DTD-Driven Bilingual Document Generation

Arantza Casillas
Departamento de Automática, Universidad de Alcalá e-mail:arantza@aut.alcala.es

Joseba Abaitua
Facultad de Filosofía y Letras Universidad de Deusto, Bilbao e-amil:abaitua@fil.deusto.es

Raquel Martínez
Departamento de Sis. Informáticos y Programación, Facultad de Matemáticas
Universidad Complutense de Madrid e-mail:raquel@eucmos.sim.ucm.es

Abstract

Extensively annotated bilingual parallel corpora can be exploited to feed editing tools that integrate the processes of document composition and translation. Here we discuss the architecture of an interactive editing tool that, on top of techniques common to most Translation Memory-based systems, applies the potential of SGML's DTDs to guide the process of bilingual document generation. Rather than employing just simple task-oriented mark-up, we selected a set of TEI's highly complex and versatile collection of tags to help disclose the underlying logical structure of documents in the test-corpus. DTDs were automatically induced and later integrated in the editing tool to provide the basic scheme for new documents.

1 Introduction

This paper discusses an approach to the architecture of an experimental interactive editing tool that integrates the processes of source document composition and translation into the target language. The tool has been conceived as an optimal solution for a particular case of bilingual production of legal documentation, but it also illustrates in a more general way how to exploit the possibilities of SGML (ISO8879, 1986) used extensively to annotate a whole range of linguistic and extralinguistic information in specialized bilingual corpora.

SGML is well established as the coding scheme underlying most Translation Memory based systems (TMBS), and has been proposed as the coding scheme for the interchange of existing Translation Memory databases - Translation Memories eXchange, TMX (Melby, 1998). The advantages of SGML have also been perceived by a large community of corpus linguistics researchers, and big efforts have been made in the development of suitable markup options to encode a variety of textual types and functions -as clearly demonstrated by the Text Encoding Initiative, TEI; (Burnard & Speberg-MacQueen, 1995). While the tag-sets employed by TMBS are simple and task-oriented, TEI has offered a highly complex and versatile collection of tags. The guiding hypothesis in our experiment has been the idea that it is possible to explore TEI/SGML markup in order to develop a system that carries the concept of Translation Memory one step further. One important feature of SGML is the DTD. DTDs determine the logical structure of documents and how to tag them accordingly. We have concentrated on the accurate description of documents by means of TEI conformant SGML markup. The markup will help disclose the underlying logical structure of documents. From annotated documentation, DTDs can be induced and these DTDs provide the basic scheme to produce new documents. We have collected a corpus of official publications from three main institutions in the Basque Autonomous Region in Spain, the *Boletín Oficial de Bizkaia* (BOB, 1990-1995), *Boletín Oficial de Álava* (BOA, 1990-1994) and *Boletín Oficial del País Vasco* (BOPV, 1995). Documents in the corpus were composed by Administration clerks and translated by translators. Both clerks and translators have been using a wide variety of word-processors, although since 1994 MSWord has been generalized as the standard editing tool. Administrative documentation shows a regular structure, and is rich in recurrent textual patterns. For each document type different document tokens share a common global distribution of elements. Official document composers learn these global structures and apply them consistently. It is also the case that composers tend to reuse old

Document Type	%
Orden Foral	53%
Decreto Foral	22%
Resolución	13%
Extracto	5.4%
Acuerdo	3.4%
Norma Foral	1.9%
Anuncio	0.4%

Table 1: Types of documents in the corpus

document files when producing new documents of the same type. Despite the fact that no SGML software was used at the editing phase, texts in the corpus show regular logical structures and consistent distribution of text segments. Our main goal in tagging the corpus was to make all them explicit (Martinez, 1997). The most common type of document in the corpus , the *Orden Foral*, was chosen (see Table 1). We analysed some 100 tokens and hand-marked the most salient elements. The heuristics to identify these elements were later expressed in a collection of recognition routines in Perl and tested against a set of 400 tokens, including the initial 100. As a result of this process of automatic tagging of structural elements we produced a TEI/SGML tagged corpus with yet no corresponding overt DTD. In Section 2 we will explain how DTDs were later induced from the tagged corpus.

Once the corpus was segmented the next step was to align it. This was conducted at different levels: general document elements (`DIV`, `SEG`, `P`), as well as sentential and intra-sentential elements, such as `S`, `RS`, `NUM`, `DATE`, etc. (Martinez, 1998b). Aligned in this way, the corpus becomes an important resource for translation. Four complementary language databases may be obtained at any time from the annotated corpus: three translation memory databases (TM1, TM2, and TM3) as well as a terminology database (termbase). The three TMs differ in the nature of the translation units they contain. TM1 consists of aligned sentences than can feed commercial TM software. TM2 contains elements which are translation segments ranging from whole sections of a document or multi-sentence paragraphs to smaller units, such as short phrases or proper names. TM3 simply hosts the whole collection of aligned bilingual documents, where the whole document may be considered the translation unit. TM3 can be construed as a bilingual document-database. Much redundancy originates from this TM collection, although it should be noticed that they are all by-products derived from the same annotated bitext which subsumes them all. Good software packages for TM1 and TM3 already exist in the market, and hence their exploitation is beyond our interest (Trados Translator's Workbench, Star's Transit, SDLX, Déjá Vú, IBM's Translation Manager) for TM1; and any SGML browsing tool for TM3). The originality of our editing tool lies in a design which benefits from joining the potentiality of DTDs and the elements in TM2, as will be shown in sections 4 and 5.

2 DTD abstraction

SGML mark-up determines the logical structure of a document and its syntax in the form of a context-free grammar. This is called the Document Type Definition (DTD) and it contains specifications for:

- Names and content for all elements that are permitted to appear in a document.
- Order in which these elements must appear.
- Tag attributes with default values for those elements.

DTDs have been abstracted away from the annotations that were automatically introduced in the corpus. Similar experiments have been reported before in the literature. (Ahonen, 1995) uses a method to build document instances from tagged texts that consists of a deterministic finite automaton for each context model. Subsequently, these automata are generalized and converted into regular expressions which are easily transcribed into SGML content models. (Shafer, 1995) combines document instances with simplification rules. Our method is similar to Shafer's, but with a modification in the way rules reduce document instances. A tool to obtain a DTD for all document instances has been developed (Casillas, 1999). Given that source and target documents show some syntactic and structural mismatches, two different DTDs are induced, one for each language, and

Spanish Text:

<div0>
<div1> ... </div1>
<seg9 id=9ES2 corresp=9EU2> Contra dicha <rs type=law id=LES12 corresp=LEU10> Orden Foral </rs>, que agota la vía administrativa podrá interponerse recurso contencioso-administrativo ante la <rs type=organization id=OES9 corresp=OEU11> Sala de lo Contencioso-Administrativo del Tribunal Superior de Justicia del País Vasco </rs>, en el plazo de dos meses, contado desde el día siguiente a esta notificación sin perjuicio de la utilización de otros medios de defensa que estime oportunos. </seg9>

<seg10 id=10ES1 corresp=10EU1> Durante el referido plazo el expediente BHI-<num num=10094> 100/94 </num>-P05-A quedará de manifiesto para su exámen en las dependencias de <rs type=place id=PES3 corresp=PEU2> Bilbao calle Alameda Rekalde </rs>, <num num=30> 30 </num>, <num num=5> 5.a </num> y <num num=6> 6.a </num> plantas. </seg10>
</div0>

<closer id=pES13 corresp=pEU13 > <name> El Diputado Foral de Urbanismo Pedro Hernández González. </name> </closer>

Basque Text:

<div0>
<div1> ... </div1>
<seg9 id=9EU2 corresp=9ES2> <rs type=law id=LEU10 corresp=LES12> Foru agindu </rs> horrek amaiera eman dio administrazio bideari; eta beraren aurka <rs type=organization id=OEU10> Administrazioarekiko </rs> auzibide-errekurtsoa jarri ahal izango zaio <rs type=organization id=OEU11 corresp=OES9> Euskal Herriko Justizi Auzitegi Nagusiko Administrazioarekiko Auzibideetarako Salari </rs>, bi hilabeteko epean; jakinarazpen hau egiten den egunaren biharamunetik zenbatuko da epe hori; hala eta guztiz ere, egokiesten diren beste defentsabideak ere erabil litezke. </seg9>

<seg10 id=10EU1 corresp=10ES1> Epe hori amaitu arte BHI-<num num=10094> 100/94 </num>-P05-A espedientea agerian egongo da, nahi duenak azter dezan, <rs type=place id=PEU2 corresp=PES3> Bilboko Errekalde zumarkaleko </rs> <num num=30> 30.eko </num> bulegoetan, <num num=5> 5 </num> eta <num num=6> 6.</num> solairuetan. </seg10>
</div0>

<closer id=pEU13 corresp=pES13> <name> Hirigintzako foru diputatua. Pedro Hernández González. </name> </closer>

Figure 1: Ilustrates a sample of the annotated bitext

are paired through a correspondence table. Correspondences in this table can be up-dated, or deleted. At present, we have six DTDs, one for each document type in each language (there are three document types; Figure 2 shows a part of one of these DTDs). By means of these paired DTDs, document elements in each language are appropriately placed. In the process of generating the bilingual document, a document type must first be selected. Each document type has an associated DTD. This DTD specifies which elements are obligatory and which are optional. With the aid of the DTD, the source document is generated. The target document will be generated with the aid of the corresponding target DTD.

3 Joining TM2 and DTD

TM2 specifically stores a type of translation segment class, which we have tagged <seg1>, <seg2>... <segn>, <title> and <rs>, and which is relevant to the DTD. Segments tagged <segn> are variable recurrent language patterns very frequent in the specialized domain of the corpus and whose occurrence in the text is well established. These <segn> tags include two attributes: `id` and `correspond` which locate the aligned segment both in the corpus and in the database (Figure 1). Segments tagged <rs> are referring expressions which have been recognized, tagged and aligned and which correspond largely to proper names (Martinez, 1998a), (Martinez, 1998b). TM2 is managed in the form of a relational database where segments are stored as records. Each record in the database consists of four fields: the segment string, a counter for the occurrences of that string in the corpus, the tag and the attributes (`type`, `id` and `corresp`). Table 2 shows how the text fragment inside

```
<!ELEMENT LEGE - - (TEXT)>
<!ELEMENT TEXT - - (BODY)>
<!ELEMENT BODY - - (OPENER, DIV0, CLOSER)>

<!ELEMENT OPENER - - (TITLE, NUM, DATE, NAME?, SEG1)>
<!ELEMENT SEG1 - - (SEG11, (#PCDATA|RS|DATE|NUM)+)>
<!ELEMENT (SEG11, NUM, DATE, RS, NAME, TITLE) - - (#PCDATA)>

<!ELEMENT (DIV0) - - ( (#PCDATA |RS |NUM |DATE |SEG4|SEG5 |SEG6|SEG7 |SEG8 |SEG12 |SEG14
|SEG15)+, SEG9?, SEG10?)>
<!ELEMENT (SEG4, SEG5, SEG6) - - (#PCDATA)>
<!ELEMENT (SEG9, SEG10, SEG7, SEG8, SEG12, SEG14, SEG15) - - (#PCDATA|RS|DATE|NUM)+>

<!ELEMENT (CLOSER) - - (PLACENAME?,DATE?, NAME?)>
<!ELEMENT (PLACENAME) - - (RS)>

<!ATTLIST RS TYPE (ORGANIZATION| LAW| PLACE| UNCAT) UNCAT>
```

Figure 2: Part of the DTD of the type document *Orden Foral*

the `</div1>`...`</div0>` tags of Figure 1 renders three records in the database. Note how the content of the string field in the database maintains only the initial `<segn>` and `<rs>` tags. Furthermore, `<rs>` tagged segments inside `<segn>` records are simplified so that their content is dismissed and only the initial tag is kept (Lange et al., 1997). The reason is that they are considered variable elements within the segment (dates and numbers are also these type of elements). The strings *Orden Foral* of record 2 marked as `<rs type=law>` and *Sala de lo Contencioso-Administrativo del Tribunal Superior de Justicia del País Vasco* of record 3 `<rs type=organization>` are thus not included in record 1 `<seg9>`, since they may differ in other instantiations of the segment. These internal elements are largely proper names that vary from one instantiation of the segment to another. The `<rs>` tag can be considered to be the name of the varying element. The value of the type attribute `<rs type=law>` constraints the kind of referential expression that may be inserted in that point of the translation segment. Table 2 shows that source and target records may not have straight one-to-one correspondences. Although this is by no means the general case; only about 5.61%, (Martinez, 1998a), such one-to-N correspondences provide good ground to explain how the TM2 is designed. The asymmetry can be easily explained. The Spanish term *recurso contencioso-administrativo* has been translated into Basque by means of a category changing operation, where the Spanish adjective *administrativo* has been translated as a Basque noun complement *Administrazioarekiko* which literally means "Administration-the-with-of" triggering its identification as a proper noun.

Table 3 shows the way in which source language units are related with their corresponding target units, which, as can be observed, can be one-to-one or one-to-N. This means that one source element can have more than one translation.

TM2 is created in three steps:

- First, non-pertinent tags are filtered out from the annotated corpus. Tags marking sentence `<s>` and paragraph `<p>` alignment are removed because they are of no interest for TM2 (recall that they are registered in TM1).
- Second, translation segments `<segn>`, `<title>` phrases and referential expressions `<rs>` are detected in the source document and looked up in the database.
- Third, if they are not already present in the database, they are stored each in its database and values of the `id` and `corresp` attributes are used to set the correspondence between source and target database.

4 Composition Strategy

Every phase in the process is guided by the markup contained in TM2 and the paired DTDs

Spanish Unit	Basque Unit
<seg9> Contra dicha <rs type=law>, que agota la vía administrativa podrá interponerse recurso contencioso-administrativo ante la <rs type=organization>, en el plazo de dos meses, contado desde el día siguiente a esta notificación, sin perjuicio de la utilización de otros medios de defensa que estime oportunos. que estime oportunos.	<seg9> <rs type=law> horrek amaiera eman dio administrazio bideari; eta beraren aurka <rs type=organization> auzibide-errekurtsoa jarri ahal izango zaio <rs type=organization>, bi hilabeteko epean; jakinarazpen hau egiten den egunaren biharamunetik zenbatuko da epe hori; hala eta guztiz ere, egokiesten diren beste defentsabideak ere erabil litezke.
<rs type=law> Orden Foral	<rs type=law> Foru agindu
	<rs type=organization> Administrazioarekiko
<rs type=organization> Sala de lo Contencioso-Administrativo del Tribunal Superior de Justicia del País Vasco	<rs type=organization> Euskal Herriko Justizi Auzitegi Nagusiko Administrazioarekiko Auzibideetarako Salari

Table 2: Source and targe language record samples in TM2

Spanish Unit	Basque Unit
<rs type=organization id= corresp=> Boletín Oficial de Bizkaia	<rs type=organization id= corresp=> Bizkaiko Aldizkari Ofizialea <rs type=organization id= orresp=> Bizkaiko Engunkari Ofizialea <rs type=organization id= corresp=> Bizkaiko Boletin Ofizialea
<seg3> dispongo	<seg3> xedatu dut <seg3> xedatzen duen

Table 3: Source language units related with their corresponding target language units

which control the application of this markup. The composition process follows two main steps which correspond to the traditional source document generation and translation into the target document. The markup and the paired DTD guides the process in the following manner:

1. Before the user starts writing the source document, he must select a document type, i.e., a DTD. This has two consequences. On the one hand, the selected DTD produces a source document template that contains the logical structure of the document and some of its contents. On the other hand, the selected source DTD triggers a target paired DTD, which will be used later to translate the document. There are three different types of elements in the source document template:

    - Some elements are mandatory and are provided to the user, who must only choose its content among some alternative usages (s/he will get a list of alternatives ordered by frequency, for example <title>). Other obligatory elements, such as dates and numbers, will also be automatically generated.
    - Some other elements in the template are optional (e.g., <seg9>). Again, a list of alternatives will be offered to the user. These optional elements are sensitive to the context (document or division type), and markup is also responsible for constraining the valid options given to the user. Obligatory and optional elements are retrieved from TM2, and make a considerable part of the source document.
    - All documents have an important part of their content which is not deter-

Word/doc.	Num. doc.	TM2
0-500	378	34.91
500-1,000	25	14.01
More 1,000	16	3.01
Weighted mean		31.8

Table 4: % generated by TM2

Word/doc.	Num. doc.	TM2	Alg.	Total
0-500	378	28.3	7.7	36
500-1,000	25	12.3	9.6	21.3
More 1,000	16	4.7	6.4	10.7
W. M.		26.5	7.6	34.2

Table 5: % translated by TM2 and algorithms

mined by the DTD (<div1>). It is the most variable part, and the system lets the writer input text freely. It is when TM2 has nothing to offer that TM1 and TM3 may provide useful material. Given the recurrent style of legal documentation, it is quite likely that the user will be using many of the bilingual text choices already aligned and available in TM1 and TM3.

2. Once the source document has been completed, the system derives its particular logical structure, which, with the aid of the target DTD, is projected into the resulting target logical structure.

5 Evaluation

Table 4 shows the number of words that make up the segments stored in TM2 from the source documents. There is a line for each document size considered. We can see that the average of segments contained in TM2 is 31.8%, on a scale from 34.91% to only 3.01%. The amount of segments dealt with in this way largely depends on the size of the document. Short documents (90.21) have about 35% of their text composed in this way. This figure goes down to 3% in documents larger than 1,000 words. This is understandable, in the sense that the larger the document, the smaller proportion of fixed sections it will contain.

Table 5. shows the number of words that are proposed for the target document. These translations are obtained from what is stored in TM2 complemented by algorithms designed to translate dates and numbers. We can see that the average of document translated is 34%. Short documents have 36% of their text translated, falling to above 11% in the case of large documents.

6 Conclusions

We have shown how DTDs derived from descriptive markup can be employed to ease the process of generating bilingual dedicated documentation. On average, one third of the contents of thedocuments can be automatically accounted for. It must also be pointed out that the part being dealt with represents the core structure, lay-out and logical components of the text. The remaining two-thirds of untreated document can still be managed with the aid of sentence-oriented TMBS, filling in the gaps in the overall skeleton provided by the target template. Composers may also browse TM3 to retrieve whole blocks for those parts which are not determined by the DTD. One of the clear targets for the future is to extend the coverage of the corpus and to test structural taggers against other document types. A big challenge we face is to develop tools that automatically perform the recognition of documents from less restricted and more open text types. However, we are not sure of the extent of the practicality of such an approach. An alternative direction we are presently considering is to establish a collection of pre-defined document types, which would be validated by the institutional writers themselves. It is a process currently being implemented in the Basque administration to define document models for writers and translators to follow. What we have demonstrated is that paired DTDs, complemented with rich language resources of the kind defined in this paper, allow for the design of optimal editing environments which would combine both document composition and translation as one single process. All the resources needed (DTDs and TMs) can be induced from an aligned corpus.

7 Acknowledgements

This research is being partially supported by the Spanish Research Agency, project ITEM, TIC-96-1243-C03-01.

References

H. Ahonen. Automatic Generation of SGML Content Models. *Electronic Publishing,* 8(2-3):195-206, 1995.

L. Burnard, C. Speberg-McQueen. TEILite: An Introduction to Text Encoding for Interchange. *URL://http://www-tei.uic.edu/orgs/tei/intros/teiu5.tei,* 1995.

Casillas A., Abaitua J., Martínez R. Extracción y aprovechamiento de DTDs emparejadas en corpus paralelos. *Procesamiento del Lenguaje Natural,* 25:33-41, 1999.

ISO 8879, Information Processing–Text and Office Systems–Standard Generalized Markup Language (SGML). *International Organization For Standards,* 1986, Geneva.

J. Langé, É Gaussier, B. Daile. Bricks and Skeletons: Some Ideas for the Near Future of MATH. *Machine Translation,* 12:39-51, 1997.

Martínez R., Abaitua J., Casillas R. Bilingual parallel text segmentation and tagging for specialized documentation. *Proceedings of the International Conference Recent Advances in Natural Language Processing (RANLP'97),* 369-372, 1997.

Martínez R., Abaitua J., Casillas A.. Bitext Correspondences through Rich Markup. *36th Annual Meeting of the Association for Computational Linguistics abd 17 International Conference on Computational Linguistics (COLING-ACL'98),* 812-818, 1998.

Martínez R., Abaitua J., Casillas A.. Aligning tagged bitexts. *Sixth Workshop on Very Large Corpora,* 102-109, 1998.

A. Melby. Data Exchange from OSCAR and MARTIF Projects. *First International conference on Language Resources & Evaluation,* 3-7, 1998.