding any type of text with a wide variety of types of markup and across different languages.[9] The system is extensible in that the markup can be expanded, refined and generally adapted to suit different purposes: it offers a tagset from which a project can utilise a subset of the elements and to which the researcher can add new categories within certain elements in order to encode the level of detail required and the key factors.[10] XML mark up is used in the TEI to encode all types of information, including large quantities of metadata (analytic, editorial, descriptive and administrative), as well as the text itself with its specific annotations regarding the relevant textual and linguistic factors: these will all be explored in detail below.

In terms of the encoding language deployed, Leech argues that XML-based systems (and those of related languages such as SGML and HTML), of which the TEI is one, will come to dominate, arguing that 'the triumph of the more advanced SGML/HTML/XML style of encoding is in the long run assured' (Leech 2005: 23), largely because, as related standardised markup languages, they have world-wide long-term currency; the advantages and problematic issues in the particular case of the TEI will be discussed below. There are of course many other well-established tagging systems used for syntactic, prosodic and pragmatic annotation, and to a lesser extent, for phonetic and stylistic encoding, as well as language acquisition (for an excellent list, see Leech 2005).

[8] See the publications of the EAGLES initiative (Expert Advisory Group on Language Engineering Standards) which has proposed a common set of morphological tags (see Delais-Roussarie 2000; Bilger 2000). Indeed, notwithstanding certain problematic issues regarding the classification of particular phenomena, an inevitable level of error and certain difficulties that can arise with spoken data (see Thompson 2005: 68), part-of-speech (POS) tagging can often be done automatically.

[9] Full details are available at www.tei-c.org. This project uses the current version, TEIP4. A new version, TEIP5, is due to be released in November 2007.

[10] TEIP5 will be even more flexible and customisable than TEIP4. There is also greater interoperability with other systems.

https://doi.org/10.1017/S0959269507003183 Published online by Cambridge University Press

3. THE TEXT ENCODING INITIATIVE: THEORY INTO PRACTICE

3.1 Metadata: the Header

Within the TEI, all metadata is contained in the 'Header', i.e. a section of each document which appears before the text itself but in the same XML language. The Header contains four key components, i.e. the file description, the encoding description, the profile description and the revision description, within each of which the researcher selects as many sub-categories from the vast choice in the TEI Guidelines as are necessary or appropriate for the project in question.[11] So, for example, to take one of the simplest sections of metadata in this project, under '**file description**' <fileDesc>, entries were inserted under the TEI elements of 'title statement' <titleStmt> (i.e. the 'title' of the *conte* <title> and the 'author' <author> (the *conteur*)), under 'publication' <publicationStmt> (the <idno> (CD number and track in the CLIO collection) and the <authority> (i.e. the researcher)), under 'source description' <sourceDesc> (a <recordingStmt> containing details of the date and equipment used):

(1)
```
<teiHeader>
  <fileDesc>

    <titleStmt>
      <title> Le Crapaud et la Tortue </title>
      <author n= "conteur" > Elisabeth Calandry </author>
    </titleStmt>
    <publicationStmt>
      <idno> CLIO-038–4 </idno>
      <authority> Janice Carruthers </authority>
    </publicationStmt>
    <sourceDesc>
      <recordingStmt>
        <recording>
          <date value= "1998-05-05" > 05 May 1998 </date>
          <equipment>
            <p> Media: CD </p>
          </equipment>
        </recording>
      </recordingStmt>
    </sourceDesc>
```

[11] In what follows, TEI elements appear, as they do in any given file, inside <> (e.g. <particDesc>) and 'Desc' is short for 'description', 'Stmt' for 'statement', 'Decl' for 'declaration', 'Class' for 'classification'. In technical terms, these TEI elements can then have attributes (these too are listed in the TEI Guidelines), and the attributes take 'attribute values', i.e. annotations or tags in common parlance.

https://doi.org/10.1017/S0959269507003183 Published online by Cambridge University Press

The same principles follow for each component of the Header, i.e., selection of elements from the TEI guidelines and embedding of these within each section of the Header. Under '**profile description**' can be listed, for example, information on the language used in the recording under <langUsage>, information on speaker variables in their 'participant description' <particDesc> such as their sex, age, identity no. (within the project), date and place of birth, educational background, their linguistic background, socio-economic status information, current and previous areas of residence and the nature of the audience, as well as information on the setting in the <settingDesc> (activity, place, building, date, length of story), and on the type of *conte* under <textClass>.[12]

By far the most expansive and complex information in a project where there is extensive tagging is contained in the '**encoding description**' section of the Header. Under <projectDesc> there is a substantial section on the nature of the project, under <samplingDecl> a detailed account of the guiding principles used in sampling (see above Section 1), and under <editorialDesc> there is information on correction of the text, on normalisation principles adopted[13] and on the rationale for the tagging structures adopted. A detailed taxonomy of all the tags then follows under <classDesc>.[14] All information is thus expressed in an XML format, including all the metadata in the Header, as well as the textual encoding.

3.2 Metadata: the tagging system

In a project such as this where we are not dealing with POS tagging, the researcher can design the particulars of their system with almost limitless flexibility. In this project, 133 different attribute values were created for the attribute 'ana', such that a given element <u> (utterance) or <seg> (segment of discourse) could have an attribute 'ana' which contains one or more attribute values from the list explained in the <classDesc>. In practice this means that each utterance will have an 'ana' which contains an attribute value from the list of forms of speech and thought presentation, and each example of negation, inversion and detachment will be wrapped in a <seg> on which there is an 'ana' attribute which contains the appropriate references for that structure.

[12] Note that inTEIP5 the <particDesc> has been expanded considerably.

[13] Normalisation concerns, for example, the spelling conventions adopted and the question of how hesitations, truncation etc. are handled. The TEI offers a number of possible solutions. In this corpus, as a broad principle, we used standard orthography where the lexical items could be unproblematically distinguished, alongside certain conventions such as ? for a rising intonation question, - for a truncated word (e.g. *pla-, fr-*). A full-stop was used to indicate a syntactic break that is evident from the intonational structure, where otherwise the construction would be open to different interpretations. For other phenomena that can be problematic in the oral medium, such as sound effects, laughter, interruptions, unclear sections etc., TEI elements such as <vocal>, <unclear>, <event> were used (see 3.3).

[14] The fourth category, <revisionDesc>, includes details relating to later changes to the file.

https://doi.org/10.1017/S0959269507003183 Published online by Cambridge University Press

3.2.1 Negation

Given previous research on the significant factors affecting use or non-use of negative *ne* (see above), the tagging system used here allows us to encode each eligible negative (encoded 'neg') for the following attribute values:[15]

- The negative form attested or not attested (i.e. presence or absence of 'ne' or 'n'', ambiguous contexts).
- The *forclusif*: (e.g. 'pas', 'plus').
- The subject of the verb (e.g. various types of pronoun, lexical phrases, contexts where there is no surface subject).
- One or more of a group of other possible relevant factors, again based on previous research (e.g. the fact that the *forclusif* precedes the *ne*, the presence of a non-subject clitic between subject and verb, the context of a frequently-used expression).

(2) il a bondi dehors il a couru à travers le plateau il a couru jusqu'à la forêt et là <seg ana= "neg zn pas cp3si" > il a pas remis ses raquettes </seg> <seg ana= "neg zn pas cp3si" > il a pas eu le temps </seg> il était trop pressé de descendre il avait gagné son pari mais ce cette nuit-là quand il s'est couché ça a été la dernière fois qu'il se couchait parce qu'<seg ana= "neg ne jamais cp3si nsc" > il ne s'est jamais relevé </seg> au bout de six mois il était mort

The first two negatives in 2 involve cases where the form 'n'' is not attested (coded 'zn') and the *forclusif* is 'pas', while the third involves 'ne' with 'jamais'. All three are coded 'cp3si' for the subject of the verb (i.e. clitic pronoun 3rd person singular 'il'), and in the case of the third example, for the non-subject clitic, 'nsc', appearing between the subject pronoun and the verb.

3.2.2 Inversion

'inv' signals that subject-verb inversion is attested in a declarative construction, and thereafter the element preceding the inversion or triggering it is encoded, as is its position at the beginning or the end of the narrative if relevant: e.g.

(3a) <seg ana= "inv adv" > à peine avait-elle dit ces mots </seg> que l'enfant transformé en corbeau s'envola

(3b) <seg ana= "inv incin" > dit le corbeau </seg>

In 3a, the inversion is coded 'inv' and the adverbial trigger 'adv', while in 3b, the tags are 'inv' and 'incin', the latter signalling an *incise* with a nominal subject (vs 'incip' where there is a pronominal subject).

[15] Some structures were eliminated form the statistical count, as they are in other surveys of negation (e.g. negatives containing a 'ne' but where there is no *forclusif*); details of the criteria used are given in the <editorialDesc> and the structures are encoded 'nl'.

https://doi.org/10.1017/S0959269507003183 Published online by Cambridge University Press

3.2.3 Detachment

Classification of detached structures encodes them as such ('det') and for the following factors shown to be significant by previous research:

- The grammatical category of the detached element and its position to the left or right of the clause.
- The replacement pronoun within the clause.
- One or more of a variety of other possible relevant factors, such as the presence of a double detachment, cases where the pronoun is not straightforwardly coreferential with the detached element, the presence of material inserted between the detached element and the main clause.

(4a) <seg ana= "det llex pca" > une petite goutte d'eau <seg ana= "neg n rien oncp" > ça n'engage à rien </seg> </seg>

(4b) <seg ana= "det rlex p3ss" > elle avait disparu la vieille </seg>

(4c) parce que <seg ana= "det lpn p1ss insd" > moi tu vois dans deux heures il faut que j'allaite l'enfant que nous élevons </seg>

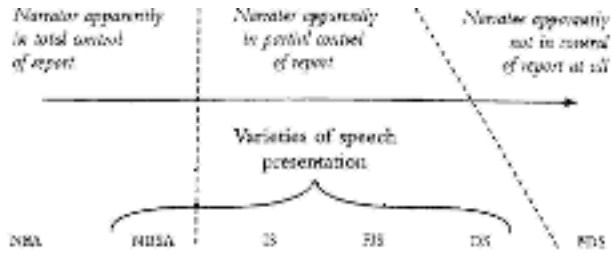
In 4a, 4b and 4c, the detached element and its position is encoded ('llex', 'rlex' for left- and right-detached lexical items and 'lpn' for a left-detached pronoun), along with the replacement pronoun ('pca' for 'ça', 'p3ss' for a 3rd person singular subject and 'p1ss' for a first person singular subject), and in 4c, the element 'insd' indicates that material is inserted between the detached element and the main clause.

3.2.4 Speech and thought presentation

The tags deployed in this project are closely allied to those used by Semino and Short (2004) in their study of speech, writing and thought presentation (henceforth SWTP) in an English-language corpus, which is in turn strongly influenced by the well-known theoretical model of speech and thought presentation proposed in the form of the continuum in Figure 2 (adapted from Leech and Short, 1981:324).

Since SWTP was the sole focus of Semino and Short's project, their tagset is more extensive, notably at the end of the spectrum that concerns the narrator. On the other hand, given the nature of oral storytelling and the interaction between *conteur* and audience, this project created attribute values such as 'NA' to indicate the narrator addressing the audience directly, and 'NF' to denote the narrator recounting elements that are part of the story's frame rather than the narrative itself. At the 'types of represented discourse' end of Leech and Short's spectrum, in addition to 'free direct discourse' ('FDD'), an attribute value of 'fDD' was added, for a particular type of free direct discourse that is found extensively in oral narration

https://doi.org/10.1017/S0959269507003183 Published online by Cambridge University Press



(NRA = narrative report of action; NRSA = narrative report of speech act; IS = indirect speech; FIS = free indirect speech; DS = direct speech; FDS = free direct speech)

Figure 2.

which in fact operates in a very similar way to direct discourse (see Carruthers 2005: 96).[16] The tagset used was thus:

- NS (narrator recounting the story)
- NA (narrator addressing audience directly)
- NF (narrator recounting frame)
- RD (a report of discourse – by either narrator or a character)
- ID (indirect discourse)
- FID (free indirect discourse)
- DD (direct discourse)
- fDD (free direct discourse operating like DD)
- FDD (free direct discourse)

and a number of tags to indicate that the interpretation of a structure is ambiguous between two or three possibilities, e.g.

- NS-FID
- ID-FID
- NS-FID-fDD

These categories are particularly important because in many instances it is the ambiguity around point of view that is of interest. For example, where there are clear indicators of FID (e.g. a deictic such as *maintenant* that is attributable to the character alongside past tenses attributable to the narrator), a structure is clearly encoded as such. However, there are instances where the narrator is clearly the *locuteur*, but it less easy to be categorical about the *énonciateur*. For example, in 5, at the point where the 'NS-FID' tag appears, there are elements that suggest that the character is the *énonciateur*, such as the exclamative 'oh', the adjective 'ravissant', the intonation and pitch used in 'ça sentait bon', but it is not always possible to be certain, especially in a context where the narrator is not a distant 3rd-person

[16] NB This project uses 'discourse' which encompasses both 'speech' and 'thought' (see the <editorialDesc> for discussion).

https://doi.org/10.1017/S0959269507003183 Published online by Cambridge University Press

narrator (where there is *séparation spatio-temporelle*, to use Koch and Oesterreicher's term (2001:586)), but one who is physically present in the discourse and frequently expresses a point of view:

(5) alors l'esprit a fait un geste de la main et tous les buissons qui entouraient la femme se sont mis à fleurir <seg ana= "NS-FID" > oh mais c'était ravissant des fleurs de toutes les couleurs et ça sentait bon </seg>

3.3 The text

Working on the premise that most transcriptions of oral language are not easily divisible into structures such as sentences, and in line with other well-known French corpora such as DELIC, a largely punctuation-free text was used, with exceptions explained in the <editorialDesc>.[17] The TEI offers a variety of possibilities for breaking the text into constituent parts. For written texts, there are more obvious breaks such as paragraphs, sentences etc., all of which have TEI elements, but with oral material, the divisions depend heavily on the nature of the analysis. Larger sections of text can be separated into divisions . Some projects separate out each clause as a separate utterance (TEI element <u> within), while others may use much longer utterances. In this case, since the differentiation between the narrator's voice and characters' voices (different *locuteurs* in *énonciation* terms) will be a key distinction for the linguistic analysis, the text is divided into <u>s that are analysed (using the attribute 'ana') as 'NS' on the one hand (in which there can be sections of 'NF' and of 'NA') and sections of 'DD' and 'fDD' on the other. For example, in 6, sections of 'NS' are interspersed with the direct discourse of characters, coded 'DD' where it is preceded or followed by a verb of speech, or 'fDD' where this is not the case:

(6) <u ana= "NS" > alors le sultan a regardé l'homme en question qui s'est mis à rougir on a vu ses mains qui commençaient à trembler </u>
<u ana= "fDD" > exprime-toi parle </u>
<u ana= "DD" > et bien </u>
<u ana= "NS RD" > <seg ana= "inv incip" > a-t-il dit </seg> </u>
<u ana= "DD" > sultan je je il y a il y a quelques années je je je vous avais invité à manger chez moi vous vous souvenez </u>
<u ana= "fDD" > oui et bien </u>
<u ana= "fDD" > et bien heu heu pendant le repas une de vos bagues a malencontreusement heu glissé sous la table et j'ai oublié de vous la rendre mais je vous en supplie sultan je vous la ramènerai demain je vous en supplie je vous ai dit la vérité </u>

Within any given <u>, there can be segments of 'RD' when discourse is reported, as well as 'ID' for segments of indirect speech, and within <u>s that are encoded as direct discourse (or 'fDD'), there can be sections of 'DD' or 'ID' embedded in the

[17] See note 13.

https://doi.org/10.1017/S0959269507003183 Published online by Cambridge University Press

direct discourse. Within <u>s encoded 'NS', there are also sections of 'FID' where the *locuteur* is the narrator and the character is the *énonciateur*.[18] Once the *conteur* (or speaking subject in *énonciation* terms) is identified by a code in the Header and by an attribute 'who' at the beginning of the first <u>, it is assumed that this remains the case unless indicated otherwise, allowing the researcher to encode instances (a small minority) where the audience becomes the speaking subject.[19]

The text is further encoded in two key ways. The first concerns elements which are all derived from the TEI Guidelines, in particular the elements recommended in the TEI chapter on spoken data covering pauses, unclear sections, vocalised non-lexical phenomena such as laughter, kinesic phenomena (e.g. gestures), non-linguistic events (e.g. background noises), and shifts or changes in vocal quality such as pitch. The TEI elements have attributes to record further details such as the reasons for unclear sections, or the nature of the change of pitch, or the precise duration of the pause etc. The extent to which a given project uses these options will depend heavily on the goals of the project. In this project, vocal phenomena such as sound effects and laughter can be very important and were thus fully marked up, whereas different lengths of pauses were not noted between clauses, except those that were exceptionally long, and as such, had a striking effect on the unfolding of the story. In example 7a, there are three unclear sections in the audience's utterances, where the coding uses the element <unclear> and gives the reason, alongside the best attempt at interpretation where possible:

(7a) <u ana= "NS" > je tourne à droite qu'est-ce qu'il y avait à droite hein? </u>
<u who= "aud" > <unclear reason= "inaudible" /> </u>
<u who= "BDS1" ana= "NS" > non </u>
<u who= "aud" > <unclear reason= "almost inaudible" > une pâtisserie </unclear> </u>
<u who= "BDS1" ana= "NS" > non </u>
<u who= "aud" > <unclear reason= "different voices" /> </u>

7b and 7c show two types of vocal phenomena, each with a description giving the nature of the vocal phenomenon:

(7b) <u ana= "DD" > oh ben attendez je m'occupe de toi <vocal desc= "chuckles" /> </u>
<u ana= "NS" > il a fermé la porte il a été voir son père </u>

(7c) <u ana= "DD" > j'ai des années d'expérience j'ai des connaissances médicales je peux invoquer bien des choses mais là pour votre fille c'est <vocal desc= "intake breath" /> mais <seg ana= "neg zne que cp1s" > je vois qu'une chose </seg> une chose pour la sauver </u>

[18] The fact that sections of FID are embedded in sections of NS poses no difficulty for searches since instances of FID are clearly encoded as such.

[19] Clearly the attribute 'who' would be very significant in a conversational corpus where there is turn-taking.

https://doi.org/10.1017/S0959269507003183 Published online by Cambridge University Press

The final layer of coding gives full stylistic and syntactic encoding, with all <u>s taking an attribute 'ana' which contains an attribute value encoding the type of speech and thought presentation (e.g. 'NS', 'DD' etc.) and all negatives, detached structures and examples of inversion wrapped in a <seg> element on which there is an 'ana' containing the relevant references for that structure. With the various layers of annotation, a fully-encoded piece of text thus appears as example 8:

(8) <u ana= "NS" > et complètement désespéré Shamsédine est rentré dans sa maison sa femme qui connaissait l'heure de son retour s'était faite elle aussi laver masser huiler parfumer épiler et elle l'attendait avec son plus gracieux sourire </u>
<u ana= "fDD" > oh soirée de félicité sur toi mon mari </u>
<u ana= "DD" > <seg ana= "neg n pas cpimpers" > il n'y a pas de félicité </seg> pour moi <seg ana= "neg n plus jamais double cpimpers insm" > il n'y aura plus jamais de félicité </seg> pour moi </u>
<u ana= "NS RD" > <seg ana= "inv incin" > a répondu Shamsédine </seg> </u>
<u ana= "DD" > mais comment donc mon époux? que manque-t-il à ton bonheur? quelle est la cause de ton chagrin? </u>
<u ana= "NS RD" > <seg ana= "inv incin" > a dit sa femme </seg> </u>
<u ana= "fDD"> la cause de mon chagrin. <seg ana= "det rlexpce" > c'est toi la cause de mon chagrin</seg>
<seg ana= "det rlex p3sdo pnnocor" > tu le sais bien la cause de mon chagrin </seg> tous les matins je vais au Suk et au Suk je vois les boutiques des autres marchands et ils ont avec eux un deux trois quatre de leurs enfants ils en sont fiers ils les présentent aux autres et <seg ana= "det lpn p1ss" > moi je suis toujours seul seul </seg> </u>

In this section, <u>s are separated out into different *locuteurs*, with sections of 'NS' where the narrator recounts the story, and sections of DD and fDD where the characters are the *locuteurs*. The two examples of negation show the nature of the negative marker, the *forclusif*, the subject (the impersonal pronoun 'il' in both cases), and other relevant factors such as the presence of a double negative and an inserted element ('insm') between the subject and verb. The two examples of inversion involve *incises* with a nominal subject 'incin', while the examples of detachment are encoded for the detached element (two right-detached lexical elements and a left-detached pronoun), the replacement pronoun (two third person singular direct object pronouns and substitute 'one example of 'ce'') and the fact that in one case, the replacement pronoun is not straightforwardly co-referential ('pnnocor').

https://doi.org/10.1017/S0959269507003183 Published online by Cambridge University Press

What then are the advantages of the TEI system? Perhaps the most significant is its international currency and the fact that the Guidelines are suitable for a wide variety of text types: the majority of texts to have used the TEI are written but the principles work equally well for transcribed oral data for which the researcher is likely to make heavy use of the subset of TEI elements relating to spoken language. Linked to this is the fact that, as has already been mentioned, the same markup language, i.e. XML, is used for all types of information, including both the metadata and textual markup. As is the case with XML-based texts, the text is normally prepared within an Editor (in this case Oxygen (www.oxygenxml.com)) that can check for XML well-formedness (i.e. that it obeys the rules of XML syntax) and for validity (i.e. that it obeys the taxonomical constructs contained in the DTD).[20]

The second major advantage is the balance between common elements available (i.e. elements that all TEI-encoded texts have in common) and the flexibility within certain attributes such as 'ana' to devise attribute values that are absolutely unique to a particular project and tailored to meet its analytical needs. So for example, in a project such as this, researchers from different disciplines could read the basic text, or the text with encoding exclusively for pauses, vocal factors etc., or could add new attribute values tailored to their own purposes under the 'ana' taxonomy. On the other hand, for those working on the four topics relevant to this project, the 'ana' taxonomy in the Header allows other researchers to interpret the encoding that has taken place, while changes are easy to make, in that the researcher can adapt the 'ana' taxonomy without difficulty.

Once the file is created, there exist a very large number of possibilities for searching it, whether for qualitative or quantitative research, including creating customised outputs such as word frequencies, collocations, statistics etc., or using linguistic and text analysis software like Xaira (www.oucs.ox.ac.uk/rts/xaira/) or plugging it into an XML database and performing more complex structural queries using XQuery. There are also search facilities within Oxygen, such as XPath. In the case of this project, broad questions can be asked, such as the % of possible negative *ne/n'* where *ne/n'* is attested (versus those where there is no *ne* or *n'*), in order to get a broad view of the use of *ne*, and the same question can be asked of individual storytellers and/or particular types of *conte* and/or different types of performance context. A quantitative enquiry concerning detachment could involve searching for examples of left- and right-detachment with different parts of speech, i.e. examining all the examples of left- and of right-detachment and in each case calculating the percentage for lexical items, pronouns etc. Any encoded linguistic phenomenon can be searched within any particular story or storyteller or type of story or performance context. Similarly, since each <seg> appears within a <u>

[20] The DTD (document type declaration) for this project was designed by Dr James Cummings of the Oxford Text Archive.

https://doi.org/10.1017/S0959269507003183 Published online by Cambridge University Press

which is encoded for speech and thought presentation, any calculation can be done for occurrence within any type of discourse presentation. In short, using this type of markup offers an extremely broad range of searching and statistical options.

On the negative side, some researchers may find the TEI overly verbose in terms of how the combination of XML language and the tagset operates. This is due to a combination of several factors, not least the fact that the elements themselves are explicit rather than consisting of typographical symbols (as is the case for example with DELIC). Moreover, each example has to be opened and closed with a <seg> and a </seg>; these opening and closing mechanisms are a basic principle of XML based systems. The sense of verbosity in the coding is most acute where there are <seg>s within <seg>s, where the <seg> element appears at least four times in the coding (e.g. to open and close two sets of attributes and attribute values). However, this verbosity is more apparent than real, in the sense that any selection of encoding elements can be hidden and, in any case, an extremely readable version can be generated through the use of customisable XSLT style sheets, with the result that the text can be viewed in a web browser with desired sections highlighted or PDF files created for publication. More importantly, the verbose encoding is actually a key design principle of the TEI in that both TEI and XML are specifically designed to be neutral and human readable (as opposed to compressed) with long-term storage of data and documentation in mind.

There is also a potentially problematic assumption in XML that the data is parsable in such a way that there will be no cross-cutting of elements (see Leech 2004: 23), or what are referred to technically as 'multiple hierarchies' or 'XML overlap'; e.g. it is not possible to have a <seg> which is tagged for negation cutting across a <seg> which is tagged for detachment (i.e., to use Leech's terminology, a situation where one finds <x> <y> </x> </y>, as opposed to an embedded <seg> which is unproblematic <x> <y> </y> </x>).[21] With XML overlap, the current solutions, although resolving the problem, add to the verbosity of the markup, although the new version, TEIP5, will propose easier solutions.

Perhaps more significantly for linguists, the TEI is most easily used for linguistic corpora where the focus is primarily on a text (whether a published text or a transcription of oral data), rather than on phonological or phonetic analysis. Some phonetic information (e.g. changes of pitch, pauses etc.) as well as some information on intonation can of course be encoded using the <vocal> element.[22] There are also multi-modal possibilities. As is always the case, these are more complex to use, but it is possible to link the text with audio and/or video recordings within the TEI.

Finally, using the TEI obviously requires a certain level of familiarity with XML and thus some investment in training and practice; it is probably a relatively

[21] Note that with an embedded <seg>, the screen shows which opening and closing elements belong together.

[22] See Johansson (1995) for an account of how the tags can be used to markup intonational features.

https://doi.org/10.1017/S0959269507003183 Published online by Cambridge University Press

labour-intensive system in the early stages of a project. On balance however, the TEI methodology provides an excellent mechanism for tailoring the encoding to the needs of a particular project while using a language which has international currency, long-term value and where the technology is constantly being enhanced.

Author's address:
Janice Carruthers
French Studies
Queen's University Belfast
Belfast BT7 1NN
Northern Ireland
UK
e-mail: j.carruthers@qub.ac.uk

REFERENCES

Armstrong, N. (2001). *Social and Stylistic Variation in Spoken French. A Comparative Approach*. Amsterdam/Philadelphia: Benjamins.

Armstrong, N. and Smith, A. (2002). The influence of linguistic and social factors on the recent decline of French 'ne'. *Journal of French Language Studies*, 12: 23–41.

Ashby, W. (1982). The drift of French syntax. *Lingua*, 57: 29–46.

Ashby, W. (1988). The syntax, pragmatics and sociolinguistics of left- and right-dislocations in French. *Lingua*, 75: 203–229.

Ashby, W. (2001). Un nouveau regard sur la chute du ne en français parlé tourangeau. *Journal of French Language Studies*, 11: 1–22.

Ayres-Bennett, W., Carruthers, J. with Temple, R. (2001). *Problems and Perspectives. Studies in the Modern French Language*. London: Longman.

Barnes, B. (1985). *The Pragmatics of Left Detachment in Spoken Standard French*. Amsterdam: Benjamins.

Beeching, K. (2002). *Gender, Politeness and Pragmatic Particles in French*. Amsterdam-Philadelphia: Benjamins.

Bilger, M. (ed.) (2000). *Corpus. Méthodologie et applications linguistiques*. Paris: Champion.

Bilger, M. and Cappeau, P. (2004). L'oral ou la multiplication des styles. *Langage et société*, 109: 13–30.

Blasco-Dulbecco, M. (1999). *Les Dislocations en français contemporain: étude syntaxique*. Paris: Champion.

Cappeau, P. (1992). *Le Sujet postposé en français contemporain*. Université de Provence: thèse nouveau régime.

Cappeau, P. and Seijido, M. (2005). *Les Corpus oraux en français*. www.culture.gouv.fr/culture/dglf/recherche/corpus_parole/Presentation_Inven taire.pdf

Carruthers, J. (2005). *Oral Narration in Modern French. A Linguistic Analysis of Temporal Patterns*. Oxford: Legenda.

Carruthers, J. and Marnette, S. (2007). Tense, voices and point of view in medieval and modern oral narration. In: E. Labeau, C.Vetters and P. Caudal (eds), *Sémantique et diachronie du système verbal français*. Amsterdam: Rodopi, pp. 177–202.

Clifford, P. (1973). *Inversion of the Subject in French Narrative Prose from 1500 to the Present Day*. Oxford: Blackwell.

https://doi.org/10.1017/S0959269507003183 Published online by Cambridge University Press

Coveney, A. (1996). *Variability in Spoken French. A Sociolinguistic Study of Interrogation and Negation*. Exeter: Elm Bank.

Coveney, A. (1998). Awareness of linguistic constraints on variable 'ne' omission. *Journal of French Language Studies*, 8: 159–187.

De La Salle, B. (2004). *Plaidoyer pour les arts de la parole*. Vendôme: CLIO.

Delais-Roussarie, E. (2000). Constitution et annotation de corpus: méthode et recommandations. In: Elisabeth Delais-Roussarie and Jacques Durand (eds.), *Corpus et variation en phonologie du français*. Toulouse: Presses Universitaires du Mirail, pp. 91–125.

Engel, D. (1999). Radio talk: French and English perfects on air. *Languages in Contrast*, 2.2: 255–277.

Johansson, S. (1995). The approach of the Text Encoding Initiative to the encoding of spoken discourse. In: G. Leech, G. Myers and J. Thomas (eds.), *Spoken English on Computer*. Harlow: Longman, pp. 82–98.

Koch, P. and Oesterreicher, W. (2001). Gesprochene Sprache und geschriebene Sprache (Langage parlé et langage écrit). In: G. Holtus, M. Metzeltin and C. Schmitt, *Lexikon der Romanistischen Linguistik*, Band 1.2. Tübingen: Max Niemeyer Verlag, pp.584–627.

Lahousse, K. (2003). NP subject inversion in French and (preposed) adverbs. *Amsterdam Studies in the Theory and History of Linguistic Science*, 4:161–176.

Lahousse, K. (2006). NP subject inversion in French: two types, two configurations. *Lingua*, 116: 424–461.

Leech, G. (2005). Adding linguistic annotation. In: M. Wynne (ed.), *Developing Linguistic Corpora. A Guide to Good Practice*. Oxford: Oxbow, pp.17–29.

Leech, G. and Short, M. (1981). *Style in Fiction: A Linguistic Introduction to English Fictional Prose*.

Lyons, J. (1981). *Language and Linguistics*. Cambridge: Cambridge University Press.

Marnette, S. (2005). *Speech and Thought Presentation in French*. Amsterdam: Benjamins.

Montelle, E. (1994). A quel temps se vouer. *Dire*, 21: 35–42.

Postal, P. (1989). *Masked Inversion in French*. Chicago: University of Chicago Press.

Semino, E. and Short, M. (2004). *Corpus Stylistics. Speech, Writing and Thought Presentation in a Corpus of English Writing*. London: Routledge.

Söll, L. (1983). *Gesprochenes und geschriebenes Französisch*. Berlin: Schmidt.

Thompson, P. (2005). Spoken language corpora. In: M. Wynne (ed.), *Developing Linguistic Corpora. A Guide to Good Practice*. Oxford: Oxbow, pp. 59–70.

https://doi.org/10.1017/S0959269507003183 Published online by Cambridge University Press