The Challenge of Parallel Text Processing

Milena Slavcheva

Institute for German Language
R5, 6-13, D-68161 Mannheim, Germany
Bulgarian Academy of Sciences
CLPII, LMD, 25 A, Acad. G. Bonchev St., 1113 Sofia, Bulgaria
milena@lml.bas.bg

Abstract. The paper presents the technology of building a large German - French parallel corpus consisting of official documents of the European Union and Switzerland, and private and public organisations in France and Germany. The texts are morphosyntactically annotated, aligned at the sentence level and marked up in conformance with the TEI guidelines for standardised representation. The multi-level alignment method is applied; its precision is improved due to the correlation with the constraints of the classical alignment method of Gale and Church. The alignment information is encoded externally to the parallel text documents.The process of creating the corpus is an interesting algorithm of applying a number of software tools and adjusting intermediate production results.

1 Introduction

Parallel texts are a basic resource for data-driven multilingual research in the domain of Human Language Technology. The first and underlying step to terminology and translation studies is the compilation of a text corpus according to the most up-to-date requirements in the field. This paper presents the technology of building a large German - French parallel corpus of proposed size of 30 million words (15 million words per language). At present, the algorithm of production has been tested on 2 million words (1 million words per language). The bilingual corpus consists of official documents of the European Union and Swutzerland, and private and public organisations in France and Germany. The parallel German and French texts of the corpus are morphosyntactically annotated, aligned at the sentence level and marked up in conformance with the TEI guidelines [1] for creating text documents in a standardised format.

2 The input and output of the parallel text processing

The initial input to the processing are texts in html or pdf format, downloaded from the WWW and converted into plain text format. The result of the conversion are text files where the structure of texts, essential for their automatic processing, is preserved to a varying extent. Paragraphs are the anchor structural units regarding the alignment procedure and the software tools used in the production chain.

Depending on the specific features of their layout and the extent to which their structure is violated, the text files are preprocessed automatically and also manually, so that their quality is improved. The preprocessing is accelerated by the "visibility" of the actual paragraph structure, since many of the texts are treaties, laws and administrative documentation containing much numbering and having conventional structure. At the same time, the texts are problematic due to their administrative style: they contain numerous lists of items that have to be delimitted in proper text portions having the form of sentences or paragraphs. A serious problem is the specific punctuation in the German and French versions or the lack of punctuation marks in the text portions containing lists of items. The proper stucturing of the texts has to be defined in such a way as to ensure the successful operation of the software tools that perform the markup assignment and the alignment. The solution is to split the list items and similar structural units into separate paragraphs. The delimitting of paragraphs interacts specifically with the tokenizing of sentences regarding the list items and the necessity to achieve parallel structuring of the corresponding German and French texts. The greater number of paragraph units, and, respectively, the smaller number of sentences within paragraphs improves the precision of the alignment process. We take the liberty to assign logical structure to the texts which enhances the automatic procedures, since there is a full change of the representation mode when compared to the multi-media visualisation of the texts in the WWW environment.

The final output of the processing are legal TEI documents consisting of two parts: 1) a header providing the metatextual information, and 2) a text body consisting of encoded basic structural units (paragraphs) and basic linguistic units (sentences and words with attached morphosyntactic tags and lemmas). The paragraphs and sentences are uniquely identified within the text. Figure 1 provides an example of a final German text document and its parallel French counterpart.

Source text (German)

```
<p id="d1p3">
<s id="d1p3s3">
<w ana="CARD">1</w>
<c ana="PUN">.</c>
<w ana="ADJD" lemma="gestuetzt">gestuetzt</w>
<w ana="APPR" lemma="auf">auf</w>
<w ana="ART" lemma="d">die</w>
<w ana="NN" lemma="unknown">Schlussfolgerungen</w>
...
</s>
</p>
```

Target text (French)

```
<p id="d1p3">
<s id="d1p3s3">
<w ana="ADJ:num">1</w><c ana="PON:sep">.</c>
<w ana="VER:pper" lemma="voir">vu</w>
<w ana="DET:def" lemma="le">les</w>
<w ana="NOM" lemma="conclusion">conclusions</w>
<w ana="PRE" lemma="sur">sur</w>
...
</s>
</p>
```

Figure 1. Parallel German and French TEI conformant documents

The alignment between corresponding segments in the bilingual parallel text documents is on the sentence level. Its encoding is externalised and is in the form of a separate document file where a link group element combines link elements indicating the correspondence of textual segments in the German and French parallel documents. Figure 2 represents the encoding of the alignment information.

```
<linkGrp type="alignment" source="/110104tagdere.txt"
crdate="empty" targOrder="Y" targType="seg" targFunc="null null"
target="/110104tagfrre.txt" domains="b1 b1"
evaluate="all">
<xptr from="ID (d1p1s1)" id="x1"/>
<xptr from="ID (d1p2s2)" id="x2"/>
<xptr from="ID (d1p3s3)" id="x3"/>
...
<link targets="d1p1s1 x1"/>
<link targets="d1p2s2 x2"/>
<link targets="d1p3s3 x3"/>
...
</linkGrp>
```

Figure 2. Alignment encoding

The particular format of the standardised representation was defined by the following factors: the structural peculiarities of the texts, the software tools available, the rich linguistic information encoded in the texts, the large amount of textual data, and the possibility for further elaboration and processing of the parallel corpus.

The markup scheme was defined so that those tags are assigned to the text structural units which are minimally required by MLAlign [2] - the program we use for aligning and for encoding the alignment information. It is also defined so that the tags can be assigned automatically using the programs available.

The elements requred by MLAlign are divisions, paragraphs and sentences supplied by the *id* attribute which identifies them uniquely within the standardised document. Each text represents one single division so that the requirements of the aligning program are fulfilled without further elaboration of the division structure. In defining the sentence boundaries and assigning the corresponding tags, accomplished automatically by MARK-ALISTeR [3], the wellknown problem between linguistic plausibility and formal convenience necessary for the automatic processing arises. In fact, the sentence delimiters are used for defining suitable portions of text which are rather sentence-like chunks enhancing the correct operation of the aligning tools. At the pragraph level, there is no other differentiation of structural elements than the paragraph.

The German and French documents also contain the appropriate TEI markup for the morphosyntactic annotation attached to each word token in the texts. The morphosyntactic tags and the lemmas are values of the respective *ana* and *lemma* attributes specifying all word and punctuation elements in the texts.

3 The alignment method and the documentation of the textual data

The multi-level alignment method [2] is applied; it is based on the logical structure of texts represented in a hierarchy of elements. The idea is to make use of the information present in texts that are already marked up and to obtain alignment results at levels of different depth in the structural hierarchy of texts (e.g., division, paragraph or sentence level). This hierarchical alignment algorithm is meant to cope with cases where the source and target texts may not have been encoded in a strictly parallel way at the intermediate levels (e.g., paragraphs) between the root element (i.e., division) and the leaves (i.e., sentences), and it allows the recombination of elements at a given level, even if they belong to two different units at the upper level [2]. The method is designed for application to independently created SGML or XML encoded text documents where the encoding scheme may vary.

The MLAlign program [2] is the software application of the multi-level alignment method which we use to align the sentence level segments in the parallel German and French texts. The input to the program are TEI conformant text documents, and the output is a link-group file with pointers to the aligned segments (i.e., sentences) in the parallel texts. Thus the alignment information is encoded externally to the text documents. This particular strategy of encoding and storing different types of information in different documents corresponds to the idea of general use and multiple reuse of standardised linguistic resources.

Although, as pointed above, the hierarchical alignment method is designed to produce correct alignments at a given level (primarily at the sentence level) in case there are discrepancies in the parallel structure at a higher level (usually the paragraph level), in reality the precision rate of the alignment operation is highly correlated to the quality of the encoding in the input text documents. In the case of the parallel text processing described in this paper, the quality of the performance of MLAlign is highly improved by the following factors: 1) the German and French TEI encoded text documents are produced simultaneously and have uniform standardised representation; 2) correlation is established with the constraints of the classical alignment method of Gale and Church [4].

In the initial steps of the production chain, the sentence tokenizer and the paragraph and sentence boundaries marker in the German and French texts is MARK-ALISTeR [3], a software system for alignment. It is based on the Gale and Church statistical method for sentence alignment whose high rate of precision relies on the equal number of paragraphs in the parallel texts. MARK-ALISTeR provides a friendly editing environment for the preprocessing of the source and target texts so that the parallel structure necessary for accomplishing the sentence alignment is obtained.

The two alignment programs have a different design, hence the different input, intermediate and final output, and different utilities for obtaining the desired results. In the production chain of textual resources described in this paper, the advantages of the two approaches to the alignment process and to the representation of the alignment results are utilised and combined in an efficient way.

4 The technology of attaining the desired final result

The process of creating the corpus is an interesting algorithm of applying a number of software tools, obtaining intermediate products and adjusting the output of a given "milestone" program so that it becomes the appropriate input to the "milestone" program coming next in the production chain. The operating peculiarites of the programs determine the order of their application so that the production chain is the most optimal and economical one.

After the initial preprocessing of the texts as described in Section 2, the texts enter MARK-ALISTeR, a system for marking, aligning and searching translation equivalents. The editing facilities of the preprocessing mode of the system are used for obtaining strict parallel paragraph structure where necessary. The program acts as a sentence boundaries delimiter of the two parallel texts. The precision of the sentence delimiting is decreased by the frequent occurence of specific abbreviations in the German and French texts due to their administrative style (e.g, the German z.B.). The program editor provides for correction of the automatically defined sentence boundaries. A valuable function of MARK-ALISTeR is the assignment of TEI markup to the paragraph and sentence units and the option of saving the two marked texts separately as an intermediate result. The output to the aligning proper function of the system is a bitext with no markup asigned which contains sequences of alignments on the sentence level. This output is stored as a result, but the production chain continues with the further processing of the files which are the output of the program marking operation.

The files in question are the input to the IMS TreeTagger [5] developed for German, French and English. It assigns a morphosyntactic tag and a lemma to each word-level token in the German and French texts.

The output files from the TreeTagger are automatically cleaned up and converted into TEI conformant format by a specially developed program. The division, paragragh and sentence elements are assigned unique identifiers. The result is shown in Figure 1. A header is attached to each text document according to the requirements of the TEI guidelines.

The TEI conformant parallel text documents are then processed by the MLAlign program, and the link-group file encoding the alignment information is generated (Figure 2) which consists of three parts: 1) external pointers to the segments of the target text that will be aligned; 2) links between the units of textual segments to be aligned that consist of more than one text unit; 3) links between aligned segments in the source and target texts. The alignment representation scheme is XML conformant.

5 Conclusion

The data-driven linguistic investigations require textual data of large quantity whose production is tedious and time-consuming, and very often at the expense of quality. The technology of producing the large bilingual parallel corpus described in this paper provides a reasonable degree of quality of the textual data.

Acknowlegments

The author is indebted to Elena Paskaleva, Martin Wynne, Alexander Genchev, Tomaz Erjavec, Helmut Schmid and Laurent Romary for the invaluable help.

References

1. Sperberg-McQueen, C.M., Burnard, L. (eds.): Guidelines for Electronic Text Encoding and Interchange. Text Encoding Initiative. Chicago and Oxford (1994)
2. Romary, L., Bonhomme, P.: Parallel Alignment of Structured Documents. In: Jean Veronis (ed.): Parallel Text Processing. Kluwer Academic Publisher (to appear)
3. Paskaleva, E., Mihov, S.: Second Language Acquisition from Aligned Corpora. In: Proceedings of the International Conference "Language Technology and Language Teaching", Groningen (1997)
4. Gale, W., Church, K.: A Program for Aligning Sentences in Bilingual Corpora. In: Computational Linguistics 19(1), (1993) 75-102
5. Schmid, H.: Probabilistic Part-of-Speech Tagging Using Decision Trees. In: Proceedings of International Conference on New Methods in Language Processing, Manchester, UK (1994)