MLIS-5016 TransAccount

WP 1

Accounting Text Database

Report
Version 1.0 from 29/03/2002
Deliverable D1.1

ALPNET Technology GmbH

AT internal

TITLE:	Accounting Text Database
PROJECT:	TransAccount / Deliverable D1.1
AUTHOR:	Waldhör
ESTABLISHMENT:	AT
ISSUE DATE:	29/03/2002
DOCUMENT ID:	TransAccount_01/01
VERSION:	1.0
STATUS:	initial
DISTRIBUTION:	TransAccount
LOCATION:	textdatabase-d-1-1.doc
KEYWORDS:	TransAccount, translation technology

Date:	Version:	Author:	Comments:
29.03.2002	1.0	Waldhör Klemens	

Contents

1 Changes to previous version

1.1 Version 1.0

New Version

Accounting Text Database

2 SUMMARY

This document describes the process how the text database as created.

3 Structure of the Document

Chapter 4 gives details about the files supplied by PwC, while chapter 5 explains how the pdf data have been converted. Chapter 6 explains the alignment results.

4 The Input Document

4.1 PwC PDF Files

The accounting document data have been delivered by PwC as **pdf files**.

4 file pairs consisting of English - French
2 files in English only
2 file in French only

The files have been moved into four directories. The CD-ROM contains those files in the directory: **TransAccountReview-2\D1.1\PwC-Data-Reports-PDF-TXT-RTF**.

Directory	Content	Comment
en	Lafarge2000annualreport.pdf Lafarge2000annualreport.rtf Lafarge2000annualreport.txt Nestle2001comptesconsos.pdf Nestle2001comptesconsos.rtf Nestle2001comptesconsos.txt Nestle2001rapportannuel.pdf Nestle2001rapportannuel.rtf Nestle2001rapportannuel.txt StGobain2000annualreport.pdf StGobain2000annualreport.rtf StGobain2000annualreport.txt	Original documents delivered as pdf files English documents where French equivalents exist
en-mono	Novartis2001annualreport.pdf Novartis2001annualreport.rtf Novartis2001annualreport.txt Novartis2001financialreview.pdf Novartis2001financialreview.rtf Novartis2001financialreview.txt	Original documents delivered as pdf files English documents where no French equivalents exist
fr	Lafarge2000annualreport.pdf Lafarge2000annualreport.rtf Lafarge2000annualreport.txt Nestle2001comptesconsos.pdf Nestle2001comptesconsos.rtf Nestle2001comptesconsos.txt Nestle2001rapportannuel.pdf Nestle2001rapportannuel.rtf Nestle2001rapportannuel.txt StGobain2000annualreport.pdf StGobain2000annualreport.rtf StGobain2000annualreport.txt	Original documents delivered as pdf files French documents where English equivalents exist
fr-mono	Valeo2000comptes.pdf Valeo2000comptes.rtf Valeo2000comptes.txt Valeo2001comptes.pdf Valeo2001comptes.rtf Valeo2001comptes.txt	Original documents delivered as pdf files French documents where no English equivalents exist

4.2 Amyot PDF Files

A similar procedure has been applied to the Amyot files. Please note that only a fraction of those files could be converted to TXT format as a lot of them have been read-only protected.

The CD-ROM contains those files in the directory: **TransAccountReview-2\D1.1\Amyot-Data**. The files which could be converted are contained in the following list. They are contained in the sub-directories **TransAccountReview-2\D1.1\Amyot-Data\en** resp. **TransAccountReview-2\D1.1\Amyot-Data\fr**.

en	fr
ACC97.txt	ACC97.txt
ACC98.txt	ACC98.txt
ALC97.txt	ALC97.txt
ALC98.txt	ALC98.txt
ALCA96.txt	ALCA96.txt
AXA97.txt	AXA97.txt
BIC97.txt	BIC97.txt
BUL96.txt	BUL96.txt
BUL97.txt	BUL97.txt
CHA97.txt	CHA97.txt
DAN97.txt	DAN97.txt
DAN98.txt	DAN98.txt
ELF97.txt	ELF97.txt
ELF98.txt	ELF98.txt
GAZ97.txt	GAZ97.txt
GUE98.txt	GUE98.txt
LAG97.txt	LAG97.txt
LEO97.txt	LEO97.txt
LEO98.txt	LEO98.txt
LVM96.txt	LVM96.txt
PRO97.txt	PRO97.txt
RHO98.txt	RHO98.txt
SAN98.txt	SAN98.txt
SEB97.txt	SEB97.txt
SOC96.txt	SOC96.txt
SYN98.txt	SYN98.txt
UNI97.txt	UNI97.txt

5 The Conversion Process

As for the alignment process (see D2.3) the documents have to be in a format like RTF, HTML or pure text format Alpnet tried to convert the PDF documents in a first step to RTF and txt format. A html conversion as also done but it turned out that those results are of a very bad quality.

A main problem is the handling of carriage return. It is not really clear under which conditions those character sequences are inserted into the converted document and when not. The problem here is that the segmenting process (done before alignment) in many cases is forced to use CR/LF as "paragraphs" which may lead to incorrect segmentation. Due to the late delivery of the data by PwC (March 22nd 2002) there was also no time to invest in correcting these problems.

As the PDF documents did not contain any structural information a special tool was used from Adobe which claims to re-create the structural information of the PDF files. This tool is called “Make Accessible”. All supplied PDF files have been converted to a new PDF document. In a next step the new PDF files have been converted to RTF, html and text format.

A first alignment was done using the RTF files (4 RTF EN - FR pairs). The alignment tool was executed using a special structural alignment mode which takes format information etc. into account. This step is executed before the actual statistical alignment is applied. Unfortunately it turned out that the results are of very poor quality because the RTF of conversion preserves the format information in a quit unpredictable way. Thus those alignments have been deleted.

In a next step the same procedure was applied to the html files. The results did not differ, thus those alignments had to be deleted too.

In the last try the pure text files have been used. The results there seem to be a little better, but it still requires manual post processing which is currently under way. Two files pairs created some ueful results, where the other results are questionable.

As a final resume the usage of PDF files cannot be re-commended.

6 The Alignment Results

6.1 PwC Aligned Files

As mentioned only the documents **Lafarge2000annualreport** and **Nestle2001comptesconsos** produced useful results.

Lafarge2000annualreport.FR.ali
Lafarge2000annualreport.FR.xml.ali

Nestle2001comptesconsos.FR.ali
Nestle2001comptesconsos.FR.xml.ali

The files are contained in the directory: **TransAccountReview-2\D1.1\AlignedFiles\PwC**

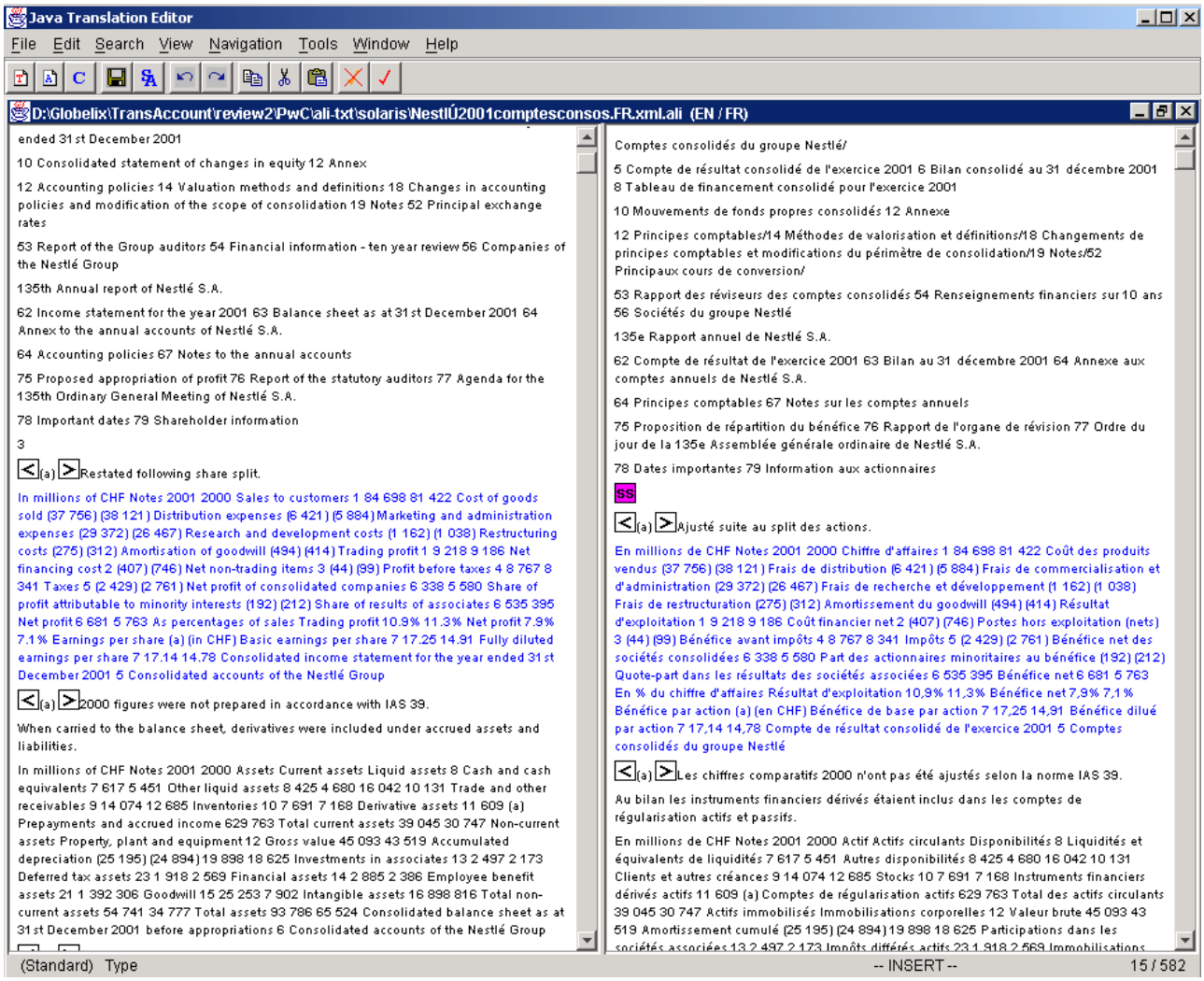
The data are available in the Globelix alignment format. Both formats are UCS-2 base, the first one based on the older alignment format produced by Globelix, the second one an xml version of the aligned files which is used by the Java Translation Editor.

The general structure of a standard alignment entry is as follows:

```
<al inst="#4AL136113" ssn="1">
 <source sn="12" status=""><s><st>Growth strategy and value
creation</st></s></source>
 <target sn="13" lan="3" tsn="1" status=""><s><st>Stratégie de croissance et
création de valeur</st></s></target>
</al>

<al inst="#4AL136113" ssn="1">
 <source sn="13" status=""><s><st>Innovation</st></s></source>
 <target sn="14" lan="3" tsn="1" status=""><s><st>Innovation</st></s></target>
</al>

<al inst="#4AL136113" ssn="1">
 <source sn="14" status=""><s><st>Environment</st></s></source>
 <target sn="15" lan="3" tsn="1"
status=""><s><st>Environnement</st></s></target>
</al>

<al inst="#4AL136113" ssn="1">
 <source sn="15" status=""><s><st>Human Resources and
Organisation</st></s></source>
 <target sn="16" lan="3" tsn="1" status=""><s><st>Ressources humaines et
Organisation</st></s></target>
</al>

<al inst="#4AL136113" ssn="2">
 <source sn="18" status=""><s><st>6</st></s> <s><st>8</st></s></source>
 <target sn="18" lan="3" tsn="1"
status=""><crit><s><st>Repères</st></s></crit></target>
</al>
```

Each entry starts with an al tag (<al> which contains an application reference (inst="#4AL136113") followed by the number of segments of the source part (ssn="1"); this indicates that the source part just contains one segment. Then the source element follows (<source>) which contains the reference to the segment number in the source document (sn="12") and a status indicator which is mainly used to store information about changes. The <s><st> element contains the actual segments. This is followed by the <target> segment which contains also a reference to the segment number in the original target document (sn="15"), a language indicator (lan="3") and the number of segments (tsn="1") in the target align part followed by a possible status information. As shown in the last example aligned segments which may be problematic a <crit> tag indication is used.

The next picture shows an aligned file in the Java Translation Editor.



Blue lines indicates corresponding alignments.

6.2 Amyot Aligned Files

The same procedure was applied to the text files which have been produced from the Amyot pdf files.

Results are contained in: **TransAccountReview-2\D1.1\AlignedFiles\Amyot**

Ali files:

ACC97.FR.ali ACC98.FR.ali ALC97.FR.ali ALC98.FR.ali
ALCA96.FR.ali AXA97.FR.ali BIC97.FR.ali BUL96.FR.ali
BUL97.FR.ali CHA97.FR.ali DAN97.FR.ali DAN98.FR.ali
GAZ97.FR.ali GUE98.FR.ali LEO98.FR.ali LVM96.FR.ali
PRO97.FR.ali SAN98.FR.ali SOC96.FR.ali SYN98.FR.ali
UNI97.FR.ali

Ali XML files:

ACC97.FR.xml.ali ACC98.FR.xml.ali ALC97.FR.xml.ali ALC98.FR.xml.ali
ALCA96.FR.xml.ali AXA97.FR.xml.ali BIC97.FR.xml.ali BUL96.FR.xml.ali
BUL97.FR.xml.ali CHA97.FR.xml.ali DAN97.FR.xml.ali DAN98.FR.xml.ali

GAZ97.FR.xml.ali GUE98.FR.xml.ali LEO98.FR.xml.ali LVM96.FR.xml.ali
PRO97.FR.xml.ali SAN98.FR.xml.ali SOC96.FR.xml.ali SYN98.FR.xml.ali
UNI97.FR.xml.ali

7 References

D2.3 Integrated thesaurus-based translators between US and FR system, manual3.pdf