Encoding a parallel corpus: The TRIS corpus experience

Carla Parra Escartín

University of Bergen

Abstract

This paper focuses on one of the many aspects to be taken into account when developing a new corpus: its encoding. During the compilation of the corpus of Technical Regulations Information System (the TRIS corpus) several encoding issues arose. In this paper the author discusses the possibilities available with regards to encoding as well as the decisions taken and the strategies followed. The author discusses standards for character encoding and corpus markup and explains how these were integrated in the compilation of the TRIS corpus.

Keywords: corpus planning, parallel corpora compilation, corpus encoding, standardization

*** Principal contact:**
Carla Parra Escartín
Marie Curie Early Stage Researcher
Language Models and Resources Research Group (LaMoRe)
Department of Linguistic, Literary and Aesthetic Studies (LLE)
University of Bergen, HF-bygget, Sydnesplassen 7 N-5007 Bergen, Norway
Tel.: +47 55 58 89 45
E-mail: carla.parra@uib.no

1. Introduction

This paper will discuss several issues related to corpus encoding and the use of available encoding standards applicable to the compilation of corpora. To illustrate this, the compilation process of the corpus of Technical Regulations Information System (in what follows the TRIS corpus) is used. The TRIS corpus is being compiled for the purposes of a larger project which aims at researching the translational correspondences between German nominal compounds and their Spanish phraseological correspondences. Details about its compilation process and its main characteristics can be found in Parra Escartín (2012).

According to the Collins Cobuild online dictionary[1], encoding in computing is "*the action of converting (characters and symbols) into a digital form as a series of impulses*". The Tech Terms Computer Dictionary[2] refers to it as "*the process of converting data from one form to another* " and specifies that *"there are several types of encoding, including image encoding, audio and video encoding, and character encoding"*. Thus, when we refer to the encoding of a corpus we may be referring to different aspects and even different kinds of encoding. My experience in compiling the TRIS parallel corpus has made me aware of this fact. This paper aims to discuss the role of encoding at different stages of a corpus compilation process. This is done to illustrate the role it plays in each phase.

The remainder of this paper is divided into sections which follow what could be considered the logical progression of a corpus compilation. At each phase the problems and challenges faced are explained and discussed as well as the strategies adopted and the decisions taken. In the next section (Section 2), I first explain the role of encoding within the compilation of a corpus. Section 3 focuses on the importance of character encoding and its role in corpora and Section 4 is devoted to the different types of markup that we may choose for a corpus.

2. The corpus encoding workflow

In order to understand the role of encoding in the compilation process of a corpus it is important to see at which stages it plays a particular role. If we take into account the definitions given in Section 1, the very first phase of the compilation process already implies several changes in the encoding of the files included in the corpus. In the case of the TRIS corpus, the files were automatically retrieved from the Database of the DG Enterprise and Industry Project[3] of the European Commission by means of a crawler (a computer program capable of performing recursive searches)[4]. After all files in outdated formats no longer available and corrupted files were disregarded, every remaining file was classified according to its original format. MS Word files were directly stored for later verification while PDF files underwent a further process. PDF "text" files were automatically converted to MS word, while PDF "scanned image" files were processed with ABBYY FineReader – an Optical Character Recognition (OCR) software – and converted to MS Word. Finally, all MS Word files were proofread and verified manually to ensure that no conversion problems had arisen. Figure 1 below illustrates the process that every crawled file underwent prior to being aligned.

[1] http://www.collinsdictionary.com/dictionary/english/encoding
[2] http://www.techterms.com/definition/encoding
[3] http://ec.europa.eu/enterprise/tris/index_en.htm
[4] For details please see Parra Escartín (2012).

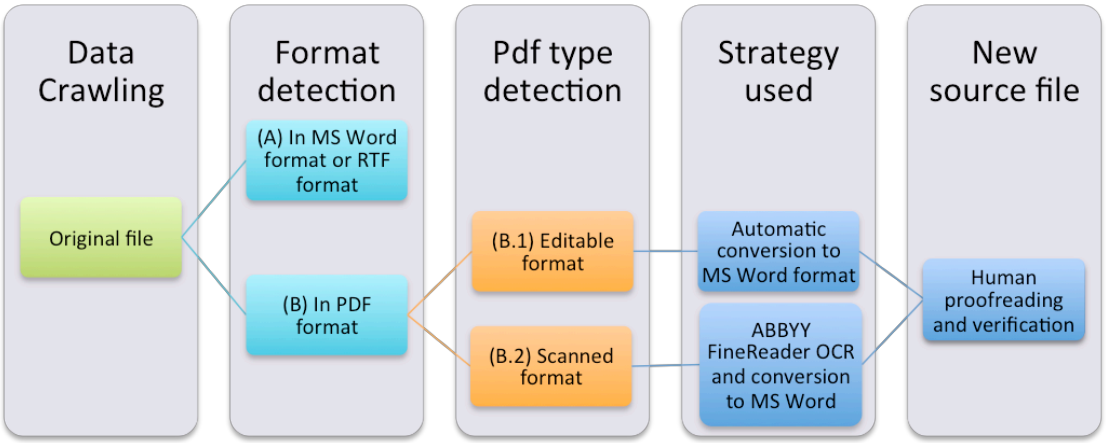

Figure 1: File selection and conversion process prior to alignment

After all files were considered ready, file pairs in German and Spanish were also verified and their formatting was checked to ensure that it matched and that it would not provoke any problems at the alignment stage. In the next phase – still in process –, MS Word files are aligned using SDL Trados WinAlign, a proprietary software programme within the suite of the Computer Assisted Translation tool (CAT tool) SDL Trados Studio 2009[5]. WinAlign automatically converts the files to RTF (Rich Text Format) and once the alignment has been manually verified and confirmed it can be exported as a translation memory in the SDL Trados proprietary format or in the *de facto* standard format TMX (Translation Memory eXchange)[6]. In the case of TRIS the translation memories corresponding to each individual file are exported in the SDL Trados proprietary format; then they are merged and converted to TMX format; and finally they are converted to TEI P5 format. Simple plain text documents with one sentence per line are also created from the TMX files. These files will be subsequently Part-of-Speech (POS) tagged. Figure 2 illustrates how the original MS Word files are transformed into different formats at the different stages of the corpus compilation.

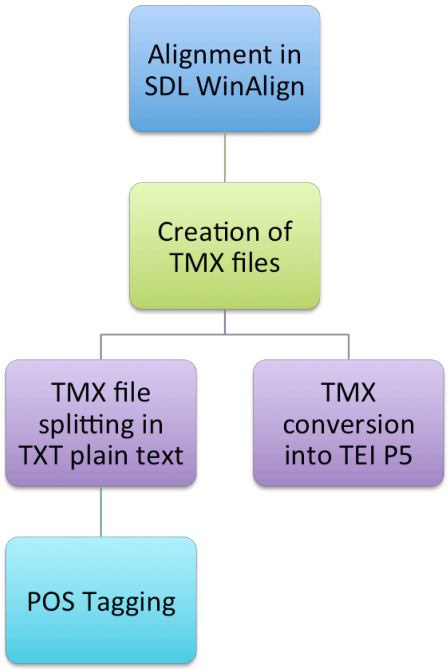

Figure 2: Different file encoding stages during the corpus compilation process

[5] http://www.sdl.com/products/sdl-trados-studio/
[6] The TMX format is explained in Section 4.2.

Finally, it is also worth mentioning that the corpus is to be released in different encoding formats to facilitate its reusability in other research projects. Concretely, the corpus will be released in plain text, POS-tagged text, TMX and TEI P5. This choice is grounded on several reasons. First of all, and as argued by Wynne (2005), it is important to avoid proprietary formats. As he points out:

> If your corpus is made up of files in a format for a commercial wordprocessing program, such as Microsoft Word, then they cannot be processed by most corpus analysis tools. What is more, the format may not be supported indefinitely into the future, and there will come a time when users won't be able to read the files any more.

Wynne (2005) continues arguing that encoding a corpus in XML is usually a good choice since it not only is appropriate for its long-term preservation but also ensures the usage of Unicode for encoding the text. The TMX and TEI P5 encoding formats are actually markup formats in XML as we shall see later in Section 4. The other two formats in which the corpus is released have been chosen to satisfy the needs of the research project in which the TRIS corpus will be first used. Generic tools often require "raw text" or plain text files to work, and thus I had to produce them for my own research. Additionally, I also needed POS-tagged files to run experiments and more concretely files in the TreeTagger[7] format. Providing these two additional formats along with the other two standard formats will enable the reusability of the corpus without requiring prior conversion processes.

3. Character Encoding: The minimal kind of encoding but yet a critical one

Character encoding may be considered the minimal kind of encoding. However, it is crucial as it will determine whether or not a text is appropriately displayed in a user's computer. McEnery and Xiao (2005) offer an extensive and clear overview of the importance of character encoding as regards corpus construction as well as of its evolution across history. As they point out, "*character encoding in a corpus must be consistent if the corpus is to be searched reliably*". In fact, something that may seem as simple as character encoding is not trivial. During the compilation of the TRIS corpus several encoding problems arose when manipulating the files in the corpus. This is something that McEnery and Xiao (2005) also mention: "*In many cases, however, multiple and often competing encoding systems complicate corpus building, providing a real problem*".

Many efforts have been made over time to ensure readability and interoperability as regards character encoding in different operating systems. The Unicode standard has been the result of these common efforts and it is commonly used nowadays in many cross-platform applications. It includes three encoding formats: UTF-8, UTF-16 and UTF-32 (Unicode Transformation Format 8 bits, 16 bits and 32 bits respectively). One of its main strengths is that it is 100% backward compatible with ASCII (McEnery and Xiao, 2005). Sasaki (2010) explains the differences between the three of them:

> The most widely used encoding form is UTF-8. If the multilingual corpus contains only Latin based textual data, UTF-8 will lead to a small corpus size, since this data can be represented mostly with sequences of single bytes. If corpus size and bandwidth are no issues, UTF-32 can be used. However, especially for web based corpora, UTF-32 will slow down data access. UTF-16 is for environments which need both efficient access to characters and economical use of storage. Finally, the aspect that an XML processor must

[7] The TreeTagger is a tool for annotating text with part-of-speech and lemma information developed at the Institute for Computational Linguistics of the University of Stuttgart. More information can be found at its website: http://www.ims.uni-stuttgart.de/projekte/corplex/ TreeTagger/.

> be able to process "only" UTF-8 and UTF-16, and not necessarily other encoding forms, should be taken into account when deciding about the appropriate encoding form.

From his reasoning it can be concluded that UTF-8 was the right choice for the TRIS corpus as it only includes Latin based textual data and therefore there was no need for using an encoding format that would imply a larger size such as UTF-16.

The files of the TRIS corpus were not originally encoded in UTF-8. The translation memory files were obtained in a Windows Operating System because the software used for alignment (SDL Trados WinAlign) is not available in other operating systems. However, when manipulating the files in another operating system –a Mac OS–, problems arose because Windows uses its own proprietary encoding (ISO Latin 1) which in turn is not compatible with Macintosh and other operating systems. This problem is easy to overcome by automatically converting the encoding format. To ensure the future readability and reusability of the TRIS corpus, the original ISO Latin 1 (also known as ISO 8859-1) encoding produced by SDL Trados WinAlign was converted to UTF-8. This was done using the command displayed in Figure 3 which instructs the computer to automatically convert from ISO-8859-1 to UTF-8 encoding all .txt files in the directory we are currently in. The character encoding conversion was done prior to the conversion of the aligned files in the Trados proprietary encoding format to the standard TMX format.

```
for file in *.txt; do iconv -f ISO-8859-1 -t UTF-8 $file > $file.utf.txt; done;
```

Figure 3: Unix command to automatically convert Latin1 files to UTF-8

4. Corpus Markup

As defined in Morrison et al. (2000), markup is *"a form of text added to a document to transmit information about both the physical and electronic resource"*. I will not discuss here the benefits of using a common and standardized markup framework as it has already been widely discussed, reasoned and agreed upon. Instead, I will focus on the different standards that are available with regards to corpus markup. In this paper, the term "standard" is not restricted to official standards such as ISO, ETSI or OASIS standards and therefore may also be used to refer to markup formats which are regularly and widely used. This section is divided in three subsections: one in which the markup languages SGML and XML are introduced (4.1), another one in which industrial standards are discussed (4.2) and a final one with a special focus on the linguistic markup of linguistic resources (4.3).

4.1. Brief introduction to markup: SGML and XML

SGML (Standard Generalized Markup Language) and XML (EXtensible Markup Language) are structured markup languages. HTML (Hypertext Markup Language), for example, is a type of SGML used to mark up text and graphics so that the most popular web browsers can interpret them. To identify the markup in a document, both SGML and XML use named elements delimited by angled brackets ("<" and ">"). As explained in (Walsh and Muellner, 1999), *"An essential characteristic of structured markup is that it explicitly distinguishes (and accordingly "marks up" within a document) the structure and semantic content of a document. It does not mark up the way in which the document will appear to the reader, in print or otherwise."* Moreover, the structure of the documents is controlled by either document type definitions (DTDs) or XML schema. A DTD is a set of declarations regarding the structure of a document, and its goal was to retain a level of compatibility with SGML for applications that might want to convert SGML DTDs into XML DTDs. It consists of a list of tag names and specifies their combination rules and it is also used to check that a particular document is appropriately structured.

While SGML was commonly used in the past, there has been a shift of markup language and nowadays it is more common to use XML. In fact, all the markup standards that will be discussed in the next subsections have either moved towards XML or were already conceived in XML.

4.2. The Translation Memory eXchange (TMX) and other LISA standards. Industrial Standards entering into Academia and beyond

TMX stands for Translation Memory eXchange and it is an XML format to encode translation memories and ensure that they can be reused and exchanged among different CAT tools without encountering any troubles. It was developed by the Localization Industry Standards Association (LISA) and after having been widely adopted in the industrial sector it has made its way into the academic and institutional sector as well. In fact some of the Language Technology Resources released by the European Commission are in this format. Examples of this are the DGT-Translation Memory[8] and the ECDC-TM; the Translation Memory of the European Centre for Disease Prevention and Control[9]. Its increasing presence as an encoding format has led to the appearance of tools to extract TMX files and convert them to simple .txt UTF-8 files if needed. This is the case of the extract-tmx-corpus tool[10], which is currently used to prepare input files for the Statistical Machine Translation System MOSES[11].

LISA was sadly dissolved in March 2011 but its contributions towards standardization in the Localization Industry were of great magnitude and some of the standards developed by them are still widely used. The body in charge of creating new standards was a specific committee called OSCAR (Open Standards for Container/Content Allowing Reuse) and as a result of their work five community standards were successfully published: the Translation Memory eXchange (TMX)[12], the TermBase eXchange (TBX)[13], the Segmentation Rules eXchange (SRX)[14], the Global information management Metrics eXchange Volume (GMX-V)[15] and the XML Text Memory (xml:tm)[16].

As can be inferred from the previous paragraph, LISA – an industrial initiative to cooperate and standardize the localization field – was a very important agent as regards standardization. It cooperated with the relevant agents in the field to ensure the success of its proposals: the ISO TC 37 group, OASIS XLIFF and the Open Architecture for XML Authoring and Localization (OAXAL). As stated in the TBX definition (Open Standards for Container/Content allowing Reuse, 2008), the TBX, for instance, is actually identical to ISO 30042.

When its dissolution was announced, the European Telecommunications Standards Institute (ETSI), worked together with LISA on a proposal to create a new Industry Specification Group (ISG) for Localisation Industry Standards (LIS), which would ensure the maintenance of the five LISA OSCAR standards mentioned above as well as the cooperation with LISA's cooperating partners. As stated in Guillemin and Trillaud (2012), *"the ETSI is a standardization institute which produces standards from information and communications technology, including fixed, mobile, radio, converged, aeronautical, broadcast and internet technologies and is officially recognized by the European Union as an European Standards Organization. ETSI is an independent, not-for-profit association with more than 700 member companies and organizations, drawn from 62 countries across five continents worldwide, that determine its work program and participate directly in its work"*. Guillemin and Trillaud (2012) offer a summarized explanation of

[8] http://ipsc.jrc.ec.europa.eu/?id=197
[9] http://ipsc.jrc.ec.europa.eu/?id=782
[10] http://code.google.com/p/extract-tmx-corpus/
[11] http://www.statmt.org/moses/
[12] http://www.gala-global.org/oscarStandards/tmx/tmx14b.html
[13] http://www.gala-global.org/oscarStandards/tbx/tbx_oscar.pdf
[14] http://www.gala-global.org/oscarStandards/srx/srx20.html
[15] http://www.gala-global.org/oscarStandards/gmx-v/gmx-v.html
[16] http://www.gala-global.org/oscarStandards/xml-tm/xml-tm.html

LISA's dissolution and what was done to ensure the continuity of the standards developed within this professional association.

As of February 2013, the ETSI has officially released the TMX as *ETSI ISG LIS GS Translation Memory eXchange (TMX)*[17] and the GMX-V as *Global information management Metrics eXchange Volume (GMX-V)*[18]. The XML Text Memory (*ETSI ISG LIS GS XML Text Memory (xml:tm)*) has reached the status of a stable draft[19], and the TBX (*ETSI ISG LIS Term-Base eXchange (TBX)*) is still an early draft[20], as is the SRX (*ETSI ISG LIS Segmentation Rules eXchange (SRX)*)[21].

The efforts made to ensure the continuity of the standards despite LISA's dissolution are a proof of the importance that they have acquired for industry, academia and the public sector. TMX and TBX are probably the two standards most related to the Natural Language Processing (NLP) field and as exemplified above, TMX is in fact starting to be a standard used for the release of new linguistic resources.

Converting the TRIS corpus into TMX

As has been mentioned in Section 2, for the alignment of the MS Word files the commercial software SDL Trados WinAlign is used. One of the reasons behind this decision is that sentence alignment can be carried out from native MS Word files and no format conversion prior to alignment is required. Moreover, the decision was taken due to practical reasons: WinAlign saves time at this stage of the process while producing bilingual files either in its own proprietary format or in TMX.

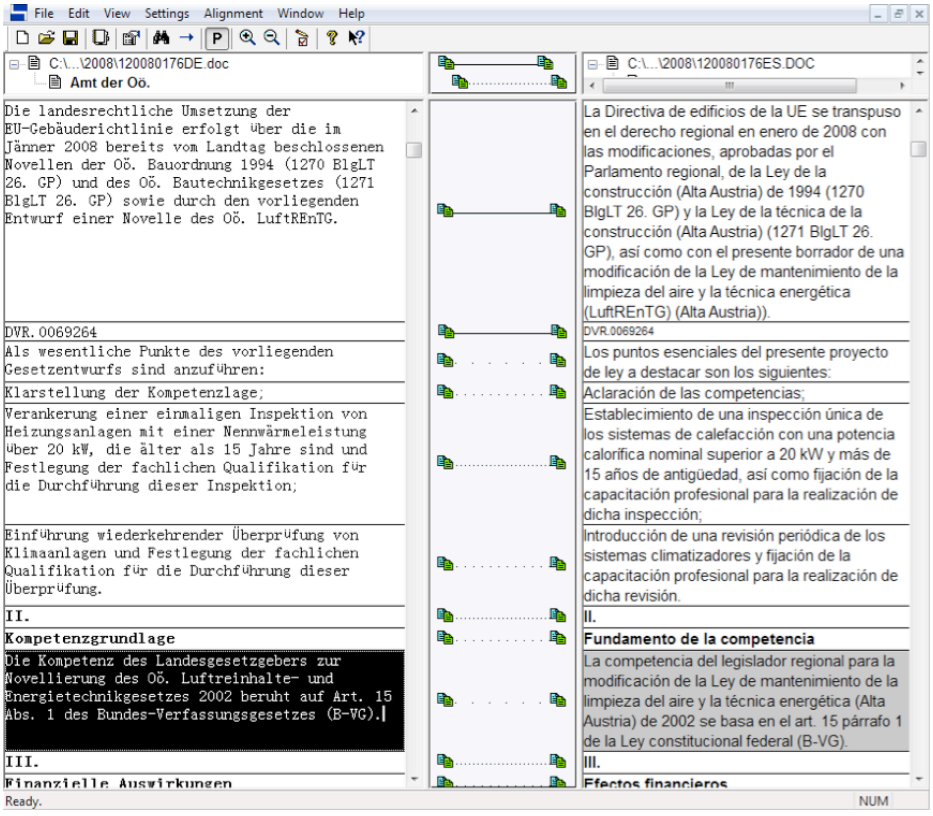

Figure 4: The SDL Trados WinAlign Interface

[17] http://www.etsi.org/deliver/etsi_gs/LIS/001_099/002/01.04.02_60/gs_ LIS002v010402p.pdf
[18] http://www.etsi.org/deliver/etsi_gs/LIS/001_099/004/02.00.00_60/gs_ LIS004v020000p.pdf
[19] http://webapp.etsi.org/WorkProgram/Report_Schedule.asp?WKI_ID=37769
[20] http://webapp.etsi.org/WorkProgram/Report_Schedule.asp?WKI_ID=37750
[21] http://webapp.etsi.org/WorkProgram/Report_Schedule.asp?WKI_ID=37767

Figure 4 shows the user interface of WinAlign. As can be seen, the program proposes automatic alignments (dotted lines), and a human validator can correct those alignments, confirm (line) or reject them (no line at all). The program also permits the user to join or split segments as well as edit them if needed. This is very useful as sometimes it is necessary to join several segments into one. This is the case, for example, when in the original MS Word file in German there is a list with the verb in a separate line at the end of the list while in the Spanish translation the verb occurs at the beginning of the list. German grammar requires that certain structures have the verb at the end and this cannot be done in Spanish.

The editing feature of WinAlign allows the user to edit the text in the segments (e.g. to correct typos not previously detected) and join/split them accordingly so that they are paired with the appropriate sentence in the other language.

Figure 5 illustrates the structure of an aligned segment produced by SDL Trados WinAlign in the .rtf format that the program uses internally. Furthermore, as mentioned earlier WinAlign also allows the user to export the alignment as a TMX file. One drawback of Trados is that the resulting translation memories (TMs) include a lot of unnecessary formatting information that has to be cleaned before further exploitation of the corresponding files. Another drawback is that when merging several TMs into one, the program filters out all duplicates and deletes them and it does not keep track of the order in which sentences appear in the text. This is because it is a Computer Assisted Translation Tool and these details are not relevant for its intended usage.

```
<TrU>
<Quality>100
<CrU>ALIGN!
<CrD>26012012, 21:11
<Seg L=DE-AT>Nach außergewöhnlichen Ereignissen, wie z. B. länger
anhaltenden extremen Temperaturen, Hochwasser, Erdbeben, Lawinen-
oder Murenabgängen, Rutschungen, Unfällen, Feuer oder Anprall von
Fahrzeugen udgl. sind die Wegweiserbrücken gezielt auf die möglichen
Auswirkungen der außergewöhnlichen Umstände hin zu besichtigen.
<Seg L=ES-ES>De haberse producido algún hecho extraordinario, como
por ejemplo, temperaturas extremas de muy larga duración, riadas,
seísmos, aludes o argayos, corrimientos, accidentes, incendios o impactos
de vehículos y similares, se deberán inspeccionar los pórticos para
mensajes de carretera específicamente en cuanto a las posibles
repercusiones de las circunstancias extraordinarias.
</TrU>
```

Figure 5: Sample of an aligned segment produced by SDL Trados WinAlign (abbreviated)

To overcome these challenges, another industrial application is used: ApSIC Xbench[22]. ApSIC Xbench supports several input formats (such as TMX and Trados' proprietary .rtf format) and allows the user to merge several translation memories without removing duplicates and respecting the order in which they appear. Thus, this tool is used to merge all single files into one file per subdomain in the corpus and convert them to TMX. Even though the TMX format is not really necessary for my research project (simple plain monolingual files with one sentence per line would have been enough), I deemed it appropriate to convert the resulting translation memories into TMX as this has become a standard in our field and would ensure interoperability and reusability in the long run. The TMX files are further processed with a python script to add additional information to each sentence in the corpus.

[22] http://www.apsic.com/en/products_xbench.html

Figure 6 illustrates the structure of the final TMX files. As can be seen, all TMX documents are divided into a header and a body element. The structure of any TMX document is appropriately described and documented in the TMX definition released by ETSI (Localization Industry Standards (LIS) ETSI Industry Specification Group (ISG), 2013). What follows is a brief summary of the information that can be found there.

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE tmx PUBLIC "-//LISA OSCAR:1998//DTD for Translation Memory eXchange//EN"
   "http://www.ttt.org/oscarstandards/tmx/tmx14.dtd">
<tmx version="1.4">

<header
creationtool="SDL Trados WinAlign 8.3.0.863"
creationtoolversion="Edition 8 Build 863"
o-tmf="SDL TM8 Format"
segtype="sentence"
adminlang="EN-US"t
srclang="DE-AT"
datatype="xml"
creationdate="June 2012"
creationid="Carla Parra, UiB"
>
</header>

<body>

<tu tuid="B00Y1999File119990211S7" creationdate="20111114T1935Z" creationid="ALIGN!">
<tuv xml:lang="DE-AT">
<seg>Das Fangsystem (beispielhaft in Abb 1 dargestellt) dient dazu, Verbrennungsgase
von Feuerstätten mit niedrigen Verbrennungsgastemperaturen (mit diesbezüglichen
Sicherheitseinrichtungen) ins Freie zu leiten.</seg>
</tuv>
<tuv xml:lang="ES-ES">
<seg>El sistema de chimenea (representado a título de ejemplo en la figura 1) sirve para
conducir al exterior los gases de combustión procedentes de hogares con baja temperatura
de los gases de combustión (con los correspondientes dispositivos de seguridad).</seg>
</tuv>
</tu>

<tu tuid="B00Y1999File119990211S8" creationdate="20111114T1935Z" creationid="ALIGN!">
<tuv xml:lang="DE-AT">
<seg>Als Feuerstätten kommen zB Brennwertgeräte jeweils mit Gas oder Heizöl extra leicht
als Brennstoff in Betracht.</seg>
</tuv>
<tuv xml:lang="ES-ES">
<seg>Los hogares a considerar son por ejemplo los equipos de índice de combustión que
empleen como combustible, respectivamente, gas o fuel extra-ligero.</seg>
</tuv>
</tu>
...
</body>
</tmx>
```

Figure 6: Sample from a TMX aligned file (abbreviated)

The header – enclosed within the <header> </header> tags – contains the metadata about the document. The body – enclosed within the <body> </body> tags – contains all the translation units in the translation memory. In the header there is information related to the Tool with which a Translation Memory has been created and its version ("*creation tool*" and

"*creationtoolversion*" respectively); the original translation memory format ("*o-tmf* "); the kind of segmentation used ("*segtype*"); the default language in which the administrative and informative elements are written ("*adminlang*"); the source language of the translations included in the translation memory ("*srclang*"); the type of data we have ("*datatype*"); the creation date of that concrete translation memory ("*creationdate*"); and the identifier for the creator of the translation memory ("*creationid*").

The body of any translation memory consists of one or more translation unit elements (enclosed within <tu> </tu>), which in turn include one or more translation unit variants (enclosed within <tuv> </tuv>). In the TRIS corpus, the translation unit element consists of two translation unit variant elements. Besides, every translation unit is described by means of three attributes: "*tuid*"; "*creationdate*"; and "*creationid*". The attribute "tuid" (translation unit identifier) offers most of the information for every single sentence. For instance, the tuid *tuid="B00Y1999File119990211S7"* in Figure 6 stands for the construction domain (*B00*), Year 1999 (*Y1999*), file name 119990211 (*File119990211*), sentence 7 (*S7*). The attribute, "*creationdate*" contains information about the date and time in which the translation unit was created and the "*creationid*" refers to the creator of the translation unit. Its value usually corresponds to the user ID of the user who created the unit. In order to specify that a translation unit comes from an alignment tool, SDL Trados WinAlign assigns itself as the creator by using the value "ALIGN!".

The translation unit variant consists of a segment element and the information corresponding to that segment for a given language. The attribute "xml:lang" refers to the language variety used in the segment that appears below. Its value must be compliant with the RFC 3066 [6][23]. Thus, in the case of TRIS "*DE-AT* " refers to German (Austria) and "*ES-ES*" to Spanish (Spain). The text between the <seg> </seg> tags is the actual text and the fact that two translation unit variants are grouped together in a translation unit indicates that one is the translation of the other.

4.3. Standards currently being fostered within the NLP field

Current European initiatives such as Meta-share[24] are making major efforts towards the usage of standards and good practices in our field. Since the TRIS corpus is to be released through Meta-Nord, the Meta-share node to which the University of Bergen belongs, their documentation was consulted to decide which standards to use with regards to corpus encoding. As stated in Deliverable 4.1 of the Meta-Nord project[25]: *Metadata descriptions and other interoperability standards*, suitable standards for corpus encoding would be TEI or (X)CES (Borin and Lindh, 2011, p.15). Therefore, I decided that my corpus would use one of these two markup languages to ensure that it would be compliant with current initiatives on standardization, curation and sustainability of Language Resources and Tools (LRTs). The next two subsections (4.3.1 and 4.3.2) briefly explain each of them, while Subsection 4.3.3 discusses which of these two standards (TEI and (X)CES) is best and reasons the decision taken. Finally, Subsection 4.3.4 provides details about the encoding of the TRIS corpus in TEI P5 format.

4.3.1. The Text Encoding Initiative (TEI)

The Text Encoding Initiative (TEI) is a non-profit organization which counts in its consortium members from academia, research projects and individual scholars from around the world. In their website[26] they offer extensive documentation about the initiative as well as guidelines and a wide range of materials. Their main goal is to collectively develop and maintain the TEI guidelines for the encoding of texts in digital form. In order to reach a wide audience their

[23] http://www.ietf.org/rfc/rfc3066.txt
[24] http://www.meta-net.eu/meta-share
[25] http://www.meta-net.eu/
[26] http://www.tei-c.org/index.xml

Guidelines are aimed for their usage in Humanities, Social Sciences and Linguistics and since 1994 they have been used in a vast number of projects, institutions and resources.

Since their first release, the TEI guidelines are periodically updated and feedback from the user community is incorporated to fulfill user needs and requirements. The last release of the TEI Guidelines for Electronic Text Encoding and Interchange was done in late January 2013 and it accounts for version 2.3.0 of the TEI P5. Besides, although the current version is the TEI P5, resources encoded in previous versions, such as the TEI P4 format, can still be used without interoperability problems thanks to the usage of the corresponding DTD. An example of a resource encoded in a prior version of the standard but still widely used nowadays is the case of the JRC Acquis (Steinberger et al., 2006), which was released in TEI P4.

4.3.2. The XML Corpus Encoding Standard ((X)CES)

Another effort towards standardization of corpus encoding is the one carried out by the Expert Advisory Group on Language Engineering Standards (EAGLES[27]). As a result of their work a first Corpus Encoding Standard (CES)[28] was developed. It started being a SGML standard compliant with the specifications of the TEI Guidelines for Electronic Text Encoding and Interchange of the Text Encoding Initiative[29]. (X)CES stands for XML Corpus Encoding Standard and it is a newer version of CES encoded in XML. It is currently more frequently used than CES because XML has become the most currently used markup language. However, is not only an XML version of CES and as pointed out by Simões and Fernandes (2011) not all corpora which claim to be encoded in (X)CES are truly encoded in (X)CES but rather in CES encoded in XML: *"... some researchers claim they are releasing their corpora in XCES format, but they are just encoding CES in XML, and XCES is more than that."*

4.3.3. TEI and (X)CES. A Comparison

TEI and XCES have become the *de facto* standards for corpus encoding and most corpora are in one of the two formats or at least easily convertible to them.

Several papers (Przepiórkowski and Bánski, 2011; Przepiórkowski, 2009; Bánski and Przepiórkowski, 2010; Simões and Fernandes, 2011) refer to TEI as the standard and reference for corpus encoding and it seems reasonable to think of it for the encoding of newly compiled corpora. For the encoding of TRIS a comparison between the two standards was made with the aim of determining which seemed best.

The first drawback found in the case of XCES is its lack of documentation and authors like Przepiórkowski (2009) and Simões and Fernandes (2011), for example, already point this out. In fact, not knowing how the encoding should actually look like makes it particularly difficult to encode a corpus from scratch in this format. Przepiórkowski (2009) also states this as follows: *"http://www.xces.org/ refers to old CES documentation as "supporting general encoding practices for linguistic corpora and tag usage" and "largely relevant to the XCES instantiation", although the CES documentation is hardly applicable to the second version of XCES"*. In the same paper, Przepiórkowski (2009) also mentions as another reason against XCES *"the potential for confusion regarding the version of the standard (in particular, for many years DTD and XML Schema specifications co-existed on XCES web pages, without clear information that they specify different representations"*. The same is pointed out in another paper: *"There is a potential for confusion regarding the version of the standard. XCES was derived from TEI version P4, but it has not been updated to TEI P5 so far"* (Przepiórkowski and Bánski, 2011). In the XCES website[30] it is stated that *"XCES is continually under development and future work will include making the XCES*

[27] http://www.ilc.cnr.it/EAGLES/home.html
[28] http://www.cs.vassar.edu/CES/
[29] More information about the origins of CES can be found at their website: http://www.cs.vassar.edu/CES/.
[30] http://www.xces.org/

compliant with TEI P5". TEI P5 was released in November 2007 and is updated every six months. The last time the XCES website was updated was June 2008[31]. This highlights the outdatedness of XCES and contrasts with the willingness of the TEI community to keep their proposed standard up to date[32].

On the other hand, a possible drawback of TEI is its extensive documentation: the current version of the guidelines (January 2013) comprises 1641 pages. As Przepiórkowski (2009) points out, *"usually there is more than one way of representing any given annotation, so designing a coherent and constrained TEI-conformant schema for linguistic corpora is a daunting task"*.

TEI P5 was the standard chosen to encode the TRIS corpus due to what is argued above. Moreover, the active support and willingness to resolve doubts and make clarifications in the TEI mailing list were also a clear advantage towards choosing TEI. Finally, it also seemed the best option with regards to the interoperability and sustainability of a resource being developed since it is also periodically reviewed and documented.

XCES is not documented enough and – as mentioned in the previous Subsection 4.3.2 – the resources available in XCES are not always truly encoded in XCES but rather represent interpretations – own XML versions – of the previous CES format or schemata based on XCES. Deliverable D.2.1 of the Let's MT project offers a good example of this last issue. As Tiedemann and Wijnitz (2010, p. 6) explain, the alignment information of their parallel corpora will be stored *"in links between sentences in external files pointing to the appropriate documents using the unique sentence IDs for identification of the aligned segments"* and for this they *"will use a simple XML format* **based** *on the XCES standard"*[33]. If resource developers create new encoding formats based in XCES, they are not using the standard any more and therefore their resources will encounter interoperability problems in the long run.

4.3.4. The TRIS corpus in TEI P5 format

In this subsection the encoding of the TRIS corpus in TEI P5 will be briefly explained. As described in (Sperberg-McQueen and Burnard, 2009, p. 139), *"a full TEI document combines metadata describing it, represented by a <teiHeader> element, with the document itself, represented by a <text> element"*. The <teiCorpus> is a variant defined for the representation of language corpora or collections of texts. It consists of one or more complete <TEI> elements (i.e. elements consisting of a <teiHeader> and a <text> element) and additionally has its own <teiHeader> describing the whole corpus. This allows for a more general description of the corpus as a whole in the <teiHeader> element prefixed to the whole corpus, and a more detailed description of every <TEI> element comprised in the <teiCorpus> in their respective <teiHeader>. Chapter 15 of the TEI P5 Guidelines (Sperberg-McQueen and Burnard, 2009) describes how to encode a corpus. In what follows the encoding of the TRIS corpus is described to exemplify the TEI P5 structure of a teiCorpus.

First of all it must be pointed out that while it was clear that the <teiCorpus> element should be used, it was also necessary to establish the inner structure of the TRIS corpus as a whole and determine how it would be encoded. The TRIS corpus includes files written in Germany, Austria and Spain, thus originally written in either German or Spanish and translated into the other language. Furthermore, we have two language variants in the case of German: Austrian and German. The corpus also includes texts from different domains and subdomains and is ordered by year of publication from 1999 to 2010[34]. So far, only the texts for a particular

[31] The last time this was verified was February 2013.
[32] The last TEI P5 release was done in January 2013 and stands for version 2.3.0 of the standard.
[33] The emphasis is my own.
[34] See Parra Escartín (2012) for detailed information about the texts in the corpus.

domain (Construction) have been released for public usage[35] but other domains will be included shortly.

When designing the TEI structure it was decided to have a general <teiHeader> for the whole corpus and then have a <TEI> element for every domain and year. This makes it relatively easy to add new files on the fly once they are ready to be added to the corpus and does not prevent the corpus from being released beforehand.

I. The <teiCorpus> header. As explained above, the <teiCorpus> element contains information about the corpus as a whole. Every TEI-conformant text must have a header prefixed to it. TEI headers consist of four major parts that must be always included:

1. A *file description* (<fileDesc>): *"a full bibliographical description of the computer file itself, from which a user of the text could derive a proper bibliographic citation (...)"* (Sperberg-McQueen and Burnard, 2009)

2. An *encoding description* (<encodingDesc>): relates to how the source files where manipulated prior to encoding.

3. A *text profile* (<profileDesc>): contains classificatory and contextual information about the text.

4. A *revision history* (<revisionDesc>): contains information about the changes done during the development of the text.

Thus, the TRIS corpus starts as follows:

```
<teiCorpus version="5.2" xmlns="http://www.tei-c.org/ns/1.0">
        <teiHeader xml:lang="en" type="corpus">
```

Figure 7: Beginning of the TRIS corpus <teiCorpus> element of the TRIS corpus header

Where *version* refers to the TEI Guidelines version used (5.2) and *xmlns* is the namespace for the Text Encoding Initiative. Within the <teiHeader> element there are two attributes: the *xml:lang* attribute, which refers to the language in which the <teiHeader> is written, and *type*, which refers to the type of document it refers to.

Figure 8, Figure 9 and Figure 10 display the information provided in the header of the TRIS corpus in TEI. Since the current release is the only one done so far in TEI there is no <revisionDesc> element so far. As the names and values of the attributes are quite self-explanatory no further details are given. If the reader wants further information about the TEI Header, please see Chapter 2 of the TEI P5 Guidelines (Sperberg-McQueen and Burnard, 2009, p. 17–53).

35 http://metashare.nb.no/repository/browse/parallel-corpus-of-documents-from-the-technical-regulations-information-system-for-german-spanish-v02/d12552021dcc11e28f61001708556d5a64b9251fd03048ecaf7fe1abdc48a2d1/

```
<fileDesc>
    <titleStmt>
   <title>Parallel Corpus of documents from the Technical Regulations Information
        System for German-Spanish (v0.2)</title>
   <funder>EU under FP7, Marie Curie Actions, SP3 People ITN, grant agreement 238405
        (project CLARA)</funder>
        <principal>Carla Parra Escartín</principal>
        <respStmt>
            <name>Carla Parra Escartín</name>
            <resp>corpus compilation, processing, encoding and markup</resp>
        </respStmt>
    </titleStmt>
    <editionStmt>
   <edition n="V0.2">Version 0.2 which extends the previous version 0.1 and fixes some
        formatting errors detected there.</edition>
        <respStmt>
          <resp>Formatting errors in v0.1 corrected and corpus enlarged</resp>
            <name>Carla Parra Escartín</name>
        </respStmt>
    </editionStmt>
    <extent>
        <measureGrp>
            <measure type="files">205</measure>
            <measure type="sentences">70648</measure>
            <measure type="words_DE-AT">638907</measure>
            <measure type="words_ES-ES">923830</measure>
        </measureGrp>
    </extent>
    <publicationStmt>
        <address>
     <addName>Institutt for lingvistiske, litterære og estetiske studier</addName>
            <addrLine>Postboks 7805</addrLine>
            <addrLine>5020 Bergen (Norway)</addrLine>
        </address>
        <date>2012</date>
        <publisher>Universitetet i Bergen</publisher>
        <pubPlace>Bergen, Norway</pubPlace>
        <distributor>Universitetet i Bergen</distributor>
        <availability status="restricted">
            <licence>Under negotiation</licence>
        </availability>
    </publicationStmt>
    <sourceDesc>
        <p>98/34/EC Directive, TRIS Database, European Commission</p>
    </sourceDesc>
</fileDesc>
```

Figure 8: <fileDesc> element of the TRIS corpus header

```
<encodingDesc>
    <projectDesc>
        <p>
     Specialized parallel corpus Spanish-German (ES-ES, DE-AT and DE-DE), texts from
           the European Commission between 1997-2010.
           The texts are technical regulations in a variety of domains.
     The original files were either in MS Word or PDF format and have been converted to
           UTF-8 plain text.
      The corpus will be used in a project involving the study of phraseological
           translation correspondences between German and Spanish.
        </p>
    </projectDesc>
    <samplingDecl>
        <p>
      All texts written in either Austria, Germany or Spain with a corresponding
           translation into German/Spanish accordingly where crawled.
          Only texts in a readable format have been included in the collection.
     Texts belong to 10 different domains and are further classified in subdomains.
     The current version includes all texts from 1999 to 2010 written in Austria and
           for the construction domain.
         Images have been omitted and only the text in them has been preserved.
           Tables were converted to plain text.
           Formulae and mathematical expressions have been also omitted.
           All included texts have been aligned at sentence level.
        </p>
    </samplingDecl>
    <editorialDecl>
        <correction>
            <p>
                Orthotypographic errors have been corrected.
        Mismatching sentences have been omitted to ensure that every sentence has an
                alignment in the other language.
            </p>
        </correction>
        <segmentation>
            <p>
        <gi>s</gi> elements mark orthographic sentences and are numbered sequentially.
            </p>
        </segmentation>
    </editorialDecl>
</encodingDesc>
```

Figure 9: <encodingDesc> element of the TRIS corpus header

```
<profileDesc>
    <creation>
        <date when="2011-2012">September 2012</date>
        <rs type="city">Bergen, Norway</rs>
    </creation>
    <langUsage>
        <language ident="DE-AT">German (Austria)</language>
        <language ident="ES-ES">Spanish (Spain)</language>
    </langUsage>
</profileDesc>
```

Figure 10: <encodingDesc> element of the TRIS corpus header

II. The <teiHeader>. After the header for the whole corpus, the teiCorpus structure requires a TEI element with its own header describing that particular element of the corpus. This header "inherits" the general characteristics from the upper one in the corpus and thus provides the specific information related to the text being encoded in its <text> attribute. Attributes and

values specified here overwrite the ones in the upper header for this particular component of the corpus. Thus, for instance the information about the number of files in the text is updated for this particular element, as well as the number of sentences and the number of words per language. Figure 11 shows an example of header for the files written in Austria in 1999 in the construction domain.

```
<teiHeader>
    <fileDesc>
        <titleStmt>
            <title>B00_CONSTRUCTION - AUSTRIA - 1999</title>
          <funder>EU under FP7, Marie Curie Actions, SP3 People ITN, grant agreement 238405
            (project CLARA)</funder>
            <principal>Carla Parra Escartín</principal>
            <respStmt>
                <name>Carla Parra Escartín</name>
                <resp>corpus compilation, processing, encoding and markup</resp>
            </respStmt>
        </titleStmt>
        <extent>
            <measureGrp>
                <measure type="files">14</measure>
                <measure type="sentences">3899</measure>
                <measure type="words_DE-AT">33698</measure>
                <measure type="words_ES-ES">48239</measure>
            </measureGrp>
        </extent>
        <publicationStmt>
            <address>
             <addName>Institutt for lingvistiske, litterære og estetiske studier</addName>
                <addrLine>Postboks 7805</addrLine>
                <addrLine>5020 Bergen (Norway)</addrLine>
            </address>
            <date>2012</date>
            <publisher>Universitetet i Bergen</publisher>
            <pubPlace>Bergen, Norway</pubPlace>
            <distributor>Universitetet i Bergen</distributor>
            <availability status="restricted">
                <licence>Under negotiation</licence>
            </availability>
        </publicationStmt>
        <sourceDesc>
            <p>98/34/EC Directive, TRIS Database, European Commission</p>
        </sourceDesc>
    </fileDesc>
    <profileDesc>
        <creation>
            <date when="2011-2012">September 2012</date>
            <rs type="city">Bergen, Norway</rs>
        </creation>
        <langUsage>
            <language ident="DE-AT">German (Austria)</language>
            <language ident="ES-ES">Spanish (Spain)</language>
        </langUsage>
    </profileDesc>
</teiHeader>
```

Figure 11: <teiHeader> element of one of the TEI elements included in the TRIS corpus

III. The <text>. The <text> element is where the actual corpus is stored. When it is created a unique id is assigned to it to enable future referencing, extraction and usage upon user needs. This id includes information about the domain covered in the group of files, the country of origin (where the files where written) and the year in which they were written. Then, in the case of TRIS, it is further subdivided into single files grouped in a <group> element which includes all

files in the corpus in the form of individual <text> elements (i.e. there are as many <text> elements as files are in the corpus). Each individual file is also assigned a unique id which includes all the information related to the domain, the year and the name of the file in the EC database from which the files were retrieved. Since every file has been sentence aligned and is presented in two different languages, the element is used to divide the text between the source language and the target language. A final link group (<linkGrp>) is included in which the sentence alignment information is given by assigning to each sentence in the source language the corresponding sentence in the target language by means of their unique ids. In order to make the alignments, each sentence is assigned a unique id in which the source language and the sentence number are specified. Figure 12 displays a shortened text of the TRIS corpus encoded in TEI P5 to illustrate the usage of the different elements described above.

```
<text xml:id="B00_AUSTRIA_1999">
    <group>
        <text xml:id="B00Y1999File119990211">
            <body>
        <div xml:id="B00Y1999File119990211_DE-AT" xml:lang="DE-AT" type="source">
           <p xml:id="B00Y1999File119990211S1_DE-AT">Magistrat der Stadt Wien</p>
           <p xml:id="B00Y1999File119990211S2_DE-AT">Magistratsabteilung 351200
                    Wien, Dresdner Straße 75</p>
           <p xml:id="B00Y1999File119990211S3_DE-AT">Verordnung des Magistrates
           der Stadt Wien über die bis zum … befristete Zulassung des Fangsystems
                    „KAMINODUR AGS".</p>
                    ...
                </div>
        <div xml:id="B00Y1999File119990211_ES-ES" xml:lang="ES-ES" type="translation">
           <p xml:id="B00Y1999File119990211S1_ES-ES">Gobierno de la Ciudad de Viena</p>
           <p xml:id="B00Y1999File119990211S2_ES-ES">Sección de Gobierno 351200
                    Viena, Dredner Strasse 75</p>
           <p xml:id="B00Y1999File119990211S3_ES-ES">Reglamento del Gobierno de la Ciudad
           de Viena relativo a la homologación temporal hasta el ... del Sistema de chimeneas
                    "KAMINODUR AGS".</p>
                …
                </div>
        <linkGrp type="alignment" domains="#B00Y1999File119990211_DE-AT #B00Y1999File119990211_ES-ES">
           <link target="#B00Y1999File119990211S1_DE-AT #B00Y1999File119990211S1_ES-ES" />
           <link target="#B00Y1999File119990211S2_DE-AT #B00Y1999File119990211S2_ES-ES" />
           <link target="#B00Y1999File119990211S3_DE-AT #B00Y1999File119990211S3_ES-ES" />
                </linkGrp>
             </body>
        </text>
        <text>
        ...
        </text>
    </group>
</text>
```

Figure 12: Shortened sample of a <text> element in the TRIS corpus

IV. Automatically converting the TMX files to TEI P5. Encoding a corpus like TRIS in TEI P5 manually would be an error prone and tedious task. For this reason a simple python script that automatically processes the tmx files was written. This script reads the TMX file to be encoded in TEI P5, stores in different variables the information needed for the different values to be assigned, and processes the files and produces an XML file with the TEI structure explained above.

5. Conclusion

In this paper several issues with regards to the encoding of a corpus have been tackled. Specifically, encoding formats – UTF-8 for character encoding and XCES, TEI and TMX for corpus encoding – have been discussed. The TRIS corpus has been used as an example to understand the storyboard of the compilation of a corpus and the different stages in which encoding plays a particular role. Due to space restrictions it has not been possible to discuss standards for Part-

Of-Speech (POS) tagging and its integration in the overall structure. This would be the next step to be done but for instance in the case of the TRIS corpus it has been decided to release separately the output files of the POS tagger used (the TreeTagger POSTagger). Thus, POS tags will not be integrated into the overall TEI P5 encoding, although it could easily be done if desired.

When the compilation of the TRIS corpus began, it was my desire as a resource developer to produce a reusable and interoperable resource. To this end it was crucial to take into consideration issues such as the encoding of the corpus discussed in this paper. Moreover, these kinds of issues should be taken into account during the corpus compilation planning phase and prior to publicly releasing the corpus, because otherwise there would be the risk of creating a resource which is not useful for the community. This planning will also avoid any other researcher interested in using a newly compiled corpus having to convert and adapt it before actually using it for his/her own research purposes. As a resource developer, I deemed it important to study the different available possibilities and decide which was the best option both for my needs and for releasing a resource which is valuable for the NLP community as a whole, notwithstanding whether the interested parties belonged to academia, industry or both. Last, but not least, I would like to highlight the importance of documenting the compilation process of any linguistic resource. This documentation is not only valuable for future reference of the resource itself, but also for guidance and reference to future resource developers looking for solutions to challenges they face or strategies followed by previously developed resources. I hope that this paper serves future corpus developers as a reference on how to encode a corpus and the different types of encoding that they may have to take into account.

6. Acknowledgements

The research reported in this paper has received funding from the EU under FP7, Marie Curie Actions, SP3 People ITN, grant agreement 238405 (project CLARA[36]).

[36] http://clara.uib.no

References

Bánski, P. and A. Przepiórkowski (2010). TEI P5 as a Text Encoding Standard for Multilevel Corpus Annotation. In *Proceedings of Digital Humanities 2010, London*.

Borin, L. and J. Lindh (2011). Deliverable D4.1: Metadata descriptions and other interoperability standards. Version 1.0, 2011-05-02. Deliverable in the METANORD project (CIP 270899).

Guillemin, P. and S. Trillaud (2012, April/May). What has become of LISA's OSCAR standards? *Multilingual*, 38–41.

Localization Industry Standards (LIS) ETSI Industry Specification Group (ISG) (2013). ETSI GS LIS 002 V1.4.2 (2013-02): Localization Industry Standards (LIS); Translation Memory eXchange (TMX).

McEnery, A. and R. Xiao (2005). *Developing Linguistic Corpora: a Guide to Good Practice*, Chapter Chapter 4: Character encoding in corpus construction. AHDS Guides to Good Practice.

Morrison, A., M. Popham, and K. Wikander (2000). *Creating and Documenting Electronic Texts: A Guide to Good Practice*, Chapter 4: Markup: The key to reusability. AHDS Guides to Good Practice. Arts and Humanities Data Service.

Open Standards for Container/Content Allowing Reuse (OSCAR) (2008). Systems to manage terminology, knowledge, and content TermBase eXchange (TBX).

Parra Escartín, C. (2012, May). Design and compilation of a specialized SpanishGerman parallel corpus. In N. C. C. Calzolari, K. Choukri, T. Declerck, M. Uğur Doğan, B. Maegaard, J. Mariani, J. Odijk, and S. Piperidis (Eds.), *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Przepiórkowski, A. (2009). TEI P5 as an XML standard for treebank encoding. In M. Passarotti, A. Przepiórkowski, S. Raynaud, and F. Van Eynde (Eds.), *Proceedings of the Eighth International Workshop on Treebanks and Linguistic Theories (TLT 8)*, Milan, Italy, pp. 149–160.

Przepiórkowski, A. and P. Bánski (2011). Which xml standards for multilevel corpus annotation? In *Proceedings of the 4th conference on Human Language Technology: Challenges for computer science and linguistics*, LTC'09, Berlin, Heidelberg, pp. 400–411. Springer-Verlag.

Sasaki, F. (2010, June). *Linguistic Modeling of Information and Markup Languages. Contributions to Language Technology*, Volume 40 of *Text, Speech and Language Technology*, Chapter 4: Markup Languages and Internationalization, pp. 67–80. Springer-Verlag.

Simões, A. and S. Fernandes (2011). XML schemas for parallel corpora. In *XATA 2011: XML, associated technologies and applications*, Vila do Conde, Portugal, pp. 59–69.

Sperberg-McQueen, M. and L. Burnard (2009, February). TEI P5: Guidelines for electronic text encoding and interchange. Technical report, The TEI Consortium.

Steinberger, R., B. Pouliquen, A. Widiger, C. Ignat, T. Erjavec, and D. Tufiş (2006). The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In *In Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'2006*, pp. 2142–2147.

Tiedemann, J. and P. Wijnitz (2010, August). Deliverable D.2.1: Specification of data formats. Deliverable in the Let's MT project.

Walsh, N. and L. Muellner (1999, October). *DocBook: The Definitive Guide*. O'Reilly & Associates, Inc.

Wynne, M. (2005). *Developing Linguistic Corpora: a Guide to Good Practice*, Chapter Chapter 6: Archiving, distribution and preservation. AHDS Guides to Good Practice.